

Article

Enhancing Wind Power Forecasting via Multi-Farm Coupling Under Data Isolation: A Physics-Guided Personalized Federated Approach

Yunjie Yang ¹, Yue Xiang ^{1,2,*} and Xinkang Liu ¹

¹ College of Electrical Engineering, Sichuan University, Chengdu 610065, China; jackyy@stu.scu.edu.cn (Y.Y.); liu67kk@163.com (X.L.)

² Clean Energy Power Systems and Equipment Key Laboratory of Sichuan Province, Chengdu 610065, China

* Correspondence: xiang@scu.edu.cn

Abstract

Accurate wind power forecasting is critical for grid stability and long-term sustainability, but mountainous wind farms face challenges from complex micro-meteorology, restricted communication, and non-IID data, exacerbated by data silos that prevent centralized learning. Most federated learning relies on data-driven averaging that ignores multi-farm coupling, or adopt complex local models that increase communication overhead. To address these, a physics-guided personalized federated approach is proposed to enhance wind power forecasting. Its core is a physics-guided aggregation mechanism that constructs a dynamic weight matrix from distance, elevation, and real-time wind direction to enable personalized aggregation capturing multi-farm coupling. The federated framework combines a shared CNN-LSTM with multi-head attention for regional patterns and a personalized layer for local microclimate. A risk-aware asymmetric loss is incorporated to penalize high-power errors, enhancing operational reliability under high-power conditions. Validation on mountainous wind farms for 3-day forecasting under typical and extreme scenarios across wet, dry, and normal seasons shows that the average R^2 exceeds 0.95, and the average RMSE is reduced by more than 24% compared to baselines, achieving high accuracy under strict privacy preservation. By enabling multi-farm coupling under data isolation, this approach achieves high forecasting accuracy on the studied wind farms, showing promise for similar ones.

Keywords: wind power forecasting; privacy protection; personalized federated learning; physics-guided dynamic weighting

Academic Editor: Carlos Vargas-Salgado

Received: 2 June 2026

Revised: 1 July 2026

Accepted: 6 July 2026

Published: 13 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

With the global low-carbon transition and sustainability drive, wind power has become increasingly important for its clean nature and key role in sustainable energy systems. In China, new wind capacity grew by 51% in 2025, with annual generation reaching 1.13 trillion kWh [1], making it a cornerstone of the country's sustainability agenda. However, its intermittent and stochastic nature poses significant challenges to the secure, stable, and sustainable operation of power systems with large-scale wind integration [2]. Accurate wind power forecasting can mitigate these impacts on grid security and economic

dispatch, supporting efficient utilization and enhancing the overall sustainability of wind power integration [3].

Wind power forecasting approaches can be grouped into four categories: physical, statistical, machine learning, and hybrid methods [4]. Physical methods estimate power via turbine models and NWP inputs, but require high-precision data and are computationally costly, limiting scalability [5]. Statistical methods (e.g., ARIMA [6], Kalman filtering [7]) capture temporal patterns but rely on stationarity assumptions and struggle with the nonlinear and stochastic nature of wind dynamics. Machine learning, especially deep learning, has become dominant in renewable energy forecasting owing to its powerful nonlinear fitting capability. Models from XGBoost [8] to LSTM [9] illustrate how data-driven paradigms consistently surpass conventional techniques [10]. Hybrid models, such as CNN-LSTM [11], improve performance by combining multiple techniques. Accurate data-driven forecasting relies on a large-scale, high-quality dataset. However, individual wind farms often suffer from limited or low-quality historical data, leading to degraded performance. While leveraging data from neighboring farms could alleviate this issue, centralized data pooling is generally infeasible due to strict energy data security policies and commercial confidentiality concerns that prevent cross-entity data sharing [12,13].

Various collaborative forecasting approaches have been explored to satisfy privacy requirements, but they struggle to jointly achieve privacy, efficiency, and accuracy. Noise-based anonymization [14] distorts data distributions and remains vulnerable to reconstruction attacks [15]. Distributed optimization approaches such as ADMM [16,17] require frequent exchanges of intermediate variables that remain vulnerable to reconstruction. Homomorphic encryption [18] introduces prohibitive computational and communication overhead, making it unsuitable for deployment in communication-constrained settings. Transfer learning and GAN-based methods [18] require external or shared data during the training of source models, which is often infeasible under strict privacy constraints.

Federated Learning (FL) has emerged as an effective privacy-preserving framework, where local models are trained on-site and only model parameters are exchanged, thus breaking data silos while keeping raw data private [19–22], but suffers from degraded forecasting accuracy under strong non-IID data distributions [23].

Existing federated learning (FL) studies for wind power forecasting primarily address statistical heterogeneity arising from non-IID data distributions. The mainstream research is built upon the FedAvg framework or its variants, and has predominantly focused on strengthening local client modeling capacity, with a secondary line of work investigating server-side aggregation strategies, including aggregation optimization and communication or parameter transmission design [24].

The main direction focuses on strengthening local client learning to better improve forecasting accuracy. Representative examples include the reinforcement-learning-based local model developed by Li et al. [21], the Transformer-based forecasting network proposed by Arooj et al. [25], and the adaptive proximal optimization strategy introduced by Wang et al. [26]. However, these methods implicitly assume that enhancing individual client capacity is sufficient to resolve global heterogeneity. This leads to a design tendency toward increasingly complex local models, which significantly increases model size and communication burden, thereby limiting their applicability in limited communication scenarios such as mountainous wind farms.

The other direction focuses on server-side aggregation strategies that aim to reconcile heterogeneous client updates through aggregation and communication design. Adaptive aggregation-based federated learning methods for statistical heterogeneity include FedMA [27], which attempts adaptive aggregation through layer-wise parameter matching, and Krum [28], which performs aggregation by automatically discarding outlier client updates before aggregation. Wang et al. [29] introduce adaptive parameter update mech-

anisms (APU) for wind power forecasting, integrating global averaging with client-specific adaptation modules to enable data-driven soft personalization at the server side. In a related wind power forecasting study, Wang et al. [30] propose a partial parameter aggregation strategy that shares only the feature extractor to reduce communication overhead. However, despite their differences in implementation, these approaches are fundamentally driven by data-driven optimization, where aggregation behavior is determined solely by observed gradients or parameter statistics. Moreover, many of these methods rely on the exchange of gradients or intermediate optimization signals during local training, which may introduce additional privacy exposure risks, as such information can potentially reveal sensitive characteristics of local data distributions.

Although several physics-informed FL approaches have been explored, their application to wind power forecasting remains relatively limited, partly because complex meteorological interactions are difficult to effectively encode within privacy-preserving federated learning. Consequently, existing studies (e.g., [31]) primarily incorporate physical knowledge into local model optimization or loss regularization, while server-side aggregation remains largely data-driven. However, wind farms exhibit inherent spatiotemporal dependencies arising from terrain characteristics and wind propagation between neighboring sites. Since these physically meaningful relationships are not explicitly reflected in purely data-driven aggregation schemes, valuable cross-site structural information may not be fully exploited during model aggregation, and the learned inter-site relationships may not necessarily align with real physical dependencies. This motivates the incorporation of physics-guided priors into the aggregation process to better leverage spatiotemporal dependencies among wind farms, thereby improving forecasting accuracy.

In summary, FL-based wind power forecasting has made progress, but it still faces clear limitations in mountainous settings, where rich wind potential [32] coexists with restricted communication, complex micro-meteorology, and pronounced non-IID data. These exhibit the following limitations:

(1) A substantial number of FL-based wind power forecasting methods increasingly rely on complex local models to handle non-IID data, which significantly increases communication overhead and reduces applicability in communication-constrained wind farm environments.

(2) Server-side aggregation remains data-driven in both adaptive FL methods and physics-informed FL, where physics is used in local optimization or loss regularization, leaving multi-farm coupling underutilized during aggregation. This may limit the physical interpretability of learned inter-site relationships.

To address the above issues, this paper proposes a physics-guided personalized federated learning framework for wind power forecasting. The core of the method is a physics-guided aggregation mechanism, implemented via a dynamic weight matrix constructed based on horizontal distance, elevation differences, and real-time wind direction, to enable personalized inter-farm model aggregation. A personalized architecture is further designed with a shared CNN-LSTM-MHA backbone for capturing regional spatiotemporal patterns and a client-specific private layer for modeling local microclimate variations. To improve robustness under critical operating conditions, a risk-aware asymmetric loss is introduced to emphasize high-power forecasting errors. The proposed model is evaluated using real-world data from four mountainous wind farms. Table 1 compares representative FL-based wind power forecasting methods.

Table 1. Comparison of representative FL-based wind power forecasting methods.

Method	Local Model	FL Strategy	Aggregation Strategy	Physical Information
FedProx [33]	Machine learning with proximal regularization	Regularized FL	FedAvg + proximal term	None explicitly (Data-driven)
FedPer [34]	Shared + private layers	Personalized FL	FedAvg on shared layers	None explicitly (Data-driven)
pFedMe [35]	Machine learning with Moreau Envelopes regularization	Personalized FL (Meta-learning)	Optimization-based aggregation	None explicitly (Data-driven)
Fed DRL [21]	Deep reinforcement learning	Standard FL	FedAvg/variants	None explicitly (Data-driven)
FedWindT [25]	Transformer-based network	Standard FL	FedAvg/variants	None explicitly (Data-driven)
Adaptive Proximal FL [26]	Neural networks	Regularized FL	FedAvg/variants	None explicitly (Data-driven)
FedMA [27]	Neural networks	Standard FL	Layer-wise matched aggregation	None explicitly (Data-driven)
Krum [28]	Machine learning	Robust FL	Outlier-filtering aggregation	None explicitly (Data-driven)
APU FL [29]	Machine learning	Personalized FL	FedAvg + APU personalized updates	None explicitly (Data-driven)
Adaptive Proximal FL [30]	Feature extractor + Predictor	Federated Transfer Learning	FedAvg/variants on feature extractor	None explicitly (Data-driven)
Physics-guided FL [31]	Physics-guided neural networks	Physics-Informed FL	FedAvg/variants	Encoded within local model to guide training
Our Work	CNN-LSTM-MHA+ private layers	Personalized FL	Physics-guided aggregation	Used to guide personalized aggregation

Thus, the major contributions of this paper can be summarized as follows:

(1) A physics-guided aggregation mechanism is proposed for personalized federated learning. A dynamic weight matrix is constructed based on horizontal distance, elevation, and real-time wind direction to generate personalized aggregation weights, enabling server-side aggregation to reflect spatiotemporal multi-farm coupling instead of relying on purely data-driven aggregation.

(2) A personalized federated learning method is developed to support the physics-guided aggregation mechanism under data isolation. A shared CNN-LSTM-MHA backbone is used to capture regional spatiotemporal patterns, while a client-specific private layer models local microclimate variation. In addition, a risk-aware asymmetric loss is introduced to enhance forecasting reliability in high-power scenarios.

The rest of the paper proceeds as follows: Section 2 details data processing and the design of the physics-guided dynamic weight matrix. Section 3 details the construction of the spatiotemporal personalized federated learning forecasting model. Section 4 presents a case study on mountainous wind farms to validate the proposed method. Section 5 concludes the paper and outlines future research directions.

2. Physics-Guided Dynamic Weight Matrix for Multi-Farm Coupling

2.1. Data Processing

Data preprocessing is indispensable for accurate wind power forecasting. Due to equipment failures, maintenance shutdowns, communication interruptions, and grid-initiated curtailment, raw measurements from actual wind farms routinely contain various anomalies. Untreated anomalies severely degrade forecasting performance, necessitating preprocessing to improve data completeness and usability. To ensure a rigorous evaluation and completely eliminate the risk of data leakage, the entire data processing pipeline in this study strictly adheres to a “split-before-processing” principle. Specifically, the raw data are first partitioned into wet, normal, and dry seasonal subsets, and then chronologically divided into training, validation, and testing sets prior to the execution of any preprocessing operations. This study defines three anomaly categories as the basis for the preprocessing strategy:

(1) Missing values: empty entries in the raw dataset. (2) Out-of-range values: values violating meteorological or electrical physical constraints. Boundaries are set as follows: wind speed is limited to 0–30 m/s (30 m/s corresponds to force 11 storm winds capable of seriously damaging turbines); wind direction follows true azimuth within 0–360°; temperature is bounded by −15–45 °C based on local historical records with margins; air pressure thresholds are altitude-adjusted with a 10% relaxation; relative humidity is 0–100%. The lower bound of wind power output is 0, combined with the cut-in and cut-out wind-speed criteria, while the upper bound is dynamically determined by each farm’s total installed capacity. (3) Consecutive repeated values: an identical feature value persisting continuously for a specified period (6 h in this study).

The preprocessing procedure is as follows. After aligning timestamps and computing the proportion of each anomaly type, consecutive repeated values are processed first. Wind direction, air pressure, and wind power output are excluded from repetition checks, as wind direction and pressure vary slowly, and power output may be in legitimate steady states such as dispatch curtailment or full generation. Wind speeds of zero, below cut-in, or above cut-out, indicating normal shutdown, are also excluded. For the remaining features, consecutive repetition exceeding 6 h is directly set to missing. Out-of-range correction is then performed. For air pressure anomalies caused by unit discrepancies (hPa vs. kPa), multiplication by 10 is attempted first; if the corrected value is reasonable, it is retained, otherwise set to missing. Negative power values and power output when wind speed is below cut-in or above cut-out are forced to zero. If any feature exceeds its boundary by no more than 5%, it is corrected to the corresponding nearest boundary; otherwise, the anomalous value is set to missing for subsequent imputation.

Finally, all resulting and originally missing values are imputed on a natural day basis. A single missing point is filled by linear interpolation. For consecutive missing segments within 2 h, considering the smooth continuity of short-term features such as wind speed and to avoid overshoot, piecewise cubic Hermite interpolation (PCHIP) is applied, as shown in Equation (1).

$$\begin{aligned}
 y(t) &= y_k H_1(s) + y_{k+1} H_2(s) + h_k d_k H_3(s) + h_k d_{k+1} H_4(s) \\
 \begin{cases} H_1(s) = 2s^3 - 3s^2 + 1; H_2(s) = -2s^3 + 3s^2 \\ H_3(s) = s^3 - 2s^2 + s; H_4(s) = s^3 - s^2 \end{cases}
 \end{aligned} \tag{1}$$

In this equation, (t_k, y_k) and (t_{k+1}, y_{k+1}) are the endpoint data of the missing segment, and $y(t)$ is the interpolated value at time t ; $h_k = t_{k+1} - t_k$ is the interval length; $s = t - t_k/h_k$ is the normalized time step; and dk is the first derivative.

