Article

Ultra-Short-Term Distributed Photovoltaic Power Probabilistic Forecasting Method Based on Federated Learning and Joint Probability Distribution Modeling

Yubo Wang [1], Chao Huo [1], Fei Xu [2,*], Libin Zheng [1] and Ling Hao [2]

[1] Beijing SmartChip Microelectronics Technology Company Limited, Beijing 102200, China; yubowang_zhixin@163.com (Y.W.); chaohuo_zhixin@163.com (C.H.); libinzheng_zhixin@163.com (L.Z.)
[2] State Key Laboratory of Power System Operation and Control (Department of Electrical Engineering), Tsinghua University, Beijing 100084, China; haolingtinghua2018@163.com
* Correspondence: xufeitsinghua@163.com

Abstract: The accurate probabilistic forecasting of ultra-short-term power generation from distributed photovoltaic (DPV) systems is of great significance for optimizing electricity markets and managing energy on the user side. Existing methods regarding cluster information sharing tend to easily trigger issues of data privacy leakage during information sharing, or they suffer from insufficient information sharing while protecting data privacy, leading to suboptimal forecasting performance. To address these issues, this paper proposes a privacy-preserving deep federated learning method for the probabilistic forecasting of ultra-short-term power generation from DPV systems. Firstly, a collaborative feature federated learning framework is established. For the central server, information sharing among clients is realized through the interaction of global models and features while avoiding the direct interaction of raw data to ensure the security of client data privacy. For local clients, a Transformer autoencoder is used as the forecasting model to extract local temporal features, which are combined with global features to form spatiotemporal correlation features, thereby deeply exploring the spatiotemporal correlations between different power stations and improving the accuracy of forecasting. Subsequently, a joint probability distribution model of forecasting values and errors is constructed, and the distribution patterns of errors are finely studied based on the dependencies between data to enhance the accuracy of probabilistic forecasting. Finally, the effectiveness of the proposed method was validated through real datasets.

Keywords: distributed photovoltaic; ultra-short-term power forecasting; federated learning; spatiotemporal correlation; joint probability distribution

1. Introduction

In recent years, with the rapid development of the economy and increased reliance on fossil fuels, including coal and oil, the earth's ecological environment has been seriously threatened [1]. To alleviate environmental pressures, countries have implemented various policies [2]. Among these, owing to its non-polluting nature, renewable energy development has emerged as a key environmental governance strategy, gradually replacing traditional fossil fuels as the cornerstone of future energy systems. Promoting new energy generation has also become a critical initiative in global environmental governance. New energy generation encompasses wind power, photovoltaic (PV) power generation, geothermal power generation, and other forms of energy. Harnessing solar power, characterized by its

plentiful supply, environmental friendliness, and sustainable qualities, has seen widespread adoption and application across the globe [3,4]. DPV power generation has rapidly gained popularity due to its decentralized, clean, and highly efficient nature [5]. According to statistics from the International Energy Agency, global DPV capacity is projected to reach 917.1 GW in 2023 and 3467.1 GW by 2030. However, PV power generation is characterized by significant intermittency, randomness, and uncertainty. The large-scale integration of DPV systems significantly complicates the operation and control of distribution networks. An accurate forecast of DPV power is crucial for the operation and management of active distribution networks and the effective utilization of DPV energy. It is vital to ensure the safe and stable operation of active distribution networks. Moreover, these forecasts enable PV power sellers to adjust their market strategies promptly, minimizing potential losses. Therefore, an accurate forecast of DPV power plant output is essential.

Ultra-short-term PV power forecasting can be categorized into three main approaches: physical modeling methods, statistical methods, and machine learning methods. The physical modeling method develops a PV power model by applying the principles of solar radiation transmission, energy conversion in PV systems, and associated physical phenomena [6–8]. This method utilizes Numerical Weather Prediction (NWP) data along with photovoltaic panel data to achieve highly accurate and interpretable forecasts. Statistical methods, in contrast, rely on historical data and statistical principles to forecast future PV power outputs. This method analyzes historical power output data along with related influencing factors to identify their correlation patterns and forecast future output power accordingly. Frequently used statistical methods encompass the Auto-Regression and Moving Average (ARMA) model, along with its variations [9–11], and the Autoregressive Integrated Moving Average (ARIMA) model [12,13]. Machine learning approaches mainly include models like Support Vector Machines (SVMs) [14], Artificial Neural Networks (ANNs) [15], and Extreme Learning Machines (ELMs) [16]. Among these, the Back Propagation (BP) neural network has gained significant attention and has been widely applied due to its flexibility and adaptability in handling nonlinear relationships [17,18]. Traditional Artificial Neural Networks face challenges in processing large amounts of input data, including issues such as gradient vanishing and explosion. Consequently, deep learning, encompassing methods like Convolutional Neural Networks and Long Short-Term Memory networks, has garnered significant attention from researchers [19,20]. These models demonstrate superior capabilities for feature learning and information extraction, achieving notable success in PV power forecasting.

While many methods developed for centralized PV plants are applicable to distributed PV power forecasting, the latter presents unique challenges. DPV plants are typically small-scale and geographically dispersed, located across diverse environments such as cities, villages, and rooftops. This distribution leads to significant spatial and temporal variability in weather, light conditions, temperature, and other environmental factors. Enhancing the accuracy of DPV power forecasting by accounting for spatiotemporal correlations between PV power plants is, therefore, a critical research focus [21–26]. Literature reference [21] utilized correlation information to improve the accuracy of power forecasting and used the power of neighboring PV sites as model inputs to forecast the power of the target station. Literature reference [22] employed Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to extract spatial and temporal features, respectively. The temporal and spatial features were then fused using LSTMs to effectively mine the spatiotemporal correlation characteristics of DPV systems. Literature reference [23] initially linked highly correlated PV power stations to construct a topological graph structure. Based on this graph, an improved spatiotemporal graph network model was constructed to comprehensively explore the spatiotemporal characteristics of regional PV power plants.

Literature reference [24] proposed a short-term regional DPV power generation forecasting method based on partitioning that considers spatiotemporal correlations. Initially, it employs Graph Convolutional Networks (GCNs) to extract spatial correlation features and LSTM networks to capture the evolution characteristics of these dynamic spatial correlations, thereby establishing power forecasting models for power stations under different weather types. Literature reference [25] applies Dynamic Directed Graph Convolutional Networks (DDGCNs) to ultra-short-term power forecasting for regional DPV systems. To capture the dynamic and directed adjacent relationships between graph nodes, a temporal attention mechanism is introduced and combined with the directed GCN model. This approach allows for considering the dynamic relationships between DPV sites.

However, existing methods for spatiotemporal correlation rely on historical operational data from all sites, achieved through centralized data sharing to construct the model. However, such a centralized data-sharing model significantly increases the risk of data leakage. Spatiotemporal correlation mining typically involves collecting historical data from multiple sites, such as power output, meteorological conditions, and equipment operation status, potentially exposing sensitive operational details of each PV site. If stored or shared centrally without synchronization or robust encryption, these data are vulnerable to privacy breaches. Additionally, distributed PV sites often belong to different electricity sellers, who may refuse to share original PV power data to safeguard their commercial interests.

Federated learning has emerged as an effective approach to address this challenge [27]. In this framework, the historical operational data of each distributed PV system are processed locally without centralized sharing or exposure to other participants. The learning model is trained locally using historical data, and only the model parameters are shared for global aggregation on a central server [28,29]. However, the shared model parameters contain limited information, making it challenging to extract spatial correlation among distributed PV sites. While existing federated learning frameworks effectively safeguard data privacy, they struggle to significantly enhance power forecasting accuracy. Therefore, an important research gap is how to improve the model's ability to mine and utilize the spatiotemporal correlation information between distributed PV power stations while ensuring data privacy under the existing federated learning framework.

Additionally, the current distributed PV power forecasting predominantly relies on traditional point forecasting methods, which yield a single deterministic value for the output power. However, distributed PV power is influenced by various instability factors, such as weather conditions, introducing significant stochasticity and uncertainty. The probabilistic forecasting of distributed PV enables the effective quantification of forecast uncertainty, offering more comprehensive forecasting information than point forecasting and providing critical guidance for the operation and regulation of active distribution networks [30,31]. Depending on the modeling approach, probabilistic forecasting methods can be categorized into parametric and nonparametric methods [32].

The parametric method presumes that the forecasting probability density function adheres to a known distribution model, such as normal or gamma distributions. It estimates the parameters of this model based on the forecasting errors, which are subsequently utilized for probabilistic forecasting [33,34]. The advantage of parametric estimation lies in leveraging the distributional form of existing data for inference, enabling the derivation of more accurate parameter estimates. However, this approach requires strict assumptions about the data's distribution. When the data deviate from the assumed distribution, significant errors can occur. In practice, PV output errors often fail to conform to such strict distributional assumptions.