For consecutive missing intervals between 2 and 12 h, the diurnal meteorological periodicity is exploited. The data are partitioned by day, and each missing point is first filled

with the average of the corresponding timestamps from the previous and following days. If either adjacent day is also missing, the average of the two preceding or two succeeding days is used instead. This step may break a long missing segment into shorter ones, which can then be processed by the aforementioned short-gap method. For continuous missing segments exceeding 12 h within a single day, that entire day is discarded to avoid introducing errors or spurious patterns from large-scale imputation.

After anomaly processing, the dataset is divided into three periods according to the Sichuan provincial policy document: the wet season (June–October), dry season (January–April and December), and normal season (May and November). Forecasting models are constructed separately for each period. Within each period, while preserving chronological order, the first 75% of data is used for training, 10% for validation, and 15% for testing. Meteorological features are then normalized to the [0,1] range within each respective dataset, as given in Equation (2).

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (2)$$

In this equation, x denotes the original value of a meteorological feature; x' is the corresponding [0, 1]-normalized value; and $x_{\max}/x_{\min}$ are the maximum and minimum values of that feature in the dataset to be normalized.

In actual wind farm operation, unit capacity expansions and dynamic growth in installed capacity are common. If the absolute power output is directly used as the forecasting target, a model trained on earlier historical data will struggle to generalize to the higher power range after a capacity upgrade. To address this, we normalize the wind power output by the total installed capacity at each moment, forming per-unit values as shown in Equation (3).

$$y_t = P_{\text{actual},t} / P_{\text{cap},t} \quad (y_t \in [0,1]) \quad (3)$$

In this equation, $P_{\text{actual},t}$ is the wind power output at time t ; and $P_{\text{cap},t}$ is the total installed capacity of the farm at time t .

The normalization of meteorological features eliminates dimensional disparities based on the statistical distribution of historical samples, preventing large-value meteorological features from dominating gradients during backpropagation and thereby ensuring training stability. The per-unit normalization of wind power output is essentially another form of [0, 1] normalization, mapping through the physical rated capacity. While safeguarding underlying training stability, it effectively preserves the forecasting model's accuracy after capacity changes. Constrained by turbine aerodynamic efficiency and wake losses, the actual per-unit output usually cannot reach the theoretical upper limit of 1. This physics-based normalization thus retains a realistic physical margin for the neural network, enabling it to implicitly learn the objective constraint that wind farms cannot always operate at full capacity.

Throughout the execution of the aforementioned anomaly correction, PCHIP interpolation, and adjacent-day imputation, all computational operations are performed solely based on historical observations within each isolated subset, strictly preventing any cross-set information exchange or access to future data. During data preprocessing, the dataset must be strictly split before any normalization to prevent data leakage into model training. Moreover, statistical parameters for normalization are estimated exclusively from the training set and subsequently applied to scale the validation and test sets. This rigorous isolation ensures strict independence among all subsets during the entire pipeline. This concludes the data preprocessing stage.

2.2. Physics-Guided Dynamic Weight Matrix Modelling

To effectively quantify the cross-farm feature transfer relevance within a wind farm cluster, this study designs a physics-guided dynamic weight matrix. This approach overcomes the limitations of graph neural networks, such as high communication costs in adverse mountainous environments and the risk of raw feature data privacy leakage across farms. It also adaptively assigns fusion weights to local models during federated aggregation, thereby enabling physics-guided dynamic aggregation of shared-layer parameters. This aggregation reflects regional meteorological evolution patterns for selective cross-farm learning from upstream to downstream wind farms.

For any two wind farms i and j , let i denote the source farm and j denote the target farm, thereby explicitly defining the direction of parameter transfer. Specifically, the source farm i provides its locally optimized shared parameters for federated aggregation, whereas the target farm j receives the aggregated parameters to update its local shared model. From a physical perspective, the source farm is regarded as the upstream node relative to the prevailing wind direction, while the target farm represents the downstream node influenced by the propagation of regional meteorological patterns. It should be noted that any wind farm can serve as either a source or a target under different source–target combinations. However, the corresponding aggregation weight is dynamically determined by the physics-guided weighting matrix, such that different source–target pairs contribute to different degrees according to their geographical distance, terrain blocking effect, elevation difference, and real-time wind direction alignment. This explicitly enables differentiated personalized aggregation among wind farms, fundamentally distinguishing it from the uniform parameter averaging conventionally utilized in standard federated learning frameworks.

To quantify the spatial correlation of a wind farm cluster in three-dimensional physical space, a spatial base weight is constructed based on the absolute positional relationship between farms. For any two wind farms i and j , it is defined as shown in Equation (4). The first product term is the reciprocal of the horizontal distance, representing the inherent attenuation of meteorological correlation by spatial separation. The second product term is an exponential decay function that quantifies the physical blocking effect of terrain undulations on airflow propagation, with output values ranging between 0 and 1.

$$W_{base}^{i \rightarrow j} = \frac{1}{d_{i,j}} \cdot \exp(-\lambda \cdot \frac{|h_i - h_j|}{d_{i,j}}) \quad (4)$$

In this equation, $d_{i,j}$ is the great-circle distance (in km) between farms i and j computed from their longitude and latitude using the Haversine formula; h denotes the farm elevation (in m). In complex mountainous terrain, the terrain gradient (the ratio of elevation difference to horizontal distance) can be substantial. Directly substituting such values into the exponential function easily causes numerical underflow, driving local weights to zero and inducing topological isolation. A smoothing coefficient λ is therefore introduced to map values into a numerically stable activation range.

Real airflow traversing complex mountain ranges is constrained not only by distance but also by the strong aerodynamic resistance of windward slopes. When air flows from a lower-elevation source farm to a higher-elevation target farm, overcoming gravity and surface friction causes significant dissipation. To quantify this nonlinear attenuation, a terrain blocking penalty term is introduced as shown in Equation (5).

$$P_{block}^{i \rightarrow j} = \exp(-\lambda \cdot \frac{\max(0, h_j - h_i)}{d_{i,j}}) \quad (5)$$

In this equation, the term $\max(0, h_j - h_i)$ ensures that the attenuation penalty is applied only when the target farm j is higher than the source farm i , for upslope airflow, while downslope or level advection unhindered by terrain remains unimpeded.

Using the real-time wind direction angle of the upstream farm i and the true geographic azimuth from farm i to farm j computed via spherical trigonometry, the dynamic weight is constructed as defined in Equation (6).

$$W_{dyn}^{(i \rightarrow j), (t)} = \begin{cases} \min(W_{base}^{i \rightarrow j} + \alpha \cdot \cos(\bar{\theta}_{wind}^{i, (t)} - \theta^{i \rightarrow j}) \cdot P_{block}^{i \rightarrow j}, 1) & \cos(\cdot) > 0 \\ \min(\beta \cdot W_{base}^{i \rightarrow j}, 1) & \text{others} \end{cases} \quad (6)$$

This dynamic weight matrix constitutes the core of directed federated parameter transfer in this study. The cosine term projects the real-time wind direction vector onto the spatial topological connection, quantifying the alignment between wind direction and inter-farm bearing. During federated learning, weight updates are synchronized with communication rounds, using the circular mean wind direction over that time interval to update the dynamic weights.

When the $\cos(\cdot) > 0$, a downwind condition prevails, and the downstream farm should anticipate and absorb the upstream wind fluctuation signatures that are about to arrive. An incremental reward, jointly constrained by alignment and terrain blocking (with gain coefficient α), is thus added to the base weight. The higher the wind direction alignment and the lower the terrain obstruction along the path, the larger the resulting conduction weight. This indicates that the downstream farm should place greater emphasis on learning the regional meteorological evolution patterns reflected in upstream shared-layer parameters at that moment, achieving more accurate wind power forecasting.

When the $\cos(\cdot) \leq 0$, an upwind or crosswind condition applies: the upstream farm's wind does not reach the downstream farm, and the upstream meteorological features become uncorrelated noise interference requiring suppression. To avoid frequent disconnections and severe oscillations in model parameters caused by localized wind direction fluctuations during training, the dynamic weight is not set directly to zero. Instead, a soft constraint is applied by attenuating the base weight with a decay coefficient ($0 < \beta < 1$), preserving training connectivity while suppressing interference from irrelevant meteorological features.

The dynamic weights between different farms satisfy the normalization condition, while each farm's self-weight is fixed at 1, reflecting the assumption that local data are reliable. Column-wise normalization is then applied to the dynamic weights and self-weights, forming an $N \times N$ physics-guided dynamic weight matrix, where N is the number of wind farms. During federated learning, the aggregation server uses this matrix to perform weighted aggregation of client parameters from each farm and distributes the personalized results. This not only overcomes the limitation of blind averaging in federated learning while preserving data privacy, but also maps regional meteorological evolution patterns reflected in shared-layer parameters into the parameter transfer mechanism, providing a solid technical foundation for the sustainable and efficient utilization of mountainous wind energy resources.

3. Spatiotemporal Personalized Federated Hybrid Forecasting Model

3.1. CNN-LSTM-MHA Forecasting Network Design

To accurately capture the severe volatility and long-range temporal dependency of wind power under complex terrain and to build on existing research in Ref. [13] on single-farm wind power forecasting, this study designs a forecasting network for each local client based on a CNN-LSTM backbone cascaded with a multi-head attention (MHA) mechanism. The structure is illustrated in Figure 1.

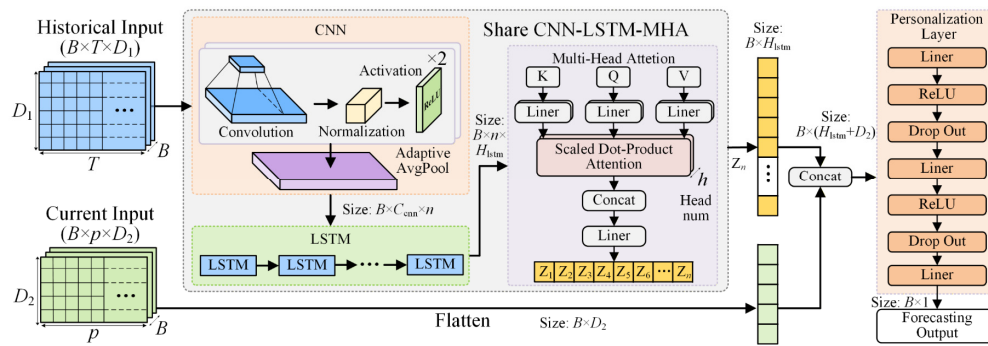


Figure 1. CNN-LSTM-MHA forecasting network.

The processed data are divided into historical data (containing historical meteorological variables and historical wind power output) and current-step meteorological data. The historical data, with dimensions $B \times T \times D_1$ (B : training batch size; T : historical time window length; D_1 : number of features), characterize the dynamic evolution of features and require temporal dependency mining. The current-step meteorological input, with dimensions $B \times p \times D_2$ (p : forecasting horizon length; D_2 : number of features), directly determines the physical boundary of the forecasting target at the current moment. If the two were mixed as input, the highly sensitive current-step features would be easily diluted or lagged by the smoothing effect of the historical sequence, degrading forecasting accuracy. Hence, the input data are partitioned accordingly.

For the historical input, a CNN comprising two stacked convolutional layers with activation functions and an adaptive average pooling layer (Adaptive AvgPool, outputting n features, $n < T$) is first applied. This filters out high-frequency random noise, captures instantaneous local abrupt features such as sudden wind-speed changes, and down-samples the lengthy original time window into a more compact high-level feature sequence. The sequence then enters a multi-layer stacked LSTM network, where the gating mechanisms within LSTM neurons capture long-range dependencies in the evolution of meteorological variables and wind power output, producing a high-dimensional hidden state sequence with encoded temporal relationships. To prevent information attenuation of long sequence features at the final time step and to accurately extract time steps strongly correlated with power variations, an MHA mechanism is cascaded after the LSTM. This mechanism maps the temporal hidden states into multiple independent attention subspaces. Through scaled dot-product operations on the query, key, and value vectors and the evaluation of the influence distribution of each time node in the historical sequence on the current power state, it leverages the global receptive field property. The final sequence node Z_n serves as the query, and the weighted sum of prior meteorological evolution nodes ($Z_1 \sim Z_{n-1}$) forms the historical state representation vector, as shown in Equation (7), thereby reducing redundant interference in long time series.

$$Z_n = \text{softmax}\left(\frac{Q_n K^T}{\sqrt{d_k}}\right) V \quad (7)$$

In this equation, Q_n is the query vector for the target time step n ; K and V are the key and value vectors containing all time steps in the sequence; and d_k is the dimensionality of Q and K in the attention subspace, where $d_k = H_{lstm}/h$.

The above CNN, LSTM, and MHA modules together constitute the Shared CNN-LSTM-MHA network whose parameters participate in server-side aggregation, with the purpose of extracting general meteorological evolution patterns of the farms. The current-step meteorological input, after flattening, is concatenated with the temporally encoded historical state representation vector along the feature dimension to form a hybrid feature

vector, which is then fed into the client's personalized private layer. This private layer is alternately constructed from fully connected layers (Linear), ReLU activation functions, and Dropout layers to prevent overfitting. It is retained locally on the client during federated training and performs nonlinear residual fitting on the fused features to learn local-specific microclimate patterns, ultimately outputting the wind power forecasting value.

Conventional deep learning-based forecasting typically adopts MSE and similar symmetric loss functions that impose uniform penalties during training. In actual operations, however, forecasting errors in the high-power operating range often cause more severe grid dispatch deviations and reserve capacity pressure. To address this, a risk-aware asymmetric training strategy is formulated, enabling the low-carbon and sustainable operation of mountainous power grids with high wind penetration. Specifically, an asymmetric load-weighted loss function, as defined in Equation (8), is constructed to impose heavier penalties on inaccurate forecasting in the high-power range, thereby ensuring forecasting accuracy in this critical region. Additionally, inspired by FedProx [33], a penalty term on the deviation of shared-layer parameters is designed to further mitigate the heterogeneity caused by non-IID data.

$$\mathcal{L}_{total} = \gamma \mathcal{L}_w + (1 - \gamma) \mathcal{L}_{huber} + \frac{\mu}{2} \left\| \Theta_{Shared}^{i,opt,(t)} - \Theta_{Shared}^{i,(t)} \right\|_2^2$$

$$\begin{cases} w_i = 1 + (\sqrt{B} - 1) \cdot \exp[\ln(\sqrt{B} + 1) \cdot y_i] \\ \mathcal{L}_w = \frac{1}{N} \sum_{i=1}^N w_i \cdot (y_i - \hat{y}_i)^2 \\ \mathcal{L}_{huber}(y, \hat{y}) = \begin{cases} \frac{1}{2} (y - \hat{y})^2, & \text{if } |y - \hat{y}| \leq \delta \\ \delta |y - \hat{y}| - \frac{1}{2} \delta^2, & \text{otherwise} \end{cases} \end{cases} \quad (8)$$

In this equation, B is the training batch size; y and $\hat{y}$ are the true and predicted per-unit wind power values; N is the number of data samples; and δ is the threshold of the Huber loss for absolute error. When the absolute error exceeds δ , a linear penalty is applied to handle extreme outliers, ensuring training stability. λ is a weighting coefficient for adjusting the proportion of the two terms; μ is the regularization coefficient for penalizing the deviation of the optimized shared-layer parameters from those distributed by the server (ensuring that extreme weather events do not disrupt the learning of regional meteorological patterns), and its value is set to approximately 0.1 times the sum of the first two terms.

Regarding the construction of the weight coefficient w : To characterize the asymmetric evolution of grid dispatch risks induced by forecasting errors, at low-to-medium power outputs grid reserves are ample and the safety risk from errors is relatively low. However, as wind power output approaches the installed capacity, the risk of frequency violations or wind curtailment caused by forecast drops or surges exhibits an exponential growth trend. An exponential mapping is therefore introduced to achieve nonlinear continuous penalization, and a constant term of 1 is added to ensure that the loss can degenerate to the standard MSE, effectively preventing metric space collapse during training.

Regarding the construction of the coefficient within w : to prevent rare full-generation samples from being diluted by the batch averaging effect during training, the extreme full-generation weight is set to be comparable to the batch size B . Meanwhile, based on stochastic optimization theory, to maintain an effective signal-to-noise ratio, the minimum weight is set as the reciprocal of the gradient noise scale $\sqrt{B}$. By substituting the boundary constraints $y_i/P_{cap,i} = 0$ and 1, the coefficients can be solved.

3.2. Spatiotemporal Federated Learning Model

To break down data silos across farms while preserving data privacy and to enable dynamic learning of beneficial cross-regional meteorological evolution patterns, this study constructs a spatiotemporal personalized federated learning model based on CNN-LSTM-MHA, as illustrated in Figure 2.

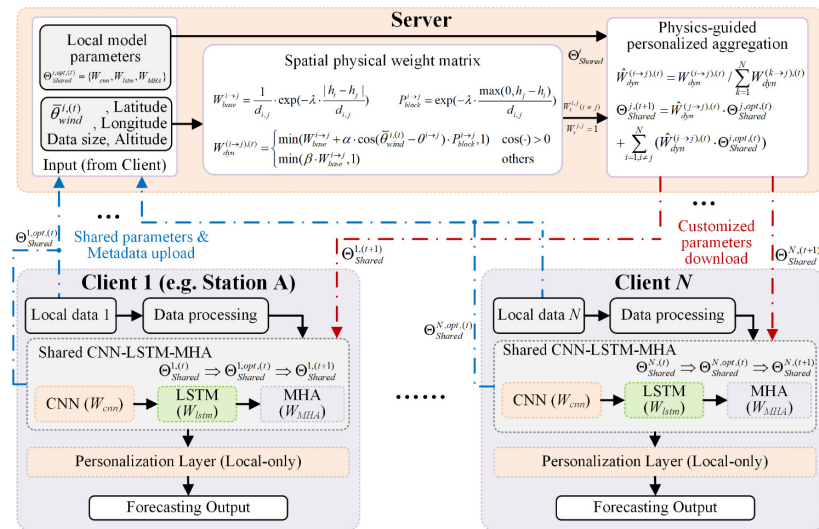


Figure 2. Architecture of the personalized federated learning model.

The forecasting network of each client is divided into two parts: a global shared layer (shared CNN-LSTM-MHA) and a personalized private layer (personalization layer). Within the same federated learning model, the neural network structure of the shared layer across all clients remains strictly identical. Let $\Theta_{Shared}^{i(t)} = \{W_{cnn}, W_{lstm}, W_{MHA}\}$ denote the parameters of the shared CNN-LSTM-MHA for the i -th farm at communication round t . The shared layer aims to generalize across wide spatial domains and extract the complex common evolution patterns of macroscale meteorology. The personalized private layer, composed of fully connected layers, activation functions, and regularization modules, performs local nonlinear residual fitting on the fused global features. This eliminates farm-specific heterogeneity induced by micro-topographic disturbances and improves local forecasting accuracy, offering a useful reference for the long-term sustainable development of mountainous renewable energy sectors.

Within the t -th federated communication round, each client first performs step-by-step forward forecasting of wind power output using a sliding historical window of length T and the shared parameters $\Theta_{Shared}^{i(t)}$ distributed from the previous round or obtained from initial forecasting, with the model parameters kept frozen. As time progresses, the client continuously accumulates newly observed ground truth into a local incremental data block. Once T_{fed} real-world time steps have been collected, the system triggers a local parameter update and the federated communication phase. The client uses this incremental data segment of length T_{fed} as the ground truth within its local training set and performs E local epochs. It jointly optimizes both the shared layer and the personalized private layer parameters via gradient descent, completing the network parameter update. Meanwhile, the client computes the circular mean wind direction over this incremental time block, as defined in Equation (9), and uploads it together with the optimized shared-layer parameters $\Theta_{Shared}^{i,opt,(t)}$ to the central server.

$$\bar{\theta}_{wind}^{i,(t)} = \{ \text{atan2}[\frac{1}{n} \sum_{i=1}^n \sin(\theta_{wind}^{i,(t)} \times \frac{\pi}{180}), \frac{1}{n} \sum_{i=1}^n \cos(\theta_{wind}^{i,(t)} \times \frac{\pi}{180})] \times \frac{180}{\pi} + 360 \} \pmod{360} \quad (9)$$

In this equation, the addition of 360° , division by 360° , and subsequent modulo operation serve to standardize the angle converted from the radian result of the atan2 function, ensuring it falls within the $0\text{--}360^\circ$ range of true wind direction azimuth.

After $\Theta_{Shared}^{i,opt,(t)}$ and $\bar{\theta}_{wind}^{i,(t)}$ are uploaded to the server, the server stores non-privacy-sensitive information such as data volume, longitude, latitude, and elevation of each farm. The physics-guided dynamic weights are computed via Equations (4)–(6), yielding an $N \times N$ dynamic weight matrix $W_{dyn}^{(i \rightarrow j),(t)}$. Column-wise normalization is then performed on this inter-farm dynamic weight matrix, as shown in Equation (10). Subsequently, personalized weighted aggregation is applied to the shared parameters of client j , as defined in Equation (11).

$$\hat{W}_{dyn}^{(i \rightarrow j),(t)} = W_{dyn}^{(i \rightarrow j),(t)} / \sum_{k=1}^N W_{dyn}^{(k \rightarrow j),(t)} (i, j \in N \cup \{i = j\}) \quad (10)$$

$$\Theta_{Shared}^{j,(t+1)} = \hat{W}_{dyn}^{(j \rightarrow j),(t)} \cdot \Theta_{Shared}^{j,opt,(t)} + \sum_{i=1, i \neq j}^N (\hat{W}_{dyn}^{(i \rightarrow j),(t)} \cdot \Theta_{Shared}^{i,opt,(t)}) \quad (11)$$

Once the personalized aggregation is executed across all target farms j in $\{1, \dots, N\}$, the globally updated parameter set seamlessly transitions to serve as the local source models i for the subsequent communication round $t + 1$, thereby completing one full federated learning loop.

The personalized weighted aggregated parameters are then distributed back to the shared layer of each client, concluding one federated communication round and commencing the next. Federated learning terminates when the L2 norm of the difference between personalized aggregated parameters from two consecutive rounds falls below a preset small threshold, or when the maximum number of communication rounds is reached. The number of federated communication rounds generally equals the total data volume divided by T_{fed} time steps. In this study, T_{fed} is set to one day, encompassing a complete diurnal cycle to capture the daily mean wind direction. This aligns with the day-ahead dispatch requirements of power grids, avoids the frequent forgetting of neural networks caused by high-frequency updates, and allows communication to take place during nighttime off-peak bandwidth in communication-constrained mountainous areas, thereby reducing overhead.

Through the personalized private layer on each local client and the personalized weighting by the physics-guided dynamic weight matrix on the server, combined with the penalty constraint on shared parameter deviation incorporated in the risk-aware asymmetric training strategy, the proposed framework achieves differential dynamic learning according to weights while preserving local microclimate adaptability. All this is done under strict data privacy protection. Consequently, effective cross-farm transfer and collaborative sharing of regional meteorological patterns reflected in shared-layer parameters are realized while raw data remain physically isolated. This mitigates non-IID heterogeneity and ultimately achieves accurate wind power forecasting.

To further enhance the reproducibility of this study, Table 2 delineates the complete training and communication procedure of the proposed physics-guided personalized spatiotemporal federated learning model. Within each communication round, each client first performs rolling forecasting using frozen parameters, and then jointly optimizes both the shared CNN-LSTM-MHA and the local personalization layer upon collecting a complete local incremental data block. Subsequently, only the locally optimized shared parameters

and non-privacy-sensitive metadata (such as the circular mean wind direction via Equation (9)) are uploaded to the central server. The server utilizes this information to construct a physics-guided dynamic weight matrix via Equations (4)–(6), facilitating a personalized weighted aggregation process as defined in Equations (10) and (11). Finally, the newly customized shared parameters are distributed back to their respective clients for the subsequent round. Throughout this entire procedure, the personalization layers remain strictly local, which not only effectively accommodates site-specific microclimate heterogeneity but also fundamentally guarantees raw data privacy.

Table 2. Physics-guided personalized spatiotemporal federated learning.

Input:	
Local datasets (Station i , $i \leq N$, total N stations)	
Initial specific shared parameters: i -th Station:	$\Theta_{Shared}^{i,(0)} = \{W_{cnn}, W_{lstm}, W_{MHA}\}$
Communication interval:	T_{fed}
Local training epochs	E
Maximum communication rounds	R
Physical metadata {Longitude, Latitude, and Elevation}	
Output:	
Personalized forecasting models,	$\Theta_{Shared}^{i,(R)}, \Theta_{Private}^{i,(R)}$
1: Initialize specific shared parameters $\Theta_{Shared}^{i,(0)}$ (Shared Layer) for each client i 2: Initialize local personalization parameters $\Theta_{Private}^{i,(0)}$ (Personalization Layer) for each client i 3: For communication round $t = 1, 2, \dots, R$ do 4: Server distributes personalized shared parameters $\Theta_{Shared}^{i,(t-1)}$ to each client i 5: for each client i in $\{1, \dots, N\}$ in parallel do 6: Perform rolling forecasting over a sliding window using frozen $\Theta_{Shared}^{i,(t-1)}$ 7: Collect continuous new samples to form local incremental data block $\mathcal{D}_i^{(t)}$ of length T_{fed} 8: Optimize $\Theta_{Shared}^{i,opt,(t)}$ and $\Theta_{Private}^{i,(t)}$ on $\mathcal{D}_i^{(t)}$ using the risk-aware asymmetric loss with proximal regularization (Equations (7) and (8)) for E local epochs 9: Compute the circular mean wind direction $\bar{\theta}_{wind}^{i,(t)}$ over $\mathcal{D}_i^{(t)}$ according to Equation (9) 10: Upload $\Theta_{Shared}^{i,opt,(t)}, \bar{\theta}_{wind}^{i,(t)}$, data size to Server 11: end for (line 5) 12: Server computes the physics-guided dynamic weight matrix $W_{dyn}^{(i \rightarrow j),(t)}$ using distance, terrain blocking, elevation difference, and real-time wind direction according to Equations (4)–(6) 13: Normalize $W_{dyn}^{(i \rightarrow j),(t)}$ column-wise to obtain $\hat{W}_{dyn}^{(i \rightarrow j),(t)}$ according to Equation (10) 14: for each client j in $\{1, \dots, N\}$ do 15: Perform personalized weighted aggregation: $\Theta_{Shared}^{j,(t+1)} = \hat{W}_{dyn}^{(i \rightarrow j),(t)} \cdot \Theta_{Shared}^{i,opt,(t)} + \sum_{i=1, i \neq j}^N (\hat{W}_{dyn}^{(i \rightarrow j),(t)} \cdot \Theta_{Shared}^{i,opt,(t)}) \text{ according to Equation (11)}$ 16: end for (line 14)	
Note: The updated target models j inherently become the new sources i for round $t + 1$, forming a closed federated loop. $\Theta_{Shared}^{j,(t+1)} \Rightarrow \Theta_{Shared}^{i,(t+1)}$	
17: if $\frac{1}{N} \sum_{j=1}^N \ \Theta_{Shared}^{i,(t)} - \Theta_{Shared}^{j,(t-1)}\ _2 < \epsilon$ (L2 norm) or $t > R$ then	
18: break	
19: end if	
20: end For (line 3)	
21: Return $\Theta_{Shared}^{i,(R)}, \Theta_{Private}^{i,(R)}$, get personalized forecasting models	

Hyperparameter settings in the local CNN-LSTM-MHA forecasting network and the personalized federated learning framework significantly influence forecasting accuracy. Manual tuning of certain hyperparameters often relies on experience. Bayesian optimization (BO), when handling the high-dimensional, non-convex hyperparameter space of deep neural networks, can achieve a dynamic balance between global optimization and precise convergence with relatively low computational overhead. Therefore, this study employs BO to provide more objective hyperparameter values for the forecasting model.