In contrast, nonparametric methods do not rely on a priori assumptions but directly model and estimate the shape of the power distribution, offering high flexibility for captur-

ing various forms of data distributions [35]. Literature reference [36] proposed constructing a nonparametric PV power probabilistic forecasting model, where independent LSTM deterministic forecasting models were developed using historical PV output data and NWP data. Nonparametric probabilistic forecasting was achieved through quantile regression averaging (QRA). Similarly, literature reference [37] effectively conducted the nonparametric probabilistic forecasting of power generation by employing a direct quantile regression method that integrates Extreme Learning Machines and quantile regression, demonstrating strong forecasting performance. Leveraging the advantages of nonparametric methods, this study recognizes that PV power fluctuations strongly influence forecasting errors: violent fluctuations tend to yield larger errors, while gentler fluctuations correspond to smaller errors. Consequently, forecasting errors do not necessarily follow an independent and identically distributed assumption [38], as their distribution law strongly correlates with the characteristics of PV power fluctuations. However, existing nonparametric probabilistic forecasting methods predominantly focus on improving forecasting models while often overlooking data dependency relationships, resulting in insufficient refinement of the error distribution characteristics.

In summary, existing DPV cluster power generation forecasts methods have three significant research gaps. Firstly, they face severe data security issues when mining and utilizing cluster-related information. Secondly, there is a lack of effective means for mining and utilizing associated features under the framework of federated learning. Thirdly, they seldom give probability forecasting results to quantify the uncertainty, and most of them ignore the correlation between forecasting results and forecasting errors.

To address these issues, this paper proposes a privacy-preserving deep federated learning method for probabilistic forecasting of ultra-short-term power generation from DPV systems and designs a collaborative feature federated learning framework. For the central server, it aggregates global models based on local model parameters and installed capacity uploaded by each station. At the same time, it gathers temporal features from various clients to generate global spatial features. For local client sites, a Transformer-based autoencoder is used to extract local time-series features. The encoder couples local time series features with global spatial features to generate spatiotemporal correlation features, which are adaptively weighted through an attention mechanism to focus the model on more critical features for forecasting. Based on point forecasting results, a joint probability distribution of forecasting values and errors is constructed, and the upper and lower bounds of errors are determined through inverse transformation, thereby obtaining the interval for probabilistic forecasting. Thus, under the premise of ensuring privacy, the dual-layer information interaction mechanism of model interaction and feature sharing is utilized to construct spatiotemporal correlation features that combine local client forecasting models, fully considering the spatiotemporal correlation relationships between different power stations to improve the forecasting accuracy of local models. By considering the dependencies between data, the distribution patterns of errors under different forecasting outcomes are refined, thus enhancing the accuracy of probabilistic forecasting.

The main contributions of this paper are as follows:

(1) In existing federated learning methods, information exchange between PV sites during the aggregation phase at the central server mostly relies on the aggregation of model parameters. This approach carries a limited amount of information, leading to insufficient shared information and failing to improve forecasting accuracy. To address the issue of low forecasting accuracy caused by inadequate information sharing in traditional federated learning methods, this paper introduces the construction of global features to fully compensate for this deficiency. By aggregating local time-series features from each site through the central server, the volume of information exchanged between sites is

increased. Under the premise of protecting data privacy, this method thoroughly considers the correlations between different sites, thereby improving power forecasting accuracy;

(2) When performing power forecasting at local stations, existing federated learning methods predominantly rely on the global model distributed by the central server, failing to account for the unique time-series characteristics of local sites. To improve this situation, this paper proposes that when conducting power forecasting at local PV sites, the Transformer model should be used to extract key time-series features specific to the local site. These extracted features are then combined with the global features issued by the central server to form spatiotemporal correlation features. On the basis of uncovering spatiotemporal correlations among sites, this approach fully considers the unique time-series characteristics of each site, further enhancing forecasting accuracy;

(3) Traditional probabilistic forecasting methods often focus solely on improving model performance while frequently overlooking the uncertainty in forecasting and lacking studies on the correlation between the distribution of forecasting errors and the fluctuation characteristics of PV output. To address this gap, this paper proposes constructing a joint probability distribution between these two aspects, modeling entirely based on the dependent relationship between forecasting values and forecasting errors. This approach fully considers the correlations within the data. Through comparisons with existing probabilistic forecasting methods, the superiority of the proposed joint distribution model has been verified.

Compared to previous recently published articles, the main differences of this paper are as shown in Table 1.

Table 1. The distinctions between this paper and other articles.

Point Forecasting Research Content			
Content of the Article	**This Paper**	**Articles [21–26]**	**Articles [27–29]**
Considering the spatiotemporal correlations between power stations	√	√	×
Considering data privacy protection	√	×	√
Effectively mining spatiotemporal correlation information while protecting data privacy	√	×	×
Considering both spatiotemporal correlation features and local time-series characteristics	√	×	×
Probabilistic forecasting research content			
Content of the article	**This paper**	**Articles [33,34]**	**Articles [35–38]**
Flexibly capturing the distribution characteristics of data	√	×	√
Considering the dependencies between data	√	×	×

2. Materials and Methods

In this paper, a privacy-preserving DPV ultra-short-term power forecasting method based on deep federated learning and probabilistic modeling is proposed, as shown in Figure 1. The proposed method consists of two main phases: point forecasting and probabilistic forecasting. In the point forecasting phase, this paper introduces a federated learning method for the cooperative processing of temporal features between the central server and the client. At the central server, two main tasks are performed. Firstly, a global model is generated using a weighted aggregation based on power station capacity. Secondly, local time-series features from individual sites are integrated via Independent Component Analysis (ICA) to generate global features that contain spatial information. These global

models and features are then distributed to each client site to enable the targeted training of client models. For the client-side, a local Transformer-based forecasting model is developed, incorporating spatiotemporal correlation features. This model uses an encoder to extract local time-series features, which are then fused with global features to form spatiotemporal correlation features. By leveraging the high correlation of historical power in neighboring time periods, key spatiotemporal features impacting the forecasted power are extracted, thereby enhancing forecasting accuracy. In the probabilistic forecasting phase, a joint probability distribution model is constructed based on the point forecasting value and forecasting error. The upper and lower boundaries of the forecasting error are obtained through inverse transformation, forming the forecasting interval. Additionally, the error distribution is further analyzed based on the dependence between the forecasting value and the forecasting error. This enables the probabilistic forecasting of DPV power.
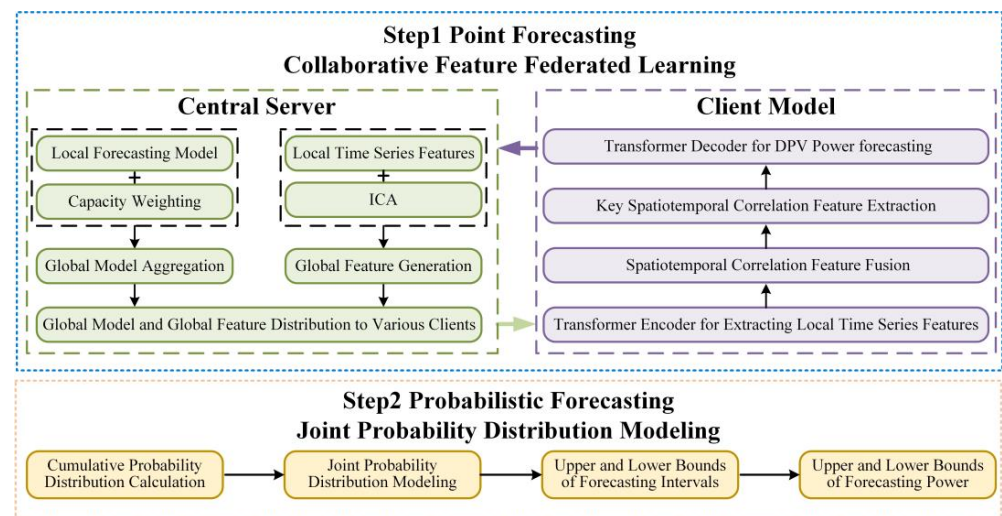

Figure 1. The flowchart of the proposed method.

2.1. Federated Learning Method for Collaborative Feature Processing

Most current federated learning frameworks rely on aggregating model parameters. However, the limited information carried by model parameters restricts their ability to enhance the accuracy of power forecasting.

2.1.1. Global Model Aggregation

Federated learning employs a decentralized training paradigm, wherein individual sites maintain data privacy by constructing a shared model through aggregating locally trained models, catering to specific user forecasting needs. However, this approach faces two main challenges: firstly, due to varying scales, different sites have differing impacts on overall error; secondly, the global model obtained is essentially a generalized model, and its direct application to forecasting at specific client sites can result in decreased forecasting performance due to the insufficient capture of the unique power characteristics of those sites. To address the aforementioned issues, two optimization measures are proposed. Firstly, during the global model aggregation phase, this paper only aggregates the parameters of shallow networks that are responsible for extracting generalized features, taking into full account the impact of capacity differences across sites during the aggregation process. Secondly, in the local client model training phase, the parameters of the shallow network are kept unchanged, while focusing on fine-tuning the parameters of deep networks that are responsible for extracting personalized features using local data. Through this approach, not only is the model's ability to effectively leverage cross-site knowledge ensured, but

also its adaptability to specific application scenarios is enhanced. The specific process is described below:

(1) The initial setup of the global model: The initial global model F^{Cen} is established through random initialization or pretraining using publicly available datasets.