4. Case Study

4.1. Dataset and Preprocessing

The dataset used in this study originates from four real wind farms affiliated with different power generation companies. The maximum great-circle distance between any two farms not exceeding 150 km. The data exhibit strong non-IID characteristics. Due to real-data privacy protection requirements, the farms are anonymized herein.

This region is subject to the combined influence of dual monsoons and local mountain-valley winds, resulting in pronounced spatial heterogeneity. (1) Station A is a river-front ridge wind farm at an average elevation of approximately 2000 m, situated at the center of the study area (closest to Station B and farthest from Station D). The interaction between the diurnal valley breeze from the river and the prevailing wind leads to highly variable wind directions and elevated turbulence intensity. (2) Station B is an open hilly ridge wind farm at an average elevation of 2100 m, located at the southwesternmost end. As the frontline of the southwest monsoon, it experiences relatively weak local topographic disturbances, with a comparably steady flow field. (3) Station C lies in a transitional zone between high mountain valleys and karst terrain at an average elevation exceeding 3000 m, at the easternmost end. The steep and fragmented underlying surface gives rise to strongly non-uniform local flow, with frequent gusty winds and severe convection. (4) Station D is situated in the Anning River Valley at the northernmost end, at an average elevation of 1500 m (farthest from Station B). The deep north-south valley produces a significant funneling effect that concentrates wind energy, but the thermally driven unevenness caused by elevation differences between the valley floor and surrounding mountains results in extremely complex local wind shear characteristics.

The raw data from the four wind farms all have a temporal resolution of 15 min and share six common features: wind speed (m/s), wind direction ($^{\circ}$, true azimuth), temperature ($^{\circ}\text{C}$), atmospheric pressure (hPa), humidity (%), and wind power output (MW). Due to different commissioning times, the data volumes vary across farms. Station A has only one year of data (the full year 2020), with an initial total installed capacity of 146 MW, later expanded to 168 MW after May 2020. Station B spans approximately 1.6 years (17 May 2019 to 30 December 2020), with an initial capacity of 198.5 MW, expanded to 233.5 MW after November 2019. Stations C and D each contain three years of data (1 January 2018 to 30 December 2020). Station C had an initial capacity of 169.5 MW, expanded to 212 MW after November 2018, while Station D underwent no capacity expansion during this period and maintained a total installed capacity of 202 MW.

Following the anomaly identification and data processing methodology defined in this study, before processing, Station A had an anomaly ratio of 30.80% (a sample is considered anomalous if any of the six features is abnormal). Among these, atmospheric pressure out-of-range values reached 15.98%. Observation of the raw data suggests this was likely caused by inconsistent unit recordings (hPa vs. kPa). Humidity consecutive repeated values accounted for 8.17%, and wind power out-of-range values reached 7.76%, with all other anomaly types each below 2%. After data processing and period division, 29 natural days were removed from Station A's dataset, leaving 12,672 samples for the wet season, 13,952 for the dry season, and 5568 for the normal season. Station B had an

overall anomaly ratio of 18.56%, with consecutive repeated values for humidity, wind power, and temperature reaching 10.57%, 5.61%, and 5.22%, respectively. After processing, 47 days were removed, yielding 25,920, 17,280, and 8352 samples for the wet, dry, and normal seasons, respectively. Station C had an anomaly ratio of 15.22%, of which temperature and pressure out-of-range values accounted for 5.62% and 5.93%, while the remaining anomaly types were all below 3.7%. After processing, 104 days were removed, resulting in 40,320, 35,040, and 16,896 samples for the three seasons. Station D had an anomaly ratio of 26.49%, with wind power out-of-range values reaching 19.58% and consecutive repeated meteorological features around 5%. After processing, 66 days were removed, leaving 40,320, 41,376, and 17,088 samples for the wet, dry, and normal seasons, respectively.

A quantitative analysis of the statistical characteristics (mean, standard deviation, minimum, maximum, skewness, and kurtosis) of the preprocessed data from four wind farms (Stations A–D) across different seasons (i.e., wet season, dry season, and normal season) reveals evident spatiotemporal variability in the meteorological variables and wind power output. In the spatial dimension, wind resources and power generation differ considerably among stations. For example, during the dry season, the average wind speed at Station C reaches 8.84 m/s (with a corresponding average power output of 118.14 MW), substantially exceeding that of Station D (5.43 m/s and 51.11 MW, respectively). In the temporal dimension, the statistical characteristics at the same station also vary noticeably across seasons. For instance, the average wind speed at Station C decreases to 4.86 m/s during the wet season, accompanied by a reduction in average power output to 40.22 MW. Furthermore, the skewness and kurtosis statistics reveal noticeable differences in the distributional characteristics of several variables. While the kurtosis of most variables remains below 5, a few variables exhibit substantially higher values. For example, atmospheric pressure reaches a kurtosis of 150.47 at Station B during the wet season, while Station D records a pressure skewness of 6.95 and a kurtosis of 34.62. Wind direction also shows noticeable asymmetric fluctuations (e.g., Station C exhibits a skewness of −2.89 and a kurtosis of 12.67 during the dry season). Overall, these statistical characteristics indicate considerable variability across wind farms and seasons, with certain variables exhibiting pronounced skewness or heavy-tailed distributional characteristics.

A comparison with the raw datasets indicates that the preprocessing pipeline effectively removed severe sensor faults and abnormal observations, correcting distorted statistical moments and physically implausible extreme values (e.g., an initial kurtosis exceeding 90,000). Nevertheless, noticeable skewness and elevated kurtosis remain in the cleaned datasets. Although most variables exhibit kurtosis values below 5, a small number of variables still present relatively high kurtosis (e.g., a pressure kurtosis of 150.47), indicating that the cleaned datasets retain noticeable skewness and heavy-tailed distributional characteristics after preprocessing rather than arising solely from obvious measurement errors. These results indicate that the preprocessing procedure improves data quality while preserving the underlying statistical characteristics of the measurements across different wind farms and seasons.

4.2. Interpretability of the Physics-Guided Dynamic Weight Matrix

To validate the effectiveness of the proposed dynamic weight matrix, Figure 3 presents the monthly averaged physics-guided dynamic weight matrix for July 2020. The daily dynamic weight matrices generated at each federated communication round are averaged over the monthly scale. This approach effectively avoids evaluation contingency introduced by localized random meteorological fluctuations on individual days, while also preventing excessive spatial weight smoothing caused by an overly long statistical

period, thereby accurately preserving the relatively stable physical asymmetry characteristics between farms.

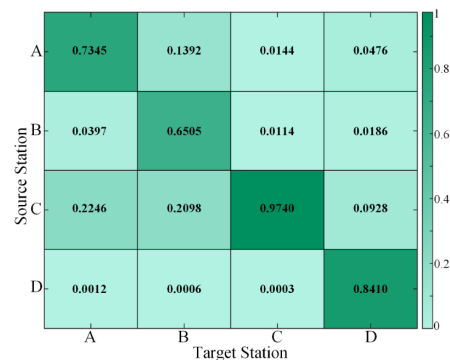


Figure 3. Monthly average physics-guided dynamic weight matrix.

Due to column-wise normalization, each column sums to 1, quantifying the target farm's absorption of regional meteorological patterns in shared-layer parameters from each source farm. Taking Station C as an example: constrained by an average elevation difference of nearly 1000 m, terrain blocking significantly weakens the meteorological influence of other farms on it. When Station C serves as the target farm, the external farm weights are all below 0.02, and its learning relies on local meteorological self-learning. However, when serving as a source farm, its potential energy characteristics can effectively radiate downstream, providing shared weights of approximately 0.2246 and 0.21 to Station A and Station B, respectively. Station B depends on the regional meteorological patterns reflected in shared-layer parameters from upstream Stations A and C to compensate for its own knowledge limitations, yet due to spatial distance attenuation, the farthest Station D contributes negligibly to its weight. The physics-guided dynamic weight matrix can attenuate shared-layer parameter learning that may cause negative transfer under up-wind or crosswind conditions. It dynamically guides each farm to learn regional meteorological patterns reflected in shared-layer parameters, ensuring that the model achieves safe and efficient aggregation guided by physical causal logic when faced with strongly non-IID meteorological data.

Based on the data from this month, Station B is used as an example to illustrate the daily variation of wind direction and the dynamic weight coefficients of its primary upstream source farms (Station A and Station C), as shown in Figure 4. This intuitively demonstrates how the weights of regional meteorological patterns reflected in shared-layer parameters from source farms vary dynamically with wind direction during the federated communication cycles.

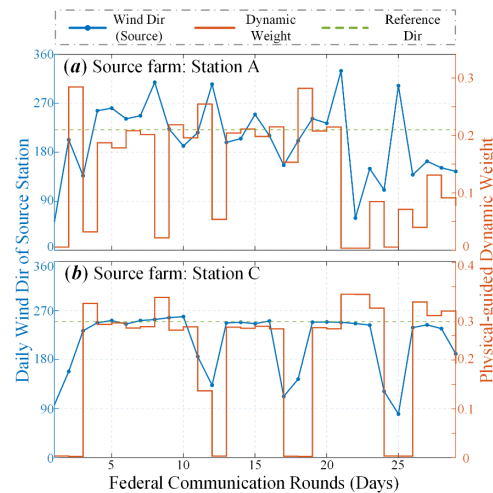


Figure 4. Time-series variation of federated dynamic weights for Station B.

In Figure 4, the blue line represents the wind direction of the source farm, the orange line represents the dynamic weight from the source farm to Station B, and the green dashed line represents the geographic azimuth angle from the source farm to Station B, which remains constant as a reference angle. Figure 4a corresponds to source farm Station A, and Figure 4b to source farm Station C. Based on the preceding analysis, the average weight from Station D as a source farm to Station B during this period is merely 0.0006, and its influence is negligible; therefore, the variation of dynamic weight from Station D to Station B with source farm wind direction is not discussed.

In the second federated communication round, i.e., the second day, the daily mean wind direction of source farm Station A closely approaches the geographic azimuth angle, causing its dynamic weight to surge to around 0.3. Although the wind direction of source farm Station C also approaches the baseline during this time, under the global competition mechanism of column-wise normalization, the weight increase of Station C is suppressed due to the superior alignment and closer spatial proximity of Station A. From the 4th to the 10th day, the wind direction of source farm Station C achieves high alignment with the geographic azimuth angle, driving its weight above 0.3, surpassing that of Station A. This confirms that the proposed dynamic weight matrix can reflect the phenomenon whereby meteorological alignment can compensate for the meteorological attenuation disadvantage caused by spatial distance. Except for passive compensation by the normalization mechanism when all source farms are in upwind conditions, the greater the wind direction deviation, the greater the attenuation of the weight coefficient, which is consistent with objective physical laws. The federated architecture with the physics-guided dynamic weight matrix aggregation can adaptively select the most valuable regional meteorological patterns reflected in shared-layer parameters from source farms at different times. This adaptive cross-site learning capability enables the sustainable and efficient deployment of distributed renewable energy systems.

To investigate the impact of key parameters in the physics-guided mechanism on collaborative forecasting, a sensitivity analysis of key parameters in the physics-guided mechanism is conducted across four wind farms. The results are shown in Figure 5.

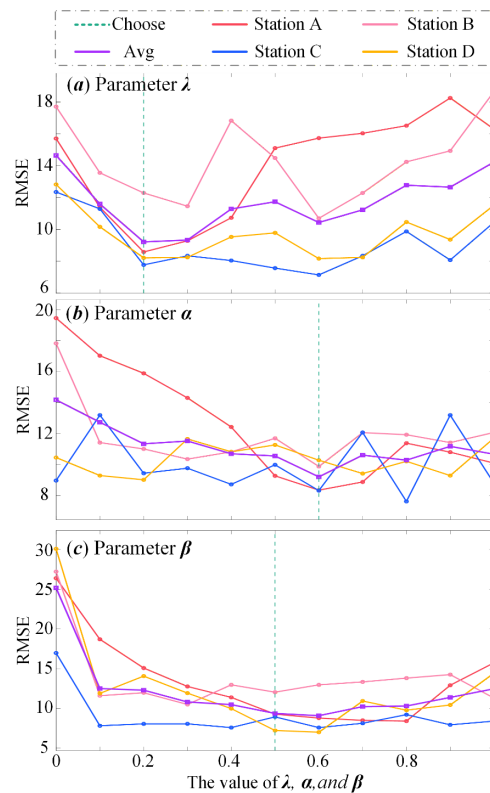


Figure 5. Sensitivity analysis in the wet season (λ , α , and β).

The smoothing coefficient λ represents terrain-induced attenuation of spatial correlation, reflecting how topography impedes airflow propagation. Results show a non-symmetric U-shaped trend. When $\lambda = 0$ (flat-terrain case), avgRMSE is high (14.64), indicating that without terrain constraints, physically blocked signals are improperly propagated across wind farms. As λ increases, forecasting error decreases and reaches its lowest level in the neighborhood of λ around 0.2 (avgRMSE = 9.20 under the discrete search grid). Clear spatial heterogeneity is observed: Station A and D perform best at λ around 0.2, while Stations B and C perform best at λ around 0.6 (RMSE 10.69 and 7.13). This reflects different local damping needs under complex terrain. When $\lambda = 1.0$, elevation effects are over-amplified, spatial correlation rapidly vanishes, and the model degenerates into near-independent learning (avgRMSE = 14.22), confirming that λ must remain within a physically valid range.

The gain coefficient α models coherent advection under downwind alignment, regulating information propagation along wind direction. When $\alpha = 0$, forecasting performance is worse (avgRMSE = 14.17). Increasing α improves performance, reaching its lowest level in the neighborhood of α around 0.6 (avgRMSE = 9.20). Beyond this range, gains saturate and fluctuate (e.g., 11.17 at $\alpha = 0.9$) due to truncation-induced saturation, where large α activates the upper bound in weight normalization, reducing directional sensitivity and limiting further improvement.

The decay coefficient β captures turbulence-induced decorrelation under upwind or crosswind conditions, acting as a damping factor to suppress spurious dependencies while preserving small positive weights for stability. When $\beta = 0$, spatial links are abruptly removed, causing severe degradation (25.16) due to broken connectivity. Setting $\beta = 0.5$ – 0.6 smooths turbulence effects and reduces avgRMSE to around 9.08. However, $\beta = 1.0$ again increases error (12.46) as weak noisy signals are over-propagated. Overall, β must

be properly constrained to balance stability and noise suppression in federated forecasting.

4.3. Results of Wind Power Forecasting

4.3.1. Parameters and Baseline Model Settings

To evaluate the forecasting accuracy of the proposed spatiotemporal personalized federated forecasting model, two baseline models are established for comparison. The first is a local independent forecasting model (Base CNN-LSTM), which adopts the identical network structure as the client in this study, but uses MSE as the loss function. It trains the forecasting model solely on local data without any parameter exchange. The second is a data-volume-weighted averaging federated learning model (Base FedAvg), which employs the same client and federated learning architecture as this study, with MSE loss and data-volume-weighted aggregation instead of uniform averaging on the server side.

In addition, two representative federated learning variants are included. FedPer shares the same client architecture and MSE loss, and performs data-volume-weighted aggregation on shared layers. Its key idea is network decoupling, i.e., separating shared global layers and local personalized layers, where only shared parameters are communicated while local layers are retained to capture client-specific patterns. FedProx also adopts the same architecture and MSE loss, and uses data-volume-weighted aggregation. It introduces a proximal regularization term to constrain local updates from deviating too far from the global model, thereby alleviating client drift under non-IID data and improving training stability.

For the proposed forecasting model in the wet, dry, and normal seasons, BO is applied to tune the hyperparameters and three coefficients: the smoothing coefficient λ in Equations (4) and (5), the gain coefficient α in Equation (6), and the penalty adjustment coefficient $\mu/2$ in Equation (8), as detailed in Table 3. The values correspond to the optimization results for the wet, dry, and normal seasons, respectively, and some values have been rounded after testing without compromising forecasting accuracy.

Specifically, the optimization objective is defined as the minimization of the average validation RMSE across the four wind farms, which guides the search for the parameters. For any hyperparameter shared with the baseline models, the search space is kept identical, and the optimal values are obtained through the same BO procedure. In the experimental design, all compared models adopt an identical optimization budget, with 30 trials per model, ensuring a fair and consistent hyperparameter tuning process across different methods. Regarding computational cost, the total runtime associated with the BO process of the proposed model ranges from 678 s to 4469 s across different tasks. This variation is mainly due to differences in seasonal dataset sizes and corresponding model configurations. This procedure is conducted offline and does not affect training or inference, introducing no additional deployment overhead.

Table 3. Optimization of hyperparameters and coefficients values.

Parameter	CNN Channels *	LSTM Hidden Units *	Attention Heads *	Private Layer Hidden Units *	LSTM Layers	λ
Range	{16, 32, 64, 128, 256}	{16, 32, 64, 128, 256}	{2, 4, 8, 16, 32}	{8, 16, 32, 64, 128}	[1, 5] (Integer)	[0, 1]
Value	128/128/64	128/64/64	8/8/2	64/32/32	1/1/1	0.2/0.3/0.2
Parameters	Dropout Rate	Initial Learning Rate	Batch Size *	Local Epochs	α	$\mu/2$
Range	[0, 0.6]	$[10^{-5}, 0.1]$	{16, 32, 64, 128, 256}	[1, 10] (Integer)	[0, 1]	$[10^{-4}, 10^{-3}]$
Value	0.19/	2.93×10^{-3} /	64/32/64	5/6/3	0.60/	4×10^{-4} /

0.15/ 0.26	$1.20 \times 10^{-4}/$ 5.32×10^{-4}	0.43/ 0.50	$6 \times 10^{-4}/$ 1×10^{-4}
---------------	---	---------------	---

* Set to align with GPU CUDA warps, avoiding underlying padding computation overhead caused by arbitrary discrete integers.

4.3.2. Model Analysis and Performance Evaluation

(1) Computational and communication costs

During training, the proposed model follows a federated learning paradigm with periodic communication and local updates driven by a streaming data mechanism. In each communication round, local training is performed on edge clients, while global aggregation is conducted at the server according to the designed personalized federated optimization strategy. In contrast, the inference stage is fully decoupled from communication and is performed solely through forward propagation of the trained local model.

To evaluate the computational and deployment characteristics of the proposed framework, several efficiency-related metrics are reported, including parameter size, floating-point operations (FLOPs), training time per communication round, inference latency, and communication cost. The model contains approximately 0.06–0.26 M trainable parameters across different seasonal configurations. The computational complexity is estimated at 1.62–6.13 MFLOPs per inference. The average training time per federated round ranges from 0.43 s to 0.71 s under the current hardware configuration, while inference latency is approximately 0.020–0.035 ms. The communication cost per round is determined by the model parameter size, ranging from approximately 0.24 MB to 0.95 MB depending on the specific configuration. The total communication load is jointly determined by the per-round communication cost and the number of federated communication rounds under the streaming update mechanism. In this study, the numbers of communication rounds are 110, 116, and 46 for the wet, dry, and normal seasons. These results are reported in detail in Table 4, covering parameter size, FLOPs, training time, inference latency, and communication cost under the same experimental configuration.

Table 4. Model efficiency and complexity analysis.

	Model	Total Params (M)	Upload per Round (MB)	Total Traffic (MB)	Inference Latency (ms)	FLOPs (MFLOPs)	Runtime (s)
Wet	FedAvg	0.0665	0.2538	223.3334	0.0245	1.6191	22.08
	FedProx	0.1213	0.4628	407.2397	0.0202	5.4555	31.97
	FedPer	0.0703	0.2432	213.9844	0.0207	1.6229	21.08
	OurWork	0.2611	0.9551	840.4688	0.0212	6.1269	62.59
Dry	FedAvg	0.0789	0.3012	279.4684	0.0212	5.221	27.34
	FedProx	0.1251	0.4773	442.9333	0.021	5.4593	44.45
	FedPer	0.1213	0.4521	419.5938	0.0201	5.4555	30.92
	OurWork	0.1213	0.4521	419.5938	0.035	5.4555	82.36
Normal	FedAvg	0.0665	0.2538	93.394	0.0378	1.6191	14.73
	FedProx	0.1213	0.4628	170.3002	0.0388	5.4555	20.12
	FedPer	0.0324	0.1167	42.9453	0.0362	1.4502	15.32
	OurWork	0.0665	0.2432	89.4844	0.0199	1.6191	19.92

(2) Residual analysis in test set

The residual plots for the test sets of the wet, dry, and normal season models are generated using the Freedman–Diaconis rule to determine the optimal number of bins, with equal-width binning applied to the predicted values. The results are sequentially presented in Figures 6–8. In these figures, the vertical axis denotes the residual values

(predicted minus actual), the horizontal axis represents the predicted wind power output, and the subplots correspond to Station A, Station B, Station C, and Station D, respectively.

As shown in Figures 6–8, the residuals for each wind farm are generally distributed around the zero line. The residual trend curves, representing conditional mean errors, fluctuate locally but remain largely centered around zero across most of the power range. In some cases, deviations from zero are observed in the high-power region, where slight overestimation or underestimation may occur. This behavior is more noticeable in extreme operating conditions, which are relatively less frequently observed in the dataset. Nevertheless, these local deviations are limited in magnitude, typically within approximately 5–10 MW, and do not alter the overall balanced distribution pattern of the residuals. For most farms, the mean errors, i.e., the error trend lines, lie within ± 5 MW. Only Station D in the dry season and Station C in the normal season exhibit mean errors approaching ± 10 MW due to fewer samples in the high-power range. However, relative to the corresponding maximum wind power output of approximately 170 MW, the errors remain within 6%, demonstrating sound forecasting accuracy. The 95% confidence intervals for most farms fall within a deviation of ± 20 MW, corresponding to a range of 10.5% to 16.7% relative to each farm's maximum output (ranging from approximately 120 MW to 190 MW), meeting the national standards [36], which require an accuracy rate exceeding 83% for short-term (3-day) forecasting.

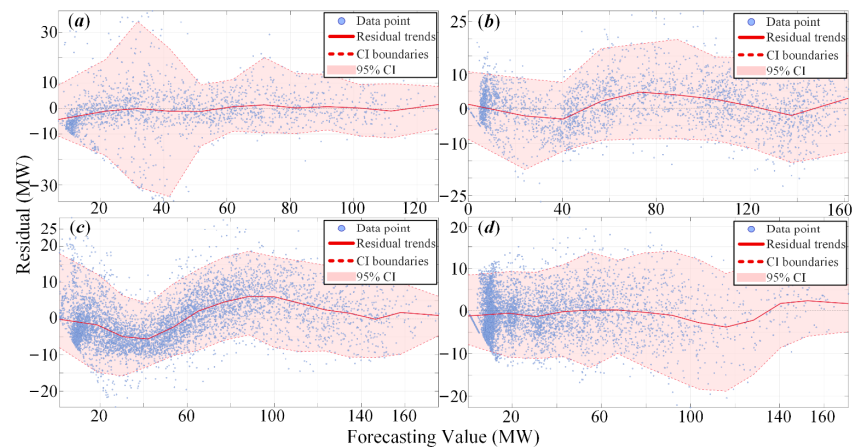


Figure 6. Forecasting residual distribution on the test set (wet season).

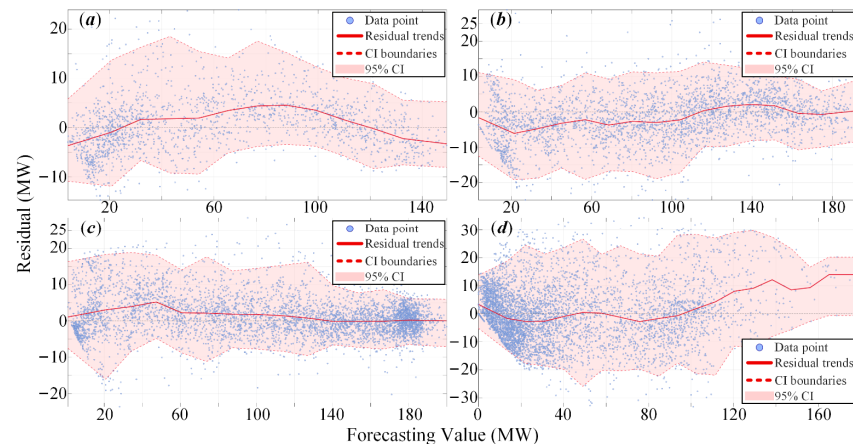


Figure 7. Forecasting residual distribution on the test set (dry season).

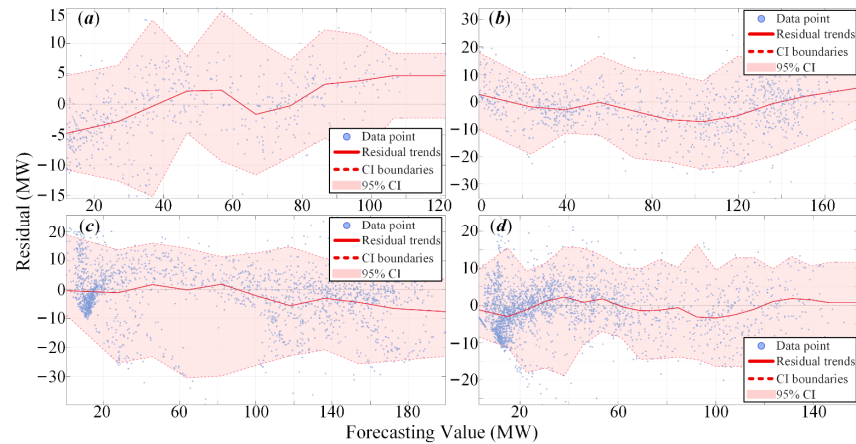


Figure 8. Forecasting residual distribution on the test set (normal season).

(3) Model Robustness and Stability

To reduce the randomness introduced by the BO process and neural network training, three independent experiments are conducted using different random seeds. The mean and standard deviation of evaluation metrics (R^2 , RMSE, MAE, and CR) are reported across all wind farms and seasons. The results in Table 5 show high consistency across independent runs, with generally small standard deviations. Specifically, the standard deviation of R^2 is less than 0.05, the RMSE exhibits limited variation (most values below 1), and the standard deviation of CR is within 0.02. These results indicate that the proposed model is stable under different random initializations and training perturbations, demonstrating robustness to stochastic factors.

From Table 5, Station D shows relatively higher forecasting errors in the dry season, which may be associated with its relatively isolated spatial location (over 90 km from the other stations). Station C exhibits higher variability in certain cases, which may be related to its higher-altitude terrain and more complex local meteorological conditions. Stations A and B also show moderate performance differences, potentially due to differences in data duration and geographical characteristics.

Table 5. Mean and standard deviation of forecasting performance in test set.

Period	Station	R^2 (Mean $\pm$ std)	RMSE (Mean $\pm$ std)	MAE (Mean $\pm$ std)	CR (Mean $\pm$ std)
Wet	A	0.923 $\pm$ 0.011	8.82 $\pm$ 0.392	5.931 $\pm$ 0.326	0.933 $\pm$ 0.004
	B	0.942 $\pm$ 0.021	11.147 $\pm$ 2.040	7.749 $\pm$ 2.143	0.937 $\pm$ 0.012
	C	0.934 $\pm$ 0.050	9.682 $\pm$ 3.474	7.181 $\pm$ 1.993	0.945 $\pm$ 0.020
	D	0.954 $\pm$ 0.004	6.856 $\pm$ 0.173	4.771 $\pm$ 0.082	0.962 $\pm$ 0.001
Dry	A	0.947 $\pm$ 0.013	9.293 $\pm$ 0.945	6.687 $\pm$ 1.038	0.937 $\pm$ 0.006
	B	0.976 $\pm$ 0.002	8.487 $\pm$ 0.428	6.313 $\pm$ 0.422	0.960 $\pm$ 0.002
	C	0.984 $\pm$ 0.002	7.539 $\pm$ 0.377	5.195 $\pm$ 0.508	0.963 $\pm$ 0.001
	D	0.897 $\pm$ 0.013	14.365 $\pm$ 0.912	11.278 $\pm$ 0.516	0.923 $\pm$ 0.006
Normal	A	0.916 $\pm$ 0.015	10.467 $\pm$ 0.513	7.258 $\pm$ 0.319	0.926 $\pm$ 0.005
	B	0.941 $\pm$ 0.008	12.660 $\pm$ 1.029	9.600 $\pm$ 0.527	0.933 $\pm$ 0.005
	C	0.958 $\pm$ 0.006	11.922 $\pm$ 0.723	9.005 $\pm$ 0.712	0.939 $\pm$ 0.004
	D	0.951 $\pm$ 0.008	7.968 $\pm$ 0.722	6.024 $\pm$ 0.574	0.953 $\pm$ 0.004

To further evaluate the statistical significance of the performance improvements, the Wilcoxon signed-rank test was conducted on the absolute forecasting errors between the proposed method and the four baseline models over the entire test set. The results indicate statistically significant improvements in the vast majority of spatiotemporal scenarios,

with most comparisons yielding p -values below 0.001. Only a few comparisons involving Station C and the personalized baselines (FedPer and FedProx) during the wet and normal seasons do not reach statistical significance ($p > 0.05$), which may be attributed to the higher-altitude location and more complex local meteorological conditions of Station C. Under such conditions, personalized federated learning methods are able to capture site-specific characteristics more effectively, resulting in forecasting errors comparable to those of the proposed method. Overall, these results demonstrate that the proposed method achieves statistically significant performance improvements in the vast majority of seasonal and spatial operating conditions.