(2) Client-side model training: The central server distributes the initialized global model to individual client sites. During the personalized training phase, the shallow network structure remains fixed, with emphasis placed on training and optimizing the deep network to improve the model's capacity for recognizing local characteristics. This allows each client site to perform personalized training on the global model using its local data, resulting in a forecast model tailored to the specific environment. The model forecasting process is presented in Equation (1), while the shallow network parameter update process is described in Equation (2).

$$[\hat{p}_{t+1}^n, \hat{p}_t^n, \ldots, \hat{p}_{t+\tau}^n] = f^n([p_{t-\sigma}^n, \ldots, p_{t-1}^n, p_t^n], \theta^{n,s}, \theta^{n,d}) \tag{1}$$

$$\theta^{n,d} \leftarrow \theta^{n,d} - \gamma \frac{\partial l^n([\hat{p}_{t+1}^n, \hat{p}_t^n, \ldots, \hat{p}_{t+\tau}^n] - [p_{t+1}^n, p_t^n, \ldots, p_{t+\tau}^n])}{\partial \theta^{n,d}} \tag{2}$$

where p_t^n and $\hat{p}_t^n$ are the forecasting and true power values of station n at time t, $f^n(\cdot)$ and $l^n(\cdot)$ are the forecasting model and loss function, $\theta^{n,s}$ and $\theta^{n,d}$ are the f^n shallow and deep network parameters, and γ is the learning rate;

(3) Global model aggregation and update: Contributions to overall forecast accuracy vary across sites due to differences in installed capacities at each location. Thus, the global model aggregation update aims to minimize overall loss, with weights assigned based on each site's capacity. This process iterates step (2) and step (3) until the model converges or reaches a predefined maximum number of iterations, producing an optimal global forecast model, as represented in Equations (3) and (4):

$$\begin{aligned} &\min_\theta L(\theta^s) = \sum_{n=1}^N \frac{C^n}{C} l^n(\theta^{n,s}, \theta^{n,d}) \\ &s.t. l^n(\theta^{n,s}, \theta^{n,d}) = \frac{1}{T^n} \sum_{t=1}^{T^n} (\hat{p}_t^n - \hat{p}_t^n)^2 \end{aligned} \tag{3}$$

$$\theta^s \leftarrow \theta^s - \gamma \frac{\partial L(\theta^s)}{\partial \theta^s} \tag{4}$$

where C and C^n are the total installed capacity of the N stations and the installed capacity of station n, respectively; T^n is the length of the training samples of station n; θ^s is the global model parameter; and $L(\cdot)$ and $l_n(\cdot)$ are the overall and the loss function of station n, respectively.

2.1.2. Global Feature Generation

Given the significant spatiotemporal correlation among DPV power plants, the temporal features of individual sites can serve as auxiliary information for neighboring sites, thereby enhancing the accuracy of power forecasting. Additionally, the features of individual stations are extracted post-model processing and cannot be easily back-extrapolated to the original power values. This ensures that sharing client-site temporal features enhances forecast accuracy while safeguarding the privacy of the original data. However, directly transmitting all temporal features from all power stations to local sites as inputs to the forecasting model, without preprocessing, may lead to feature redundancy. This not only increases the computational load on the local model but also degrades its performance. To address this, this paper introduces a federated learning approach leveraging global feature generation. Temporal features from local clients are first aggregated at the central server

using ICA to produce global features. This step extracts spatial correlation information, reduces feature dimensionality, and efficiently utilizes critical information. Subsequently, each local client receives the global features provided by the central server and integrates them with local time-series features to derive spatiotemporal correlation features. These are then utilized for local power forecasting.

During global feature generation, the significant fluctuations in DPV power output and the complex nonlinear correlations among PV systems in the same region are considered. ICA effectively separates statistically independent signal sources from multidimensional data, reducing dimensionality while preserving the nonlinear characteristics. By applying ICA to the multidimensional dataset composed of time-series features from each distributed PV site, global features capturing the spatial correlations among PV systems are generated, enhancing the subsequent power forecasting task.

For N DPV plants, each plant has time-series data at T time points, and the data at each time point can be represented as a D-dimensional vector (comprising power data, meteorological factors, etc.). These data are organized into a $T \times N \bullet D$ matrix $X = [x_1, x_2 \dots x_n]$, where each column corresponds to the time-series data of a single plant, and each row represents the data for all plants at a specific time point. The goal of ICA is to recover the potential independent components S from the data matrix X, that is, to identify a mixing matrix A and an independent component matrix S such that

$$X \approx AS \tag{5}$$

where S is the unobserved independent component matrix and A is the unknown mixing matrix.

In this paper, the Fast ICA algorithm is used to find the independent component S by maximizing the likelihood estimation, which is calculated as follows:

(1) Firstly, the data are centralized by subtracting the meaning:

$$\widetilde{X} = X - \frac{1}{T}\sum_{t=1}^{T} X_t \tag{6}$$

where $\widetilde{X}$ represents the data matrix after centralization, X denotes the original data matrix, and $\frac{1}{T}\sum_{t=1}^{T} X_t$ indicates the calculation of the mean values for the data in that column;

(2) Whitening the data using Principal Component Analysis (PCA) results in its covariance matrix C being an identity matrix, as shown in Equations (7)–(10):

$$\widetilde{X}_{white} = W\widetilde{X} \tag{7}$$

$$C = \frac{1}{T}\widetilde{X}\widetilde{X}^T \tag{8}$$

$$C = U\Lambda U^T \tag{9}$$

$$W = U\Lambda^{-\frac{1}{2}}U^T \tag{10}$$

where W is the whitening matrix, which can be obtained by calculating the covariance matrix C of $\widetilde{X}$ and performing eigenvalue decomposition on it;

(3) A random vector w_0 is selected as the initial weight vector, and a fixed-point algorithm is employed to perform iterations until w_i converges. These steps are repeated to determine the independent component $S = [s_1, s_2, \dots s_k]$ representing the global feature, as shown in Equations (11)–(13):

$$w_{i+1} = E_1\left[\widetilde{X}_{white}g(w_i\widetilde{X}_{white})\right] - E_2\left[g'(w_i\widetilde{X}_{white})\right]w_i \tag{11}$$

$$w_{i+1} - w_i < \varepsilon \tag{12}$$

$$s_i = w_i \widetilde{X}_{white} \tag{13}$$

where g is the nonlinear activation function, E_1 denotes the average of the element-by-element product, and E_2 denotes the computational average.

2.1.3. Local Client Transformer Self-Encoder Forecasting Modeling

Relying on the above federated learning framework, a local client model is constructed for power forecasting at each site based on spatiotemporal correlation features formed by coupling local temporal features and global features. To avoid the issue of feature redundancy when the decoder processes spatiotemporal correlation features, a Transformer model with an attention mechanism is introduced into the autoencoder power forecasting model, enabling the model to focus on the global features that are more critical for output power. The model structure is shown in Figure 2. The autoencoder model in this paper is primarily based on a Gate Recurrent Unit (GRU) model and a Fully Connected Neural Network (FC) connected in series as the encoder and decoder. Among them, GRU serves as a shallow network for extracting global generalized features, and FC acts as a deep network for further extracting personalized features of local client sites. The specific forecasting steps are as follows:
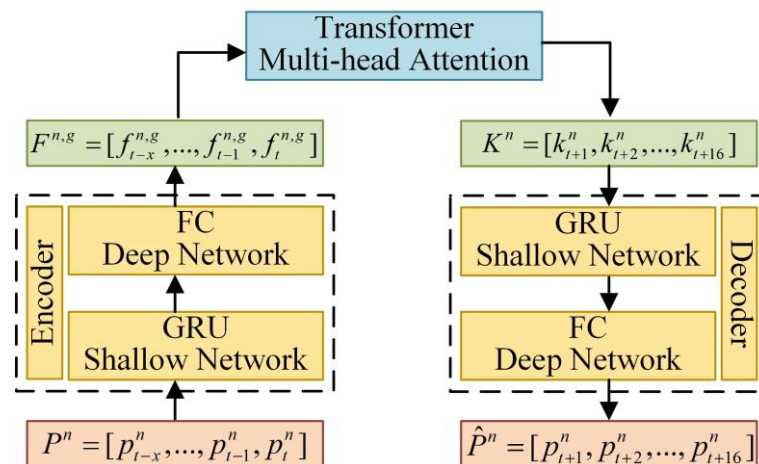

Figure 2. Autoencoder model structure.

(1) At each PV site, the ultra-short-term power forecast is executed at the moment t. Initially, the historical power sequence $P^n = [p_{t-x}^n, \ldots, p_{t-1}^n, p_t^n]$ from the preceding x moments is input, and the encoder F computes the feature vector at moment t based on the power at moment i and the feature from moment $i-1$, as shown in Equation (14):

$$f_i^n = f^{Enc}(p_i^n, f_{i-1}^n) \tag{14}$$

Subsequently, the temporal features are coupled with the global features S to obtain temporal correlation features $F^{n,g} = [f^n; S]$, which are used as decoder inputs;

(2) During the forecast execution at the moment t, to obtain input features that are critical for forecasting output power at target moments $t + 1$ to $t + 16$, it is assumed that the power remains relatively stable over short periods. The decoder output at moment $t - 1$, specifically, the forecasting power values from t to $t + 15$ (denoted as $\hat{P}_{t:t+15}^n = [\hat{p}_t^n, \hat{p}_{t+1}^n, \ldots, \hat{p}_{t+15}^n]$), is used as the query vector. The spatiotemporal correlation features serve as the key vector. These vectors are utilized in training to compute the

attention weights a_{t+1} during the forecast execution at the moment of *t*, as illustrated in Equations (15) and (16).

$$e_{t+1,i} = v_s^\top \tanh(w_s F_i^{n,g} + u_s \hat{P}_{t:t+15}^n + b_s) \tag{15}$$

$$a_{t+1,i} = \frac{\exp(e_{t+1,i})}{\sum\limits_{i=t-\sigma}^{t} \exp(e_{t+1,i})} \tag{16}$$

where $e_{t+1,i}$ is the attention score, w_s, u_s, b_s, and v_s are trainable parameters, $\tanh(\cdot)$ is the activation function, and $\exp(\cdot)$ is the power exponential function.

Subsequently, the attention weight matrix is weighed and summed with the key value matrix to obtain the final attention representation k_{t+1}^n, as shown in Equation (17). This is used as the decoder input sequence to further improve the forecast accuracy.

$$k_{t+1}^n = a_{t+1} F_{t-x:t}^{n,g} \tag{17}$$

where $a_{t+1,i}$ is the attention weight, $a_{t+1,i} \in a_{t+1}$, and $\exp(\cdot)$ are exponential functions with e as the base;

(3) Using the decoder model $f^{Dec}(\cdot)$, the power forecasts for future time periods $t + 1 \sim t + 16$ can be calculated, as shown in Equation (18).

$$\hat{p}_{t+\tau}^n = f^{Dec}(\hat{p}_{t+\tau-1}^n, k_{t+\tau}^n) \tag{18}$$

2.2. Joint Probability Distribution Modeling

Current research on probabilistic forecasting of PV power output often overlooks the uncertainty in forecasting and lacks studies on the correlation between the distribution of forecasting errors and the fluctuation characteristics of PV output. To address these issues, this paper proposes a probabilistic forecasting method for ultra-short-term DPV power based on joint distribution modeling. This method integrates joint distribution models with nonparametric methods, fully modeling based on the dependent relationship between forecast values and forecasting errors. It can effectively describe the dependency structure between forecasting values and forecasting errors, thereby more accurately quantifying the uncertainty in forecasting. The Copula function is a mathematical function used to establish dependent relationships between multidimensional random variables. It offers a flexible approach to describing the correlation between variables without presupposing the form of their marginal distributions, making it especially suitable for uncovering correlations in nonlinear data. Therefore, this paper treats the point forecasting results and forecasting errors of PV power as two random variables, using a Copula function to fit the deterministic point forecasting and forecasting errors. This allows the width of the forecasting intervals to better accommodate the characteristics of the forecasting error distribution, introducing uncertainty information into the forecasting process, thereby enhancing the accuracy of probabilistic forecasting for PV power generation. The specific steps for forming the probabilistic forecast interval are as follows:

(1) The joint probability distribution between the deterministic point forecasting result $\hat{P}$ and the forecasting error *e* is constructed based on the Copula function, and the joint distribution function *F* (*p,e*) is shown in Equation (19) for the marginal cumulative probability distribution functions $F_P(p)$ and $F_E(e)$ for the point forecasting value and forecasting error:

$$F(p,e) = C_{P,E}[F_P(p), F_E(e)] \tag{19}$$

where the functions $C_{P,E}$ are called the Copula probability distribution function of the point forecasting values and forecasting errors;

(2) After obtaining the joint probability distribution of point forecasting value and forecasting error, the conditional probability distribution function of the cumulative probability distribution of forecasting error is obtained by taking the value according to the cumulative probability distribution of the point forecasting value, as in Equation (20):

$$F(F_E(e)|F_P(p);\theta) = \frac{\partial C_{P,E}}{\partial F_P(p)} \tag{20}$$

where θ is the main parameter of the Copula function;

(3) Based on the desired confidence level, the corresponding forecast interval is obtained. Let the confidence probability of the power curve be α, such that the α of the data falls within the probability interval. Let $\beta = 1 - \alpha$. Let the asymmetry coefficient of the confidence interval be k. The quantile probabilities β_1, β_2 of the upper and lower boundaries of the confidence interval indicate that the probability of a data point being higher than the upper boundary is β_1 and that the probability of a data point being lower than the lower boundary is β_2, as shown in Equations (21) and (22).

$$\beta_1 = (1 - k)\beta \tag{21}$$

$$\beta_2 = k\beta \tag{22}$$

(4) Based on the obtained upper and lower boundary quantile probabilities and the corresponding quantile probability of the point forecast value p_i, the conditional probability distribution function of the cumulative probability distribution of the forecast error is utilized to calculate the quantiles F_{1i}, F_{2i} corresponding to β_1, β_2, as in Equations (23)–(25).

$$F_{1i} = C_{(p_i)}^{-1}(\beta_1) \tag{23}$$

$$F_{2i} = C_{(p_i)}^{-1}(\beta_2) \tag{24}$$

$$C_{(p_i)}(F(e_i)) = \frac{\partial C_{P,E}(F_E(e), F_P(p) = p_i)}{\partial F_P(p)} \tag{25}$$

(5) According to the obtained upper and lower boundaries of the corresponding quartiles F_{1i}, F_{2i}, the corresponding upper and lower boundaries of the forecast error values e_{1i}, e_{2i} are obtained by using the inverse of the cumulative probability distribution function of the forecast error $F_E(e)$, as in Equations (26) and (27).

$$e_{1i} = F_E^{-1}(F_{1i}) \tag{26}$$

$$e_{2i} = F_E^{-1}(F_{2i}) \tag{27}$$

The upper and lower boundary values of the error are correspondingly superimposed onto the corresponding forecasting value $\hat{p}$ to obtain the upper and lower boundaries of the forecasting interval p_{1i}, p_{2i}, and realize the DPV power probability forecasting.

3. Results and Discussion

3.1. Dataset Description

The experimental data originate from the power generation data of 30 DPV stations located in a certain area in northern China. The data span from 1 July 2023 to 1 July 2024, with a time resolution of 15 min. To ensure data quality and reduce the impact of outliers on power forecasting performance, the original data underwent cleaning and

repair. Specifically, abnormally large values exceeding capacity, abnormally small values such as negative outputs, and missing power values were corrected. Data repair was achieved by replacing the abnormal points with the average power values of the three normal points before and after them. Subsequently, the cleaned dataset was divided into training, validation, and test sets at a ratio of 8:1:1, respectively used for deterministic point forecasting training, constructing a joint distribution model using validation set data, and testing the effectiveness of the proposed method with test set data. During the forecasting process, all power values less than zero were set to zero.

3.2. Evaluation Metrics

During the point forecasting phase, this paper adopts the Normalized Root Mean Square Error (NRMSE) and Accuracy (ACC) as evaluation metrics, as shown in Equations (28) and (29). NRMSE considers the relative error between forecasting values and actual values, while ACC evaluates the model's performance over the entire forecasting period by comprehensively considering the 16-step results of forecasting, thus avoiding the limitation of assessing model performance based solely on single-point forecasting results.

$$NRMSE = 1 - \frac{1}{p^{\max}} \sqrt{\frac{1}{T} \sum_{t=1}^{T} (p_t - \hat{p}_t)^2} \times 100\%$$ (28)

$$ACC = \frac{\sum_{t=1}^{n} \left(1 - \frac{\sqrt{\frac{\sum_{i=1}^{16} \left[(p_i - \hat{p}_i)^2 \cdot \frac{|p_i - \hat{p}_i|}{\sum_{t=1}^{T} |p_i - \hat{p}_i|}\right]}}{p^{\max}}\right)}{n} \times 100\%$$ (29)

where $p^{\max}$ represents the maximum power, T indicates the length of the forecasting time, $\hat{p}_t$ and p_t are the forecasted and true values at time t, and $\hat{p}_i$ and p_i are the forecasted and true values at the *ith* step of forecasting at time t, with n being the number of forecasting points evaluated for that day.