(4) Ablation study and component analysis

To evaluate the impact of each component on the overall model performance, we constructed four ablation variants. Model I replaces the physics-guided aggregation mechanism with FedAvg. Model II enables parameter sharing in the private layer to evaluate the effect of the shared-private architecture. Model III replaces the asymmetric weighted loss with standard mean squared error (MSE) and removes the proximal term. Model IV removes the multi-head attention mechanism in the client-side shared layer.

Table 6 shows the forecasting results of the complete model and the four ablated variants across four wind farms and three seasonal settings. Overall, the complete model achieves the best performance in most cases. Each component corresponds to a different stage of the federated learning pipeline, leading to distinct degradation patterns under ablation. Specifically, the physics-guided aggregation mechanism governs cross-site knowledge sharing, the private layer preserves local characteristics, the asymmetric loss improves optimization under complex error distributions, and the multi-head attention enhances temporal feature learning.

Table 6. Results of ablation studies.

Model	Wet Season		Dry Season		Normal Season	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
I (Station A–D)	0.6774	18.3635	0.9369	10.053	0.8211	15.8786
	0.9031	14.5189	0.8316	22.3139	0.7665	25.0411
	0.9615	7.6237	0.924	16.3849	0.9447	13.3307
	0.8947	10.2827	0.8673	15.9062	0.919	10.238
II (Station A–D)	0.8884	10.8012	0.9405	9.7663	0.8782	13.1042
	0.9108	13.9306	0.8498	21.0732	0.9603	10.3242
	0.941	9.4401	0.9413	14.4042	0.9317	14.8217
	0.8052	13.9833	0.8253	18.2457	0.8355	14.5907
III (Station A–D)	0.6803	18.2813	0.9518	8.7895	0.8447	14.7929
	0.9479	10.6469	0.9717	9.1534	0.9528	11.2635
	0.9343	9.961	0.9145	17.3772	0.9731	9.2927
	0.9395	7.7893	0.8626	16.1829	0.904	11.1431
IV (Station A–D)	0.9025	10.094	0.9510	8.8635	0.908	11.3859
	0.9589	9.4512	0.9640	10.3206	0.958	10.6214
	0.9697	6.7649	0.9625	11.5047	0.9658	10.4789
	0.9329	8.2049	0.8954	14.118	0.8994	11.41
Full (Station A–D)	0.923	8.82	0.947	9.293	0.916	10.467
	0.942	11.147	0.976	8.487	0.941	12.660
	0.934	9.68	0.984	7.539	0.958	11.922
	0.954	6.85	0.897	14.365	0.951	7.968

In particular, Model I exhibits the most significant degradation across wind farms. For example, during the wet season, the R^2 at Station A decreases by approximately 25%,

while RMSE nearly doubles; Stations B and D show around 5% R^2 degradation. This indicates that FedAvg, relying solely on statistical weighting, has limitations under strong spatial heterogeneity. Stations A and B have approximately 1 year and 1.6 years of data, while Stations C and D each have about 3 years of observations, leading the global model to be biased toward data-rich clients. Moreover, the four stations are located in distinct micro-meteorological environments, including valley wind regions, monsoon-affected areas, alpine canyon terrain, and narrow valleys, resulting in strong spatial variability. Under such conditions, purely data-driven aggregation struggles to capture cross-site relationships, causing degraded performance on data-scarce and highly local stations. In contrast, the proposed physics-guided aggregation mechanism leverages distance, terrain, and real-time wind direction to establish a dynamic weight matrix, thereby mitigating negative transfer and improving overall performance.

In comparison, Model II shows that removing the private layer leads to clear station-dependent degradation. In particular, Station D, located in the Anning River Valley, is affected by channeling effects, wind shear, and thermal contrasts, with R^2 decreasing from above 0.9 to below 0.85 and RMSE increasing to nearly twice its original size. This indicates that the private layer helps preserve site personalization on top of shared knowledge.

Furthermore, Models III and IV correspond to the optimization and temporal feature refinement stages in the federated learning pipeline, respectively. Replacing the asymmetric loss with standard MSE leads to performance degradation, more evident under complex wind conditions or at specific stations. For example, in the dry season, Station C shows an R^2 decrease of approximately 7%, while in the wet season, Station A decreases by about 25%. This suggests that the asymmetric loss improves optimization stability under complex and asymmetric error distributions. After removing the multi-head attention mechanism, the CNN-LSTM backbone still maintains competitive performance, with only moderate degradation under challenging conditions.

Overall, the four components correspond to knowledge sharing, personalized modeling, optimization, and temporal feature refinement stages in the federated learning pipeline. They exhibit distinct degradation patterns under ablation. The complete model consistently achieves the best performance across all stations and seasons, indicating that these components are complementary and jointly enhance the model's ability to capture spatial heterogeneity and temporal dynamics in complex mountainous wind farm environments.

4.3.3. Comparison and Analysis of Forecasting Results

This study performs short-term wind power forecasting for 72 h (288 steps). The 72 h (3-day) forecasting horizon is adopted because it is directly relevant to practical power system operations, including unit commitment, generation scheduling, maintenance planning, and electricity market participation, while also providing a more challenging benchmark due to the increasing uncertainty associated with longer forecasting horizons. For typical scenarios, K-means clustering is first applied to historical wind power time series of each station to identify representative diurnal patterns, from which a cluster-based representative curve over a 3-day horizon is obtained. A sliding window is then applied over the full chronological test set, and the continuous 3-day segment with the minimum RMSE to the representative curve is selected as the typical scenario.

Moreover, extreme scenarios are identified based on data variability. The variance of ground-truth wind power is computed within each 3-day sliding window in the test set, and the segment with the highest variance is selected to represent high-volatility conditions with rapid ramps and large fluctuations. The selection is performed using only statistical properties of the test set prior to model evaluation, ensuring consistency with the intrinsic data distribution.

The results are shown in Figures 9–14 (the subplots (a)–(d) correspond to Station A–D, respectively) and Tables 7–12, using R^2 , RMSE, MAE, and the forecasting accuracy rate (C_R) defined by the national standard [36], calculated by Equation (12). C_i denotes the operating capacity of the wind farm at time i .

$$C_R = 1 - \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{C_i} \right)^2} \quad (12)$$

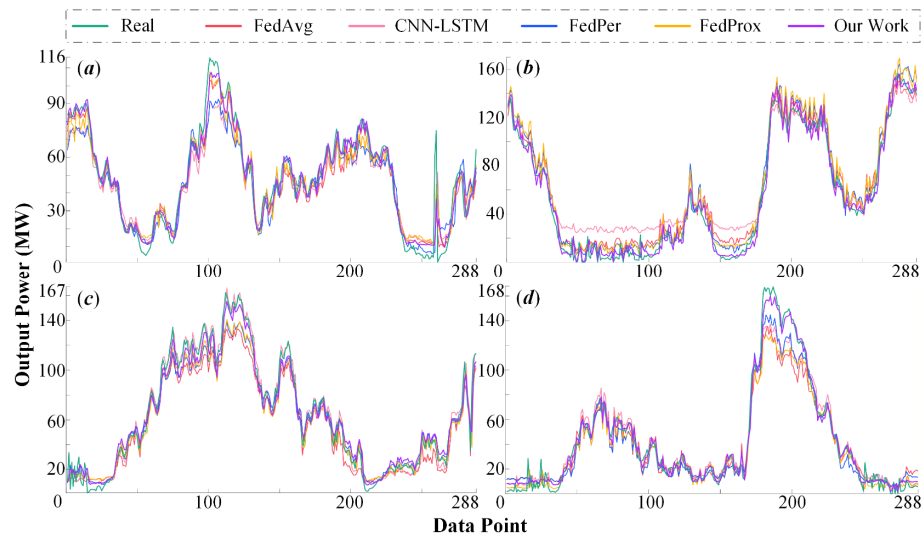


Figure 9. Forecasting results of different models under a typical scenario (wet season).

Table 7. Forecasting performance comparison under a typical scenario in the wet season.

Model		R^2	RMSE	MAE	C_R
Station A	Base CNN-LSTM	0.8633	10.2199	7.6352	92.18%
	Base FedAvg	0.9286	7.3865	5.3794	94.35%
	Base FedPer	0.8671	10.0775	7.3059	92.29%
	Base FedProx	0.9179	7.9190	5.7925	93.94%
	Our Work	0.9457	6.4402	4.3693	95.07%
Station B	Base CNN-LSTM	0.9106	14.6176	12.1742	91.66%
	Base FedAvg	0.9595	9.8407	8.5087	94.39%
	Base FedPer	0.9620	9.5225	8.0464	94.57%
	Base FedProx	0.9519	10.7156	9.0816	93.89%
	Our Work	0.9818	6.5870	5.0817	96.24%
Station C	Base CNN-LSTM	0.9763	6.8507	5.1675	96.13%
	Base FedAvg	0.9308	11.6993	9.2457	93.38%
	Base FedPer	0.9547	9.4652	7.2280	94.65%
	Base FedProx	0.9516	9.7791	7.4381	94.47%
	Our Work	0.9727	7.3441	5.7224	95.85%
Station D	Base CNN-LSTM	0.9452	10.2068	7.0642	94.41%
	Base FedAvg	0.9104	13.0527	8.8673	92.86%
	Base FedPer	0.9469	10.0485	7.6760	94.50%
	Base FedProx	0.9253	11.9207	7.3544	93.48%
	Our Work	0.9824	5.7837	4.3928	96.83%

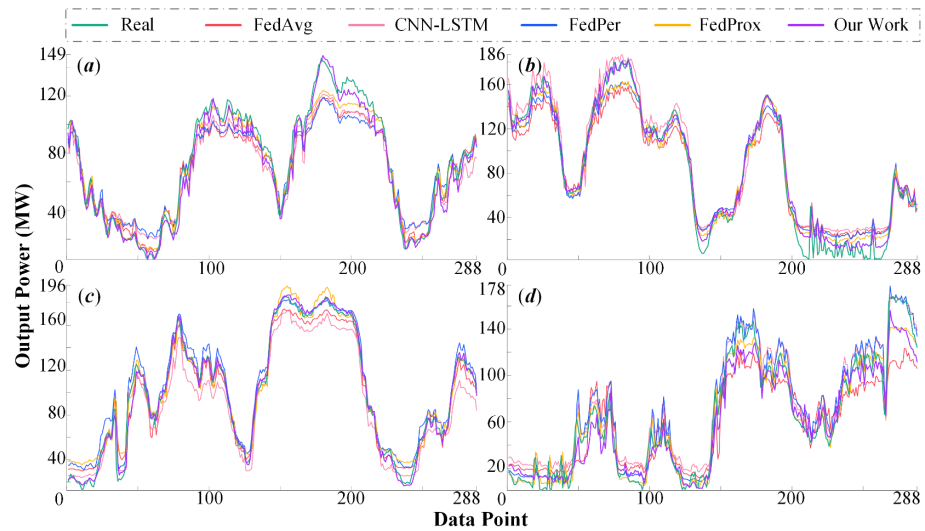


Figure 10. Forecasting results of different models under a typical scenario (dry season).

Table 8. Forecasting performance comparison under a typical scenario in the dry season.

	Model	R^2	RMSE	MAE	C_R
Station A	Base CNN-LSTM	0.9040	12.2495	10.7710	91.72%
	Base FedAvg	0.9201	11.1789	8.8623	92.44%
	Base FedPer	0.9018	12.3928	10.2742	91.62%
	Base FedProx	0.9633	7.5786	5.7126	94.88%
	Our Work	0.9675	7.1273	5.5022	95.18%
Station B	Base CNN-LSTM	0.9458	13.2665	10.4972	93.69%
	Base FedAvg	0.9354	14.4811	12.2829	93.11%
	Base FedPer	0.9560	11.9597	9.7280	94.31%
	Base FedProx	0.9691	10.0261	7.8582	95.23%
	Our Work	0.9798	8.0964	6.0432	96.15%
Station C	Base CNN-LSTM	0.9310	13.7954	11.8915	93.10%
	Base FedAvg	0.9718	8.8206	6.8949	95.59%
	Base FedPer	0.9693	9.2002	7.3099	95.40%
	Base FedProx	0.9669	9.5570	6.8518	95.22%
	Our Work	0.9886	5.6182	3.5860	97.19%
Station D	Base CNN-LSTM	0.9270	13.2096	11.2029	93.04%
	Base FedAvg	0.8302	20.1470	16.2396	89.38%
	Base FedPer	0.9367	12.3016	10.0214	93.52%
	Base FedProx	0.9582	9.9982	7.0992	94.73%
	Our Work	0.9346	12.5055	9.9501	93.41%

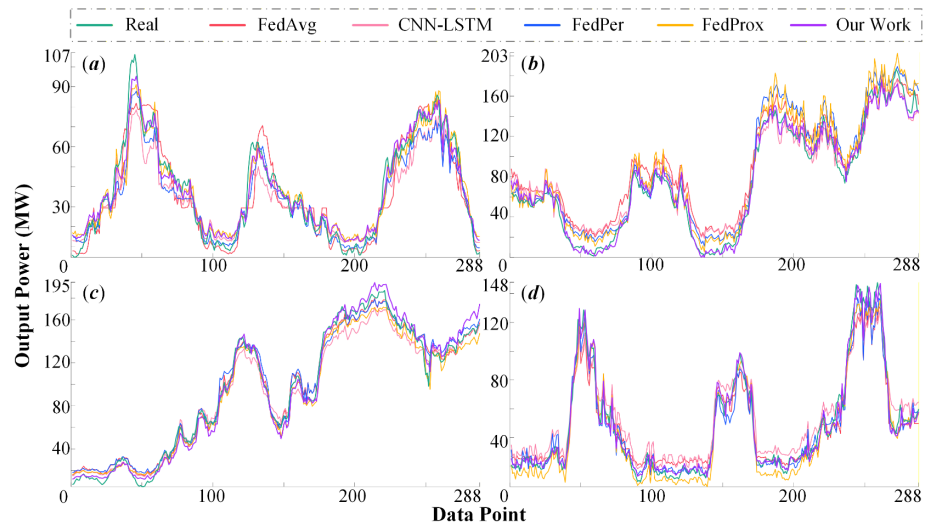


Figure 11. Forecasting results of different models under a typical scenario (normal season).