During the probabilistic forecasting phase, the Prediction Interval Coverage Probability (PICP), Prediction Interval Normalized Average Width (PINAW), and Skill Score (SS) are adopted as evaluation metrics, as shown in Equations (30)–(32). PICP represents the percentage of actual values that fall within the forecasting intervals, reflecting the accuracy of the model. PINAW measures the width of the forecasting intervals, indicating the sensitivity of the model. SS provides a comprehensive evaluation, rewarding narrower intervals but penalizing if the target value falls outside the forecasting interval. The value of this metric is non-positive, and, the closer it is to 0, the better the probabilistic forecasting performance.

$$PICP = \frac{1}{N} \sum_{i=1}^{N} C_i, C_i = \begin{cases} 1, \ p \in [p_{2i}, p_{1i}] \\ 0, \ p \notin [p_{2i}, p_{1i}] \end{cases}$$ (30)

$$PINAW = \frac{1}{N} \sum_{i=1}^{N} (p_{1i} - p_{2i})$$ (31)

$$SS = \frac{1}{N} \sum_{i=1}^{N} S_i, \ S_i = \begin{cases} -2\alpha(p_{1i} - p_{2i}) - 4(p - p_{1i}), \ p > p_{1i} \\ -2\alpha(p_{1i} - p_{2i}), \qquad\qquad p \in [p_{2i}, p_{1i}] \\ -2\alpha(p_{1i} - p_{2i}) - 4(p_{2i} - p), \ p < p_{2i} \end{cases}$$ (32)

where N represents the sample size. p_{1i} and p_{2i} represent the upper and lower boundaries of the forecasted samples.

3.3. Experimental Setup

To demonstrate the effectiveness of the proposed method in this paper, four additional models are set up as control models in addition to the method proposed in this paper:

Proposed method: A joint distribution probabilistic forecasting model based on collaborative feature federated learning and a Transformer autoencoder;

Method 1: A joint distribution probabilistic forecasting model based on a traditional federated learning framework and a Transformer autoencoder model;

Method 2: A joint distribution probabilistic forecasting model using a Transformer autoencoder built solely on local data for each station;

Method 3: A joint distribution probabilistic forecasting model using a GRU built solely on local data for each station;

Method 4: A traditional probabilistic forecasting model based on collaborative feature federated learning and a Transformer autoencoder.

In the settings of these methods, Method 1 involves only the exchange of model parameter information without aggregating global features. Methods 2 and 3, while protecting the data privacy of clients, sacrifice inter-station information sharing and do not fully utilize the spatiotemporal correlations between stations. Method 4 considers only the distribution patterns of forecasting errors without accounting for the dependencies between forecasting values and forecasting errors. By comparing the proposed method with Methods 1 and 2, we verify that the proposed method achieves more accurate forecasting results while ensuring data privacy. Comparing Method 2 with Method 3 verifies the forecast performance of the local client Transformer autoencoder model. Comparing the proposed method with Method 4 validates the accuracy of probabilistic forecasting based on joint probability distribution. The input to each forecasting model is the power from the previous 96 time steps of the forecast period, and the output is the power for the next 16 time steps, achieving ultra-short-term power forecasting.

In the point forecasting phase, this paper employs grid search to optimize the selection of hyperparameters (learning rate, batch size, number of iterations) for the forecasting model. Initially, based on the scale of the experimental data and the complexity of the model, three different values are set for each hyperparameter. Then, by means of a grid search, the method seeks the optimal combination of hyperparameters within the predefined hyperparameter space. Among all combinations of hyperparameters, the one that yields the best performance on the validation set is selected as the final hyperparameter configuration. Through comparison, this paper selects a learning rate of 0.001, a batch size of 128, and 100 epochs for the Transformer point forecasting model. For the GRU point forecasting model, a learning rate of 0.001, a batch size of 128, and 150 epochs are chosen.

3.4. Experimental Results Analysis

In the point forecasting phase, the proposed method and Methods 1, 2, and 3 are used to conduct ultra-short-term power forecasting for each client station. Since Method 4 is identical to the proposed method in the point forecasting phase, no repeated verification is performed for Method 4. Under a time scale of 1 h ahead to 4 h ahead, the average NRMSE metric values for each forecasting method across all stations are statistically analyzed, as shown in Table 2. The average ACC metric values for each forecasting method across all stations are presented in Table 3. It can be seen that the proposed method demonstrates superiority in both evaluation metrics. Compared with Methods 1 and 2, the proposed method improves the NRMSE by 0.55% and 0.41%, respectively, at the 1 h ahead time scale, and by 1.17% and 2.59%, respectively, at the 4 h ahead time scale. The ACC is improved by 1.61% and 1.74%, respectively, indicating that the collaborative feature federated learning method proposed in this paper, compared to traditional parameter-

aggregated federated learning, effectively mines the spatiotemporal correlations between different client stations by introducing their temporal features and completing a dual-layer aggregation of information with the global model, thereby enhancing forecasting accuracy while ensuring the privacy of individual station data. Method 1 shows improved accuracy over Method 2 by leveraging spatiotemporal correlation information to some extent through the aggregation of model parameters: though its forecasting results at the 1 h ahead time scale are not as good as those of Method 2, its overall forecasting performance, evaluated by the 16-step results, is better. By comparing Methods 2 and 3, the autoencoder built on the Transformer model outperforms the GRU neural network by 2.45% and 2.14% in the NRMSE metric, and by 1.42% in the ACC metric, at the 1 h ahead and 4 h ahead time scales, respectively. This demonstrates that the proposed autoencoder can effectively learn long-term dependencies within local station data and focus on more critical temporal features through the multi-head attention mechanism, acquiring more important information during training and thus achieving more accurate forecasting results compared to traditional neural networks. Furthermore, by calculating the decrease in the NRMSE metric for each method at the 1 h ahead and 4 h ahead time scales, the proposed method shows a reduction of only 2.21%, while other methods show reductions ranging from 2.83% to 4.39%. This indicates that the proposed method performs more stably in ultra-short-term 16-step forecasting, with the forecasting model better capturing trends in future data, resulting in superior forecasting performance.

Table 2. NRMSE metrics for each forecasting method.

Metrics	Time Scale	Proposed Method	Method 1	Method 2	Method 3
NRMSE	1 h	94.15%	93.60%	93.74%	91.29%
	2 h	93.63%	92.58%	91.72%	90.92%
	3 h	92.90%	91.89%	89.98%	89.02%
	4 h	91.94%	90.77%	89.35%	87.21%

Table 3. ACC metrics for each forecasting method.

Metrics	Proposed Method	Method 1	Method 2	Method 3
ACC	93.62%	92.01%	91.88%	90.46%

Taking Station No. 15 as an example, the forecasting results and forecasting errors of the four methods are shown in Figure 3. From the figures, it can be observed that the curve generated by the proposed method closely aligns with the actual values, and the overall error is relatively low. Furthermore, it is evident that the distribution characteristics of forecasting errors vary under different forecasting output results and fluctuation patterns. For example, as shown in the figure, the first day is sunny, with relatively smooth power output fluctuations, and the amplitude of forecasting errors is relatively small. The second day is cloudy, with more intense fluctuations in PV power output, leading to similarly intense fluctuations in error. The third day is overcast, with lower forecasting power values and consequently smaller error values. This indicates that there is a relationship between the distribution patterns of forecasting errors and the characteristics of power output fluctuations. Moreover, the distribution patterns of forecasting errors differ from those of forecasting values under different weather conditions. This insight also provides a basis for constructing joint probability distributions in subsequent probabilistic forecasting research.
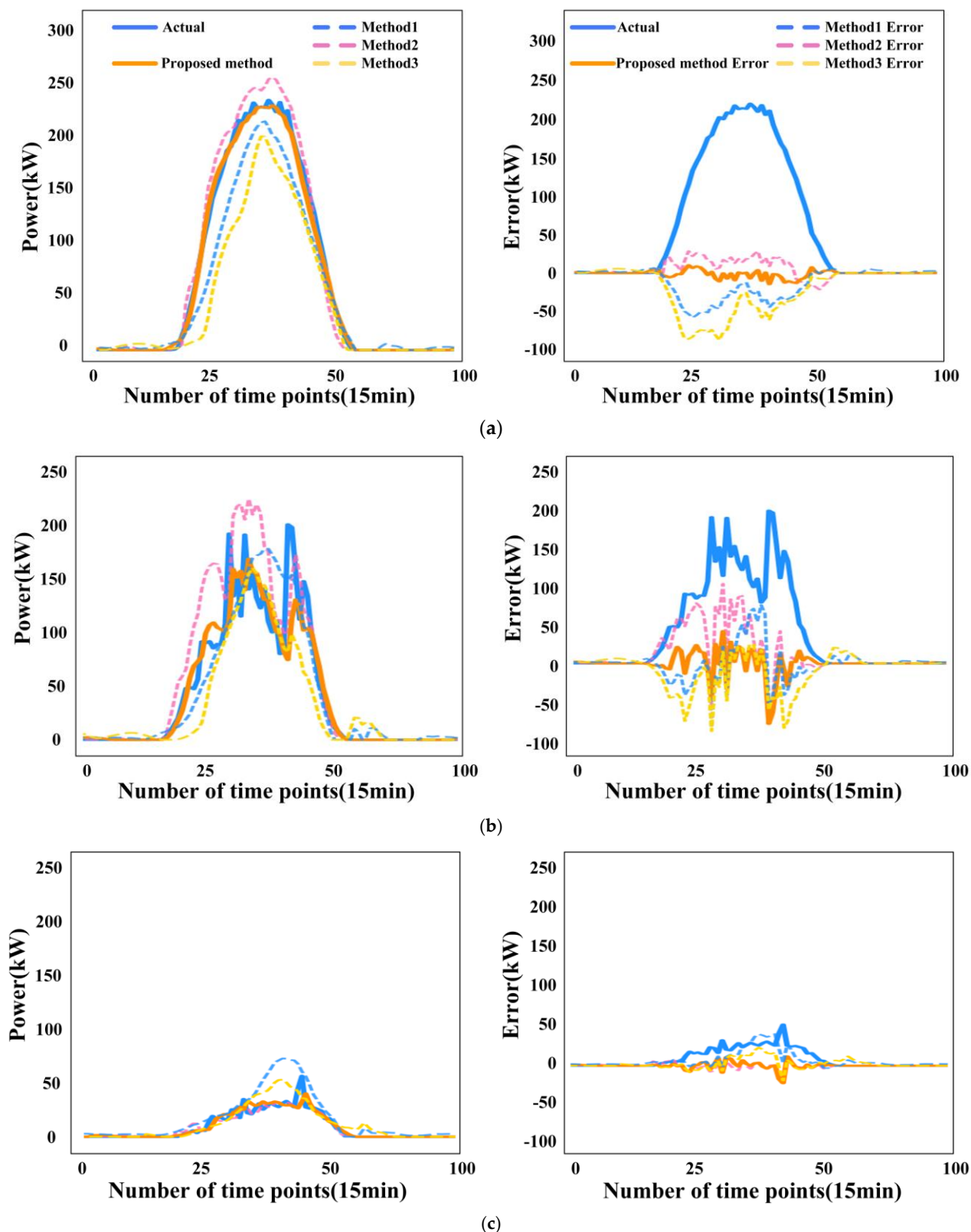
Figure 3. Point forecasting results and forecasting errors under three different weather conditions (the experimental results of Method 4 are identical to those of the proposed method). (**a**) Sunny; (**b**) Cloudy; (**c**) Overcast.

Based on the above discussion, to further demonstrate the effectiveness of the proposed method in this paper, the average values of forecasting evaluation metrics for each station under a 4 h ahead time scale were statistically analyzed for three typical weather conditions,

as shown in Table 4. Through comparison, it is evident that the proposed method in this paper achieves good forecasting results in all typical weather scenarios. Under sunny conditions, the proposed method in this paper achieved an NRMSE of 93.57%, which is 1.15%, 2.61%, and 1.56% higher than Method 1, Method 2, and Method 3, respectively. In terms of the ACC metric, our method reached 94.88%, outperforming the other methods by 1.76%, 2.86%, and 1.71%, respectively. This indicates that the proposed method offers higher forecasting accuracy and stability under sunny conditions. The analysis shows that, due to the relatively regular changes in irradiance and stable power output from DPV systems under clear skies, all methods achieve good forecasting results, with our proposed method being the most effective. In overcast conditions, the proposed method's NRMSE was up to 4.01% better than other methods, and its ACC was 3.93% higher. Under cloudy conditions, the proposed method outperformed others by up to 3.12% and 2.7% in the two evaluation metrics, respectively. Through comparative analysis, it is evident that the rapid movement and changes of cloud cover in overcast and cloudy weather lead to highly irregular patterns of solar irradiance. Solar irradiance can drop sharply from high levels to low within a few minutes, causing significant fluctuations in PV system output power over short periods. This unstable irradiance condition makes it difficult for traditional forecasting models to accurately capture the trends of PV power output, leading to decreased forecasting accuracy, especially on short time scales (such as very short-term forecasting). The proposed method addresses these challenges by considering the spatiotemporal correlations between this station and nearby stations, obtaining more information about irradiance and power output changes. Stations at different geographic locations may experience similar weather patterns but with certain lags or advances in timing. By sharing this information, we can gain a more comprehensive understanding of current and future irradiance trends, thereby improving forecasting accuracy. Furthermore, the application of the multi-head attention mechanism enhances the model's ability to focus on key feature changes, effectively capturing critical information essential for accurate forecasting. This helps mitigate the increased uncertainty brought by overcast and cloudy conditions, resulting in superior forecasting performance compared to other methods.

Table 4. Forecasting accuracy of various methods under different weather conditions.

		Proposed Method	Method 1	Method 2	Method 3
Sunny	NRMSE	93.57%	92.42%	90.96%	92.01%
	ACC	94.88%	93.12%	92.02%	93.17%
Cloudy	NRMSE	89.93%	89.67%	86.77%	85.92%
	ACC	91.85%	91.61%	88.74%	87.92%
Overcast	NRMSE	90.95%	87.83%	89.40%	87.91%
	ACC	92.96%	90.26%	91.65%	91.30%

Based on the aforementioned analysis, the proposed method demonstrates superior point forecasting performance. On one hand, the Transformer-based autoencoder model constructed in this paper possesses superior time series modeling capabilities, effectively enhancing the accuracy of power forecasting by focusing on key temporal features. On the other hand, the proposed method recognizes that information sharing among neighboring DPV stations is an effective way to improve forecasting accuracy, achieving a deeper exploration of spatiotemporal correlations through the dual-layer information interaction of global models and global temporal features. Consequently, under the premise of protecting the data privacy of local clients, the proposed method fully considers the relationships between stations, thereby improving the accuracy of point forecasting.

To further demonstrate the effectiveness of the proposed method, the ACC accuracy of different forecasting methods across all client stations was statistically analyzed, as shown in Figure 4. It is evident that the proposed method achieves the best forecasting performance in the majority of stations, indicating that, by constructing a collaborative feature federated learning framework, by establishing a more comprehensive information sharing mechanism between the central server and local clients, and by building a Transformer model for local temporal feature extraction, the proposed method not only effectively mines the spatiotemporal correlations between different client stations but also fully considers the personalized features of local sites, thereby achieving higher forecasting accuracy. Method 1, compared to Method 2, considers more of the spatiotemporal correlations between stations, and, despite sharing less information, still manages to improve forecasting accuracy. Method 2, compared to Method 3, considers the temporal characteristics of the power station itself, enabling the model to better focus on the fluctuation patterns of local power data. It should be noted that the proposed method does not yield the best forecasting results for Stations 13, 18, 25, and 26. This might be due to weaker power output correlations between these stations and others, which not only makes it difficult for the global model to effectively capture specific features of these stations during training, but also introduces a significant amount of redundant information into the local forecasting model, reducing its forecasting accuracy. However, overall, the proposed method still achieves the best forecasting results, followed by Method 1, Method 2, and Method 3. This also verifies the advantages of introducing global feature aggregation and key spatiotemporal correlation feature generation in the collaborative feature federated learning during the forecasting process.
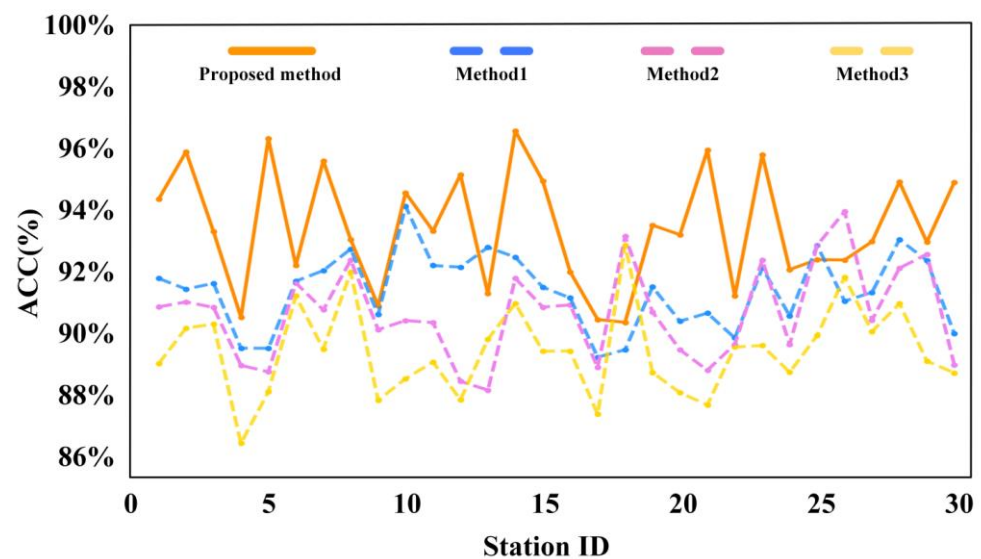

Figure 4. Comparison of ACC (accuracy) for each method across different stations (the experimental results of Method 4 are identical to those of the proposed method).

In the probabilistic forecasting phase, the proposed method and the four other methods were used to conduct ultra-short-term probabilistic forecasting for each power station. In this paper, the Clayton Copula function was selected to fit the cumulative probability distribution of deterministic point forecasting results and forecasting errors, with the primary parameter θ of the Copula function being estimated using maximum likelihood estimation.

Taking Station No. 15 as an example again, the accuracy comparisons of various probabilistic forecasting methods at different confidence levels are presented in Table 5. By comparing the experimental results of different methods, it can be seen that, under three different confidence levels, the proposed method exhibits the best SS, effectively narrowing

the forecasting intervals according to the distribution patterns of forecasting errors while maintaining high coverage, thus achieving a balance between reliability and sensitivity in probabilistic forecasting.

Table 5. The accuracy metrics of probabilistic forecasting.