Table 9. Forecasting performance comparison under a typical scenario in the normal season.

	Model	R^2	RMSE	MAE	C_R
Station A	Base CNN-LSTM	0.8641	9.0029	7.0488	93.49%
	Base FedAvg	0.8721	8.7346	6.3629	93.68%
	Base FedPer	0.9061	7.4820	5.4653	94.59%
	Base FedProx	0.9356	6.1948	4.8461	95.52%
	Our Work	0.9496	5.4807	4.3782	96.04%
Station B	Base CNN-LSTM	0.9369	13.2622	11.2405	93.01%
	Base FedAvg	0.8988	16.8019	15.1564	91.14%
	Base FedPer	0.9199	14.9458	12.5309	92.12%
	Base FedProx	0.9262	14.9458	11.9765	92.44%
	Our Work	0.9806	7.3478	5.7389	96.12%
Station C	Base CNN-LSTM	0.9725	9.4731	7.6320	95.05%
	Base FedAvg	0.9868	6.5580	5.0528	96.58%
	Base FedPer	0.9814	7.6312	6.2720	95.93%
	Base FedProx	0.9822	7.6312	6.3856	96.02%
	Our Work	0.9848	7.0534	5.6105	96.32%
Station D	Base CNN-LSTM	0.9022	11.4710	9.8440	93.27%
	Base FedAvg	0.9310	9.6352	7.2702	94.35%
	Base FedPer	0.9491	8.2737	5.8577	95.15%
	Base FedProx	0.9559	7.7068	6.2159	95.48%
	Our Work	0.9733	5.9953	4.3442	96.48%

Seasonal variations across wet, normal, and dry periods induce heterogeneous spatiotemporal distribution shifts in wind power data, exhibiting distinct statistical characteristics across different seasonal partitions. These regimes represent typical operating conditions of the wind power system, characterized by different levels of variability and spatial concentration, leading to non-stationary and uneven data distributions among wind farms. Such heterogeneity increases the difficulty of achieving stable forecasting performance under privacy-preserving constraints.

Under these typical scenarios, FedAvg exhibits noticeable performance degradation, as uniform aggregation tends to obscure station-specific patterns, particularly in sites with distinct spatiotemporal characteristics. Although FedPer and FedProx introduce person-

alization or proximal regularization to improve training stability, their performance still varies across seasonal regimes, indicating limited adaptability under pronounced distribution shifts. In contrast, the proposed framework differs from FedAvg, FedPer, and FedProx in three aspects: it adopts a physically guided aggregation strategy instead of uniform or static weighting, employs a structured separation between shared and client-specific components rather than fully shared or shallowly split architectures, and introduces an asymmetric loss function with a mild proximal regularization term to better handle output-sensitive forecasting regions.

Across all evaluated typical scenarios, the proposed method consistently achieves better forecasting performance among all compared approaches, maintaining R^2 values above 0.9346 (most R^2 values exceed 0.95, and the average R^2 is 0.970), C_R values exceeding 93.41%, and the average C_R (calculated as the arithmetic mean of C_R values for different stations and periods) improves by 1.37–2.69% compared with the baseline methods, which is calculated as $(C_{R,ours} - C_{R,base})/C_{R,base}$. Compared to the baseline methods, it consistently reduces the average RMSE. Specifically, it achieves a relative reduction ranging from approximately 24.97% to 38.21%, which is calculated as $(RMSE_{base} - RMSE_{ours})/RMSE_{base}$.

This stable improvement pattern, rather than isolated gains in specific scenarios, indicates consistent effectiveness in handling heterogeneous spatiotemporal conditions within the scope of this study.

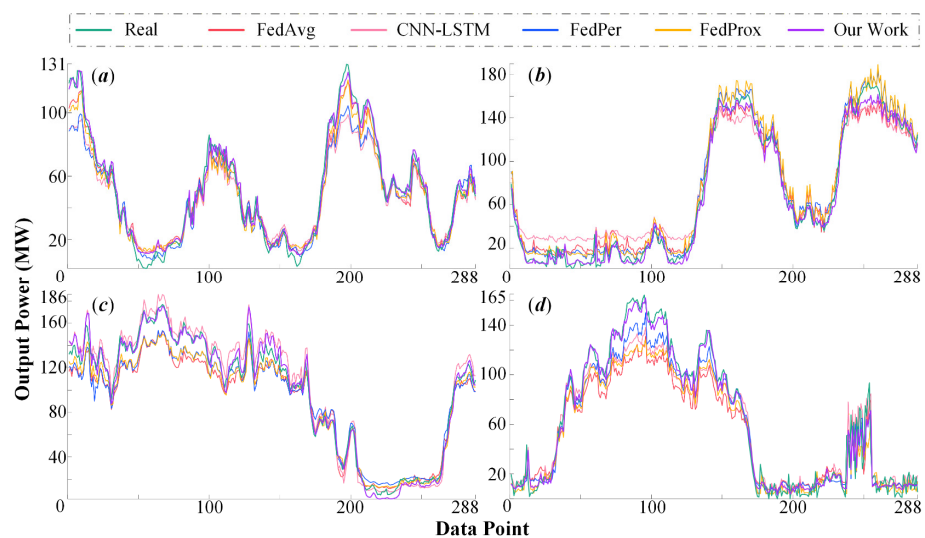


Figure 12. Forecasting results of different models under extreme scenarios (wet season).

Table 10. Forecasting performance comparison under extreme scenarios in the wet season.

Model		R^2	RMSE	MAE	C_R
Station A	Base CNN-LSTM	0.8845	11.5217	8.6120	91.18%
	Base FedAvg	0.9605	6.7419	5.2800	94.84%
	Base FedPer	0.9066	10.3622	7.5028	92.07%
	Base FedProx	0.9511	7.4995	5.9874	94.26%
	Our Work	0.9766	5.1811	4.0664	96.03%
Station B	Base CNN-LSTM	0.9380	14.7680	12.8420	91.58%
	Base FedAvg	0.9691	10.4223	8.5600	94.05%
	Base FedPer	0.9759	9.2029	7.5439	94.75%
	Base FedProx	0.9677	10.6571	8.9432	93.92%
	Our Work	0.9874	6.6459	4.9032	96.21%
Station C	Base CNN-LSTM	0.9664	9.4911	7.9177	94.63%

Station D	Base FedAvg	0.9317	13.5255	10.9806	92.35%
	Base FedPer	0.9401	12.6692	10.5155	92.84%
	Base FedProx	0.9530	11.2175	8.7337	93.66%
	Our Work	0.9856	6.2034	4.6766	96.49%
	Base CNN-LSTM	0.9157	15.5487	11.3763	91.49%
	Base FedAvg	0.8675	19.4918	15.1641	89.33%
	Base FedPer	0.9411	12.9971	9.7820	92.89%
	Base FedProx	0.9008	16.8646	12.2200	90.77%
	Our Work	0.9707	9.1609	6.0262	94.99%

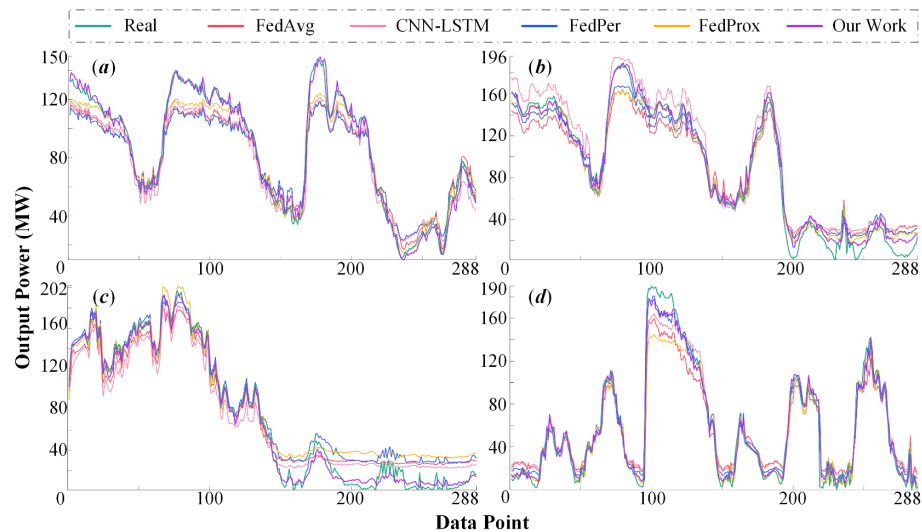


Figure 13. Forecasting results of different models under extreme scenarios (dry season).

Table 11. Forecasting performance comparison under extreme scenarios in the dry season.

	Model	R^2	RMSE	MAE	C_R
Station A	Base CNN-LSTM	0.9134	11.5629	9.8895	92.18%
	Base FedAvg	0.9036	12.2002	10.0462	91.75%
	Base FedPer	0.8842	13.3714	11.2466	90.96%
	Base FedProx	0.9568	8.1628	5.9270	94.48%
	Our Work	0.9845	4.8878	3.6329	96.69%
Station B	Base CNN-LSTM	0.9376	14.8424	12.8756	92.94%
	Base FedAvg	0.9305	15.6600	13.8068	92.55%
	Base FedPer	0.9591	12.0177	10.3312	94.28%
	Base FedProx	0.9710	10.1213	7.5667	95.18%
	Our Work	0.9837	7.5803	5.8562	96.39%
Station C	Base CNN-LSTM	0.9418	16.0446	14.4877	91.98%
	Base FedAvg	0.9476	15.2184	12.9032	92.39%
	Base FedPer	0.9438	15.7720	12.5704	92.12%
	Base FedProx	0.9189	18.9385	15.0475	90.53%
	Our Work	0.9926	5.7070	4.5490	97.15%
Station D	Base CNN-LSTM	0.9379	13.2241	10.1188	93.03%
	Base FedAvg	0.9193	15.0702	11.7576	92.06%
	Base FedPer	0.9643	10.0240	7.5334	94.72%
	Base FedProx	0.9145	15.5153	9.5048	91.82%
	Our Work	0.9692	9.3018	6.6586	95.10%

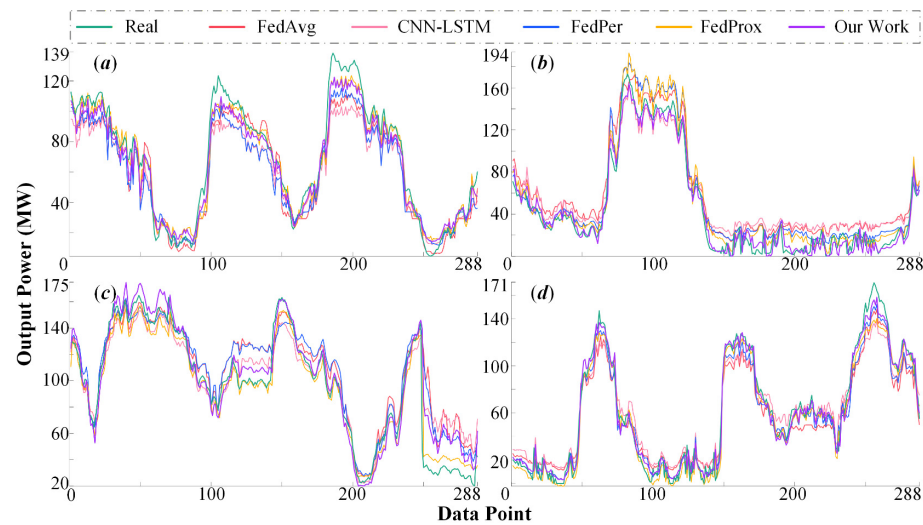


Figure 14. Forecasting results of different models under extreme scenarios (normal season).

Table 12. Forecasting performance comparison under extreme scenarios in the normal season.

	Model	R^2	RMSE	MAE	C_R
Station A	Base CNN-LSTM	0.8466	14.7252	11.4233	89.35%
	Base FedAvg	0.8928	12.3087	9.5931	91.10%
	Base FedPer	0.8822	12.9019	10.7062	90.67%
	Base FedProx	0.9555	7.9281	6.0650	94.27%
	Our Work	0.9503	8.3801	6.6224	93.94%
Station B	Base CNN-LSTM	0.8851	16.2257	14.4010	91.44%
	Base FedAvg	0.8716	17.1491	15.6191	90.96%
	Base FedPer	0.9117	14.2223	11.4601	92.50%
	Base FedProx	0.9289	12.7627	10.0863	93.27%
	Our Work	0.9730	7.8579	6.0783	95.86%
Station C	Base CNN-LSTM	0.8488	16.6203	11.6452	91.32%
	Base FedAvg	0.8163	18.3207	12.0289	90.44%
	Base FedPer	0.8366	17.2829	12.3573	90.98%
	Base FedProx	0.9526	9.3070	6.6284	95.14%
	Our Work	0.9093	12.8724	8.8590	93.28%
Station D	Base CNN-LSTM	0.9202	13.5020	11.3580	92.08%
	Base FedAvg	0.9250	13.0904	11.3027	92.32%
	Base FedPer	0.9589	9.6928	7.9932	94.31%
	Base FedProx	0.9610	9.4383	6.9311	94.46%
	Our Work	0.9786	6.9973	5.3447	95.60%

Under extreme scenarios, the wind power system exhibits intensified spatiotemporal fluctuations, where rapid meteorological variations further amplify distribution shifts across wind farms. Compared with typical operating conditions, this highly non-stationary setting imposes more stringent requirements on the robustness and stability of forecasting models, effectively serving as a stress test for federated learning under severe data heterogeneity.