	95%			90%			85%		
	PICP/%	PINAW/kW	SS/kW	PICP/%	PINAW/kW	SS/kW	PICP/%	PINAW/kW	SS/kW
Proposed method	90.63%	39.98	−54.03	87.44%	37.48	−50.69	82.90%	35.75	−46.58
Method1	88.49%	44.36	−58.93	85.33%	40.69	−55.74	80.06%	37.90	−51.08
Method2	92.82%	48.02	−65.31	88.29%	45.92	−60.22	83.92%	42.61	−57.88
Method3	81.72%	40.33	−77.20	76.02%	34.57	−71.94	73.28%	30.40	−67.32
Method4	93.05%	54.87	−69.51	90.41%	49.65	−64.32	85.91%	46.03	−61.13

Comparing the proposed method with Methods 1, 2, and 3 reveals that, based on achieving high-precision point forecasting, the proposed method also realizes high-precision probabilistic forecasting. Under the three confidence levels, the SS of the proposed method is improved by 30.01%, 29.54%, and 30.81%, respectively. Moreover, combining the results of point forecasting, it is found that higher point forecasting accuracy also leads to higher probabilistic forecasting accuracy. This indicates that the accuracy of point forecasting plays a crucial role in the precision of probabilistic forecasting. High-precision point forecasting can more accurately reflect actual values, narrow the range of error distribution, stabilize forecasting errors, and reduce the occurrence of extreme values and sharp fluctuations, aiding probabilistic forecasting models in more accurately estimating uncertainties, thereby achieving highly accurate probabilistic forecasting. Comparing the proposed method with Method 4, it is evident that probabilistic forecasting based on joint probability distribution yield better results than those considering only error distribution. Under the three confidence levels, although the proposed method reduces the PICP by 2.42%, 2.97%, and 3.01%, respectively, it effectively narrows the PINAW by 27.14%, 24.51%, and 22.33%, respectively, thereby improving the SS by 22.27%, 21.29%, and 23.80%, respectively. This suggests that joint probability distribution, by simultaneously considering the information of forecasting values and errors, captures the intrinsic structure and mutual relations of data more comprehensively. By exploring the dependency structure between forecasting values and errors, it refines the study of error distribution patterns, achieving higher-precision probabilistic forecasting.

At a 90% confidence level, the probabilistic forecasting results of each method are shown in Figure 5. Method 4 only considers the distribution characteristics of errors, so the width of its forecasting interval is entirely determined by the quantiles of error boundaries and is unaffected by changes in the forecasting values. This approach assumes that the error distribution remains constant, leading to a forecasting interval width that is a fixed constant value. While this fixed interval width is simple and intuitive, it overlooks the dynamic nature of actual power output over time. Consequently, Method 4's probabilistic forecasting lacks sensitivity to real-time power fluctuations and fails to effectively capture rapid changes in the short term, resulting in poorer predictive acuity. In contrast, the method proposed in this paper, which is based on joint probability distribution modeling, exhibits significant advantages. Our proposed method not only considers the distribution of errors but also incorporates the patterns of change in point forecasting results. This allows the width of the forecasting intervals to dynamically adjust according to actual conditions, closely aligning with the true power values. This characteristic enables the model to better reflect changes in the distribution of forecasting errors caused by differences

in power fluctuation characteristics. By introducing Copula functions, our method can flexibly describe the complex dependency relationships between PV output and forecasting errors.
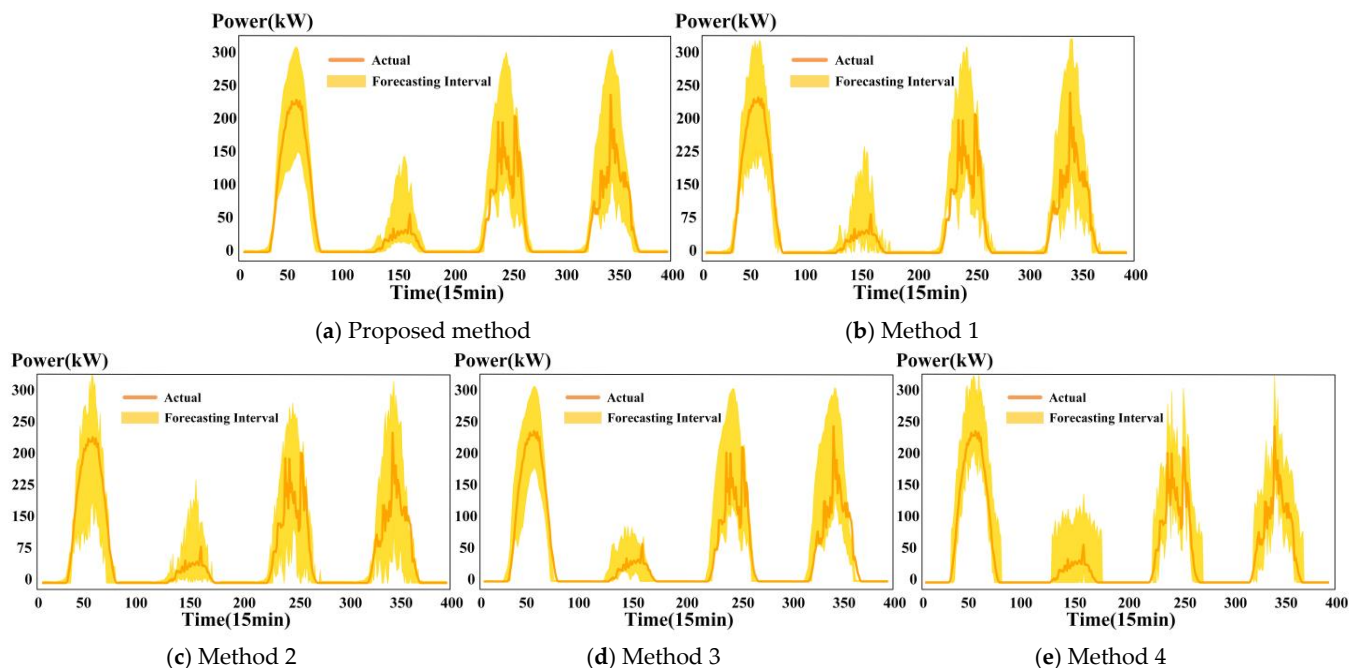

Figure 5. Probabilistic forecasting results at 90% confidence level for different methods (**a**–**e**).

In summary, the proposed method achieves dynamic adjustment of forecasting intervals through joint probability distribution modeling. This ensures both the accuracy of probabilistic forecasting and significantly enhances sensitivity to real-time power fluctuations.

4. Conclusions

To provide more comprehensive reference information for power system dispatch, this paper proposes a privacy-preserving deep federated learning method for the probabilistic forecasting of ultra-short-term power generation from DPV systems. This method achieves the enhancement of power forecasting accuracy for target stations by mining the correlation information of power output between stations while protecting the data privacy of each client. In the point forecasting phase, a collaborative feature federated learning method is proposed. Initially, the central server aggregates model parameters from each client and generates global features, which are then distributed to each station to facilitate the interaction and sharing of global and local information. Subsequently, a Transformer-based autoencoder model is established for each client to deeply mine the local temporal features of the station. This dual-layer information interaction in federated learning effectively enhances the accuracy of point forecasting, providing reliable data support for probabilistic forecasting. In the probabilistic forecasting phase, the Copula function is used to construct a joint probability distribution model between forecasting values and forecasting errors, allowing for a detailed study of error distribution patterns and obtaining more adaptive forecasting intervals to improve the sensitivity and accuracy of probabilistic forecasting.

For future research, the following improvements can be made:

(1) In this paper, the aggregation of global features is introduced during the traditional federated learning process, but no improvements have been made to the aggregation of model parameters. Therefore, future work will further explore global model aggregation methods that are more suitable for the characteristics of DPV power generation, to enhance the performance of federated learning;

(2) When constructing the joint probability distribution model in this paper, only the Clayton Copula function is used for fitting. Thus, future work will continue to investigate other types of fitting functions to find those that better fit PV data, further improving the performance of probabilistic forecasting;

(3) Future research will consider introducing a weather anomaly detection mechanism to identify extreme weather events and assess the model's performance under different weather anomalies and grid instability conditions.

References

1. Wang, Y. The analysis of the impacts of energy consumption on environment and public health in China. *Energy* **2010**, *35*, 4473–4479. [CrossRef]
2. Liu, J. China's renewable energy law and policy: A critical review. *Renew. Sustain. Energy Rev.* **2018**, *99*, 212–219. [CrossRef]
3. Shi, Y.; He, W.; Zhao, J.; Hu, A.; Pan, J.; Wang, H.; Zhu, H. Expected output calculation based on inverse distance weighting and its application in anomaly detection of distributed photovoltaic power stations. *J. Clean. Prod.* **2020**, *253*, 119965. [CrossRef]
4. Mangherini, G.; Diolaiti, V.; Bernardoni, B.; Andreoli, A.; Vincenzi, D. Review of Façade Photovoltaic Solutions for Less Energy-Hungry Buildings. *Energies* **2023**, *16*, 6901. [CrossRef]
5. Wang, F.; Ge, X.; Dong, Z.; Yan, J.; Li, K.; Xu, F.; Lu, X.; Shen, H.; Tao, P. Joint energy disaggregation of behind-the-meter PV and battery storage: A contextually supervised source separation approach. *IEEE Trans. Ind. Appl.* **2022**, *58*, 1490–1501. [CrossRef]
6. Mayer, M.J.; Gróf, G. Extensive comparison of physical models for photovoltaic power forecasting. *Appl. Energy* **2021**, *283*, 116239. [CrossRef]
7. Malinkovich, Y.; Sitbon, M.; Lineykin, S.; Dagan, K.J.; Baimel, D. A Combined Persistence and Physical Approach for Ultra-Short-Term Photovoltaic Power Forecasting Using Distributed Sensors. *Sensors* **2024**, *24*, 2866. [CrossRef] [PubMed]
8. Wang, J.; Zhong, H.; Lai, X.; Xia, Q.; Wang, Y.; Kang, C. Exploring key weather factors from analytical modeling toward improved solar power forecasting. *IEEE Trans. Smart Grid* **2019**, *10*, 1417–1427. [CrossRef]
9. Abedinia, O.; Bagheri, M.; Agelidis, V.G. Application of an adaptive Bayesian-based model for probabilistic and deterministic PV forecasting. *IET Renew. Power Gener.* **2021**, *15*, 2699–2714. [CrossRef]
10. Jiang, Y.; Zheng, L.; Ding, X. Ultra-short-term prediction of photovoltaic output based on an LSTM-ARMA combined model driven by EEMD. *J. Renew. Sustain. Energy* **2021**, *13*, 046103. [CrossRef]
11. Lu, J.; Wang, B.; Ren, H.; Zhao, D.; Wang, F.; Shafie-khah, M.; Catalao, J.P.S. Two-Tier Reactive Power and Voltage Control Strategy Based on ARMA Renewable Power Forecasting Models. *Energies* **2017**, *10*, 1518. [CrossRef]
12. Zhang, J.; Liu, Z.; Chen, T. Interval prediction of ultra-short-term photovoltaic power based on a hybrid model. *Electr. Power Syst. Res.* **2022**, *216*, 109035. [CrossRef]
13. Pieri, E.; Kyprianou, A.; Phinikarides, A.; Makrides, G.; Georghiou, G.E. Forecasting degradation rates of different photovoltaic systems using robust principal component analysis and ARIMA. *IET Renew. Power Gener.* **2017**, *11*, 1245–1252. [CrossRef]
14. Gong, B.; An, A.; Shi, Y.; Guan, H.; Jia, W.; Yang, F. An interpretable hybrid spatiotemporal fusion method for ultra-short-term photovoltaic power prediction. *Energy* **2024**, *308*, 132969. [CrossRef]
15. Meng, X.; Gao, F.; Xu, T.; Zhou, K.; Li, W.; Wu, Q. Inverter-Data-Driven Second-Level Power Forecasting for Photovoltaic Power Plant. *IRE Trans. Ind. Electron.* **2021**, *68*, 7034–7044. [CrossRef]
16. Wang, M.; Wang, P.; Zhang, T. Evidential extreme learning machine algorithm-based day-ahead photovoltaic power forecasting. *Energies* **2022**, *15*, 3882. [CrossRef]

17. Hu, Y.; Lian, W.; Han, Y.; Dai, S.; Zhu, H. A seasonal model using optimized multi-layer neural networks to forecast power output of PV plants. *Energies* **2018**, *11*, 326. [CrossRef]

18. Liu, L.; Sun, Q.; Wennersten, R.; Chen, Z. Day-Ahead Forecast of Photovoltaic Power Based on a Novel Stacking Ensemble Method. *IEEE Access* **2023**, *11*, 113593–113604. [CrossRef]

19. Nie, Y.; Sun, Y.; Chen, Y.; Orsini, R.; Brandt, A. PV power output prediction from sky images using convolutional neural network: The comparison of sky-condition-specific sub-models and an end-to-end model. *J. Renew. Sustain. Energy* **2020**, *12*, 046101. [CrossRef]

20. Li, Q.; Xu, Y.; Chew, B.S.H.; Ding, H.; Zhao, G. An Integrated Missing-Data Tolerant Model for Probabilistic PV Power Generation Forecasting. *IEEE Trans. Power Syst.* **2022**, *37*, 4447–4459. [CrossRef]

21. Zhen, H.; Niu, D.; Wang, K.; Shi, Y.; Ji, Z.; Xu, X. Photovoltaic power forecasting based on GA improved bi-LSTM in microgrid without meteorological information. *Energy* **2021**, *231*, 120908. [CrossRef]

22. Dai, Q.; Huo, X.; Hao, Y.; Yu, R. Spatio-temporal prediction for distributed PV generation system based on deep learning neural network model. *Front. Energy Res.* **2023**, *11*, 1204032. [CrossRef]

23. Deng, S.; Cui, S.; Xu, A. Power Prediction of Regional Photovoltaic Power Stations Based on Meteorological Encryption and Spatio-Temporal Graph Networks. *Energies* **2024**, *17*, 3557. [CrossRef]

24. Lai, W.; Zhen, Z.; Wang, F.; Fu, W.; Wang, J.; Zhang, X.; Ren, H. Sub-region division based short-term regional distributed PV power forecasting method considering spatio-temporal correlations. *Energy* **2023**, *288*, 129716. [CrossRef]

25. Wang, Y.; Fu, W.; Zhang, X.; Zhen, Z.; Wang, F. Dynamic directed graph convolution network based ultra-short-term forecasting method of distributed photovoltaic power to enhance the resilience and flexibility of distribution network. *IET Gener. Transm. Dis.* **2024**, *18*, 337–352. [CrossRef]

26. Wang, F.; Chen, P.; Zhen, Z.; Yin, R.; Cao, C.; Zhang, Y.; Duic, N. Dynamic spatio-temporal correlation and hierarchical directed graph structure based ultra-short-term wind farm cluster power forecasting method. *Appl. Energy* **2022**, *323*, 119579. [CrossRef]

27. Huang, W.; Wang, D.; Ouyang, X.; Wan, J.; Liu, J.; Li, T. Multimodal federated learning: Concept, methods, applications and future directions. *Inf. Fusion* **2024**, *112*, 102576. [CrossRef]

28. Lin, J.; Ma, J.; Zhu, J. A privacy-preserving federated learning method for probabilistic community-level behind-the-meter solar generation disaggregation. *IEEE Trans. Smart Grid* **2022**, *13*, 268–279. [CrossRef]

29. Hosseini, P.; Taheri, S.; Akhavan, J.; Razban, A. Privacy-preserving federated learning: Application to behind-the-meter solar photovoltaic generation forecasting. *Energy Convers. Manag.* **2023**, *283*, 116900. [CrossRef]

30. Zhou, N.; Xu, X.; Yan, Z.; Shahidehpour, M. Spatio-temporal probabilistic forecasting of photovoltaic power based on monotone broad learning system and copula theory. *IEEE Trans. Sustain. Energy* **2022**, *13*, 1874–1885. [CrossRef]

31. Cheng, Z.; Liu, Q.; Zhang, W. Improved Probability Prediction Method Research for Photovoltaic Power Output. *Appl. Sci.* **2019**, *9*, 2043. [CrossRef]

32. David, M.; Ramahatana, F.; Trombe, P.J.; Lauret, P. Probabilistic forecasting of the solar irradiance with recursive ARMA and GARCH models. *Sol. Energy* **2016**, *133*, 55–72. [CrossRef]

33. van der Meer, D.W.; Widén, J.; Munkhammar, J. Review on probabilistic forecasting of photovoltaic power production and electricity consumption. *Renew. Sustain. Energy Rev.* **2018**, *81*, 1484–1512. [CrossRef]

34. Fernandez-Jimenez, L.A.; Monteiro, C.; Ramirez-Rosado, I.J. Short-term probabilistic forecasting models using Beta distributions for photovoltaic plants. *Energy Rep.* **2023**, *9*, 495–502. [CrossRef]

35. Ma, X.; Du, H.; Wang, K.; Jia, R.; Wang, S. An efficient QR-BiMGM model for probabilistic PV power forecasting. *Energy Rep.* **2022**, *8*, 12534–12551. [CrossRef]

36. Mei, F.; Gu, J.; Lu, J.; Lu, J.; Zhang, J.; Jiang, Y.; Shi, T.; Zheng, J. Day-ahead nonpara-metric probabilistic forecasting of photovoltaic power generation based on the LSTM-QRA ensemble model. *IEEE Access* **2020**, *8*, 166138–166149. [CrossRef]

37. Wan, C.; Lin, J.; Wang, J.; Song, Y.; Dong, Z. Direct quantile regression for nonparametric probabilistic forecasting of wind power generation. *IEEE Trans. Power Syst.* **2016**, *32*, 2767–2778. [CrossRef]

38. Golestaneh, F.; Pinson, P.; Gooi, H.B. Very short-term nonparametric probabilistic forecasting of renewable energy generation—With application to solar energy. *IEEE Trans. Power Syst.* **2016**, *31*, 3850–3863. [CrossRef]