In this context, traditional federated baselines experience noticeable performance deterioration. For instance, FedAvg suffers from significant performance degradation due to aggregation inconsistency under extreme distribution shifts, with its RMSE increasing sharply to 19.4918 and 15.0702 at Station D in distinct extreme events. This suggests that uniform aggregation becomes less reliable when local data distributions diverge substan-

tially across clients. While advanced variants such as FedPer and FedProx partially alleviate this issue through personalization and proximal regularization, their performance remains highly sensitive to extreme variations, indicating a limited ability to maintain consistent behavior under severe non-stationarity.

Under certain localized extreme variations, FedProx occasionally achieves competitive or slightly better accuracy at individual stations (e.g., yielding marginally lower RMSEs than the proposed method at Stations A and C in Table 10). This phenomenon suggests that stronger proximal constraints may improve local adaptation in specific high-deviation scenarios, but this comes at the cost of reduced consistency across other stations within the same period.

Compared with these baselines, the proposed framework maintains relatively stable forecasting performance under extreme conditions, showing limited sensitivity to abrupt distribution shifts across clients. Across all evaluated extreme scenarios, the method maintains R^2 values above 0.9093 (most R^2 values exceed 0.95, and the average R^2 is 0.972) and C_R values above 93.41%. In terms of RMSE, it achieves relative reductions ranging from approximately 33.87% to 47.24% compared with baseline models, while the average C_R improves by 2.3–4%.

These results indicate that, although extreme distribution shifts may induce localized performance fluctuations, the proposed model is able to avoid severe degradation and maintain relatively stable predictive performance across stations within the scope of the conducted extreme scenarios.

Across both typical and extreme scenarios, the proposed method demonstrates improved forecasting performance compared with all baseline methods. Across different seasons and wind farms, the average RMSE, average R^2 , and average C_R are all improved, while effectively mitigating the performance degradation of conventional federated learning methods under extreme distribution shifts. The results from both typical and extreme scenario evaluations demonstrate that the proposed method improves forecasting performance under data privacy constraints across multiple wind farms in the complex meteorological environment, providing a useful reference for collaborative wind power forecasting in mountainous regions with complex meteorological conditions.

5. Conclusions

Addressing the demand for data privacy protection and cross-farm collaborative forecasting of multiple wind farms in complex mountainous terrain, this paper proposes a physics-guided personalized spatiotemporal federated hybrid forecasting model. Compared with baseline models, the proposed model achieves significantly higher multi-farm wind power forecasting accuracy while preserving data privacy, advancing the sustainable deployment of distributed wind energy systems.

Actual operational data from four mountainous wind farms, including historical meteorological variables and wind power output, are first collected and processed through anomaly removal and normalization to form local training sets. A CNN-LSTM-MHA forecasting network consisting of a shared layer and a personalized private layer is then designed for each client. The risk-aware asymmetric training strategy comprises an asymmetric load-weighted loss function, which captures the forecasting risk escalation with output by imposing increasing penalties on the high-power range, and a proximal term constraining local updates from deviating. On this basis, a federated learning framework integrating the shared layer and the personalized private layer is established, where the shared layer extracts regional meteorological patterns reflected in shared-layer parameters and the private layer fits local microclimate residuals. Furthermore, a physics-guided dynamic weight matrix is designed, incorporating spatial parameters such as geographical distance, terrain blocking, and real-time wind direction alignment, and performing

personalized parameter aggregation through daily communication rounds. This design avoids the high interaction costs and privacy risks of graph neural networks under communication-constrained mountainous conditions.

The validation results for the datasets from the four wind farms demonstrate that the proposed model outperforms the baseline methods in both forecasting accuracy and overall performance. Specifically, all forecasting cases achieve an average R^2 exceeding 0.95, while the average RMSE is reduced by more than 24% compared with the baseline methods. It effectively integrates physics-guided cross-farm regional meteorological patterns reflected in shared-layer parameters and mitigates non-IID heterogeneity while keeping raw data physically isolated. These findings may provide useful insights for wind power forecasting in similar mountainous environments under data privacy constraints.

Several limitations should also be acknowledged. First, the proposed model relies on the availability of key meteorological variables, particularly wind direction, and its forecasting performance may decrease when these inputs become inaccurate or unavailable. Second, the collaborative performance gain may diminish when participating wind farms are geographically distant and exhibit weak physical or meteorological correlations. While the designed physics-guided weight matrix effectively captures spatial topologies, it remains a simplified mathematical approximation of the highly non-linear, three-dimensional aerodynamic turbulence and wake effects inherent to complex mountainous environments. Third, although the proposed model performs well on the current wind farm cluster, further improvements in communication efficiency and system robustness are still required to support substantially larger client populations. In addition, the present study evaluates the proposed model under an observed-meteorology setting. Future investigations incorporating operational numerical weather prediction (NWP) products are expected to provide a more practical assessment of end-to-end forecasting performance.

Future research may further investigate dynamic transfer mechanisms involving additional physical factors, validate the proposed model across broader geographical regions, climatic conditions, and terrain types, systematically compare different forecasting horizons, investigate NWP uncertainty modeling, and extend the proposed approach to photovoltaic and hybrid renewable energy forecasting.

Author Contributions: Conceptualization, Y.Y. and Y.X.; writing—original draft preparation, Y.Y. and Y.X.; writing—review and editing, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The authors do not have permission to share data. The data are not publicly available due to privacy restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. National Energy Administration. Grid-Connected Operation of Renewable Energy in 2025. Available online: <https://www.nea.gov.cn/20260212/742b8c6a078347b0b39de676c05c5d58/c.html> (accessed on 5 May 2026). (In Chinese)
2. Dolatabadi, A.; Mohammadi-ivatloo, B.; Abapour, M.; Tohidi, S. Optimal stochastic design of wind integrated energy hub. *IEEE Trans. Ind. Inform.* **2017**, *13*, 2379–2388.
3. Arora, P.; Jalali, S.M.J.; Ahmadian, S.; Panigrahi, B.K.; Suganthan, P.N.; Khosravi, A. Probabilistic wind power forecasting using optimized deep auto-regressive recurrent neural networks. *IEEE Trans. Ind. Inform.* **2022**, *19*, 2814–2825.

4. Arjomandi-Nezhad, A.; Ahmadi, A.; Taheri, S.; Fotuhi-Firuzabad, M.; Moeini-Aghaie, M.; Lehtonen, M. Pandemic-aware day-ahead demand forecasting using ensemble learning. *IEEE Access* **2022**, *10*, 7098–7106.
5. Sweeney, C.; Bessa, R.J.; Browell, J.; Pinson, P. The future of forecasting for renewable energy. *Rev. Energy Environ.* **2020**, *9*, e365. <https://doi.org/10.1002/wene.365>.
6. Liu, X.; Lin, Z.; Feng, Z. Short-term offshore wind speed forecast by seasonal ARIMA—A comparison against GRU and LSTM. *Energy* **2021**, *227*, 120492.
7. González-Cagigal, M.A.; Rosendo-Macias, J.A.; Gómez-Expósito, A. Parameter estimation of wind turbines with PMSM using cubature Kalman filters. *IEEE Trans. Power Syst.* **2020**, *35*, 1796–1804.
8. Bae, D.J.; Kwon, B.S.; Song, K.B. XGBoost-based day-ahead load forecasting algorithm considering behind-the-meter solar PV generation. *Energies* **2021**, *15*, 128.
9. Joseph, L.P.; Deo, R.C.; Prasad, R.; Salcedo-Sanz, S.; Raj, N.; Soar, J. Near real-time wind speed forecast model with bidirectional LSTM networks. *Renew. Energy* **2023**, *204*, 39–58.
10. Zhao, J.; Wang, J.; Guo, Z.; Guo, Y.; Lin, W.; Lin, Y. Multi-step wind speed forecasting based on numerical simulations and an optimized stochastic ensemble method. *Appl. Energy* **2019**, *255*, 113833.
11. Tang, Q.W.; Xiang, Y.; Dai, J.K.; Li, Z.; Sun, W.; Liu, J. Wind farm power transfer forecasting method based on CNN-LSTM. *Adv. Eng. Sci.* **2024**, *56*, 91–99. <https://doi.org/10.15961/j.jsuese.202201165>. (In Chinese)
12. Li, Y.; Yang, Z.; Li, G.; Mu, Y.; Zhao, D.; Chen, C.; Shen, B. Optimal scheduling of isolated microgrid with an electric vehicle battery swapping station in multi-stakeholder scenarios: A bi-level programming approach via real-time pricing. *Appl. Energy* **2018**, *232*, 54–68.
13. National Energy Administration. Notice of the National Energy Administration on Issuing the Administrative Measures for Data Security in the Energy Sector (Trial). Available online: https://www.gov.cn/zhengce/zhengceku/202512/content_7051044.htm (accessed on 12 May 2026). (In Chinese)
14. Zhang, Y.; Wang, J. A distributed approach for wind power probabilistic forecasting considering spatio-temporal correlation without direct access to off-site information. *IEEE Trans. Power Syst.* **2018**, *33*, 5714–5726.
15. Gonçalves, C.; Bessa, R.J.; Pinson, P. Privacy-preserving distributed learning for renewable energy forecasting. *IEEE Trans. Sustain. Energy* **2021**, *12*, 1777–1787.
16. Zhang, T.; Zhu, Q. Dynamic differential privacy for ADMM-based distributed classification learning. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 172–187.
17. Sommer, B.; Pinson, P.; Messner, J.W.; Obst, D. Online distributed learning in wind power forecasting. *Int. J. Forecast.* **2021**, *37*, 205–223.
18. Gonçalves, C.; Bessa, R.J.; Pinson, P. A critical overview of privacy-preserving approaches for collaborative forecasting. *Int. J. Forecast.* **2021**, *37*, 322–342.
19. Bousbiat, H.; Bousselidj, R.; Himeur, Y.; Amira, A.; Bensaali, F.; Fadli, F.; Mansoor, W.; Elmenreich, W. Crossing roads of federated learning and smart grids: Overview, challenges, and perspectives. *arXiv* **2023**, arXiv:2304.08602.
20. Li, T.; Sahu, A.K.; Talwalkar, A.; Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process. Mag.* **2020**, *37*, 50–60.
21. Li, Y.; Wang, R.; Li, Y.; Zhang, M.; Long, C. Wind power forecasting considering data privacy protection: A federated deep reinforcement learning approach. *Appl. Energy* **2023**, *329*, 120291.
22. Ahmadi, A.; Talaei, M.; Sadipour, M.; Amani, A.M.; Jalili, M. Deep federated learning-based privacy-preserving wind power forecasting. *IEEE Access* **2023**, *11*, 39521–39530.
23. Lu, N.; Liu, S.; Wen, Q.; Chen, Q.; Sun, L.; Wang, Y. Federated domain separation for distributed forecasting of non-IID household loads. *IEEE Trans. Smart Grid* **2024**, *15*, 4271–4283.
24. Arooj, Q. FedWindT: Federated learning assisted transformer architecture for collaborative and secure wind power forecasting in diverse conditions. *Energy* **2024**, *309*, 133072.
25. Grataloup, A.; Jonas, S.; Meyer, A. A review of federated learning in renewable energy applications: Potential, challenges, and future directions. *Energy AI* **2024**, *17*, 100375.
26. Wang, Y.; Guo, Q. Privacy-preserving and adaptive federated deep learning for multiparty wind power forecasting. *IEEE Trans. Ind. Appl.* **2025**, *61*, 1352–1362.
27. Wang, H.; Yurochkin, M.; Sun, Y.; Papailiopoulos, D.; Khazaeni, Y. Federated learning with matched averaging. *arXiv* **2020**, arXiv:2002.06440. <https://doi.org/10.48550/arXiv.2002.06440>.

28. Blanchard, P.; El Mhamdi, E.M.; Guerraoui, R.; Stainer, J. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 119–129.
29. Wang, H.; Zheng, Z.Y.; Yan, J.; Zhang, N.; Zhang, J.; Liu, Y. A novel federated learning framework with adaptive parameter updater for wind power forecasting. *IEEE Trans. Sustain. Energy* 2025, *early access*.
30. Wang, L.; Fan, W.; Jiang, G.; Xie, P. An efficient federated transfer learning framework for collaborative monitoring of wind turbines in IoE-enabled wind farms. *Energy* 2023, *284*, 128518.
31. Stadtmann, F.; Furevik, E.R.; Rasheed, A.; Kvamsdal, T. Physics-guided federated learning as an enabler for digital twins. *Expert Syst. Appl.* 2024, *258*, 125169.
32. Chinese Academy of Environmental Planning. *Assessment of Wind and Solar Power Potential in China (2024)*; Chinese Academy of Environmental Planning: Beijing, China, 2024. (In Chinese)
33. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated optimization in heterogeneous networks. *Proc. Mach. Learn. Syst.* 2020, *2*, 429–450.
34. Arivazhagan, M.G.; Aggarwal, V.; Singh, A.K.; Choudhary, S. Federated learning with personalization layers. *arXiv* 2019, arXiv:1912.00818. <https://doi.org/10.48550/arXiv.1912.00818>.
35. Dinh, C.T.; Tran, N.; Nguyen, T.D. Personalized federated learning with Moreau envelopes. *Adv. Neural Inf. Process. Syst.* 2020, *33*, 21394–21405.
36. GB/T 19963.1-2021; Technical Specification for Connecting Wind Farm to Power System—Part 1: On Shore Wind Power. State Administration for Market Regulation; Standardization Administration of the P.R.C.: Beijing, China, 2021. (In Chinese)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.