A Review of Short-Term Photovoltaic Power Forecasting Driven by Big Data: From LSTM to Hybrid Intelligent Models

Ruijia Huang
Chongqing University of Posts and Telecommunications, International college, ChongQing, China

Abstract: The dual carbon goals are compelling power systems to accelerate the integration of high proportions of renewable energy. However, photovoltaic output fluctuates dramatically due to weather randomness, introducing significant uncertainty into dispatch, trading, and energy storage deployment. Over the past five years, deep learning models represented by LSTM have rapidly become the mainstream tool for short-term photovoltaic power forecasting (STPF) due to their superior ability to capture long-range dependencies in time series data. Concurrently, the big data environment has fostered a new generation of hybrid forecasting frameworks integrating "signal decomposition, feature extraction, and parameter optimization." This paper systematically reviews the latest global advancements in data infrastructure, model evolution, algorithm optimization, and engineering implementation. It distills a universal paradigm-"VMD/CNN/Attention-LSTM combined with Population-Based Intelligence"-while emphasizing that challenges such as the depth of multi-source data fusion, model interpretability, transferability across sites, and the implementation of online learning remain critical areas for future research.

Keywords: Photovoltaic Power Generation; Short-Term Power Forecasting; LSTM; Population-based Optimization Algorithms; Big Data

1. Research Background and Significance of the Project

The global energy structure is undergoing profound transformation centered on clean energy. In 2024, China's new photovoltaic (PV) installations reached 278 gigawatts, bringing the cumulative installed capacity to 886 gigawatts-a 45% year-on-year increase. However, PV output shows strong intermittency and non-stationarity due to the combined influence of cloud cover, temperature, aerosols, and other environmental factors, which directly threaten grid frequency and voltage stability [1]. According to the Guidelines for Power System Security and Stability, reducing prediction errors by 1% can equate to a 0.8% reduction in spinning reserve, resulting in annual coal cost savings of nearly 1.2 billion yuan. Therefore, enhancing day-ahead and intraday forecasting accuracy has become a critical bottleneck that the emerging power system urgently needs to address.

Traditional physical models rely on numerical weather prediction (NWP) and component physical parameters, offering high accuracy but requiring substantial computational resources. Statistical models like ARMA and SVM struggle with nonlinearity. With the proliferation of big data and AI technologies, end-to-end data-driven methods based on Long Short-Term Memory (LSTM) networks have demonstrated outstanding performance in 15-minute to 4-hour forecasting, reducing the Mean Absolute Percentage Error (MAPE) to 4%–6%. However, LSTMs exhibit sensitivity to hyperparameters and struggle to capture multi-channel meteorological features. To address this, researchers propose a hybrid framework combining "signal decomposition + CNN feature extraction + population optimization." This approach reduces RMSE by over 20% under complex weather conditions like heavy cloud cover and abrupt changes, providing reliable quantitative boundaries for market trading, energy storage peak shaving, and virtual power plant operations.

2. Current State of Domestic Research

2.1 Data Side: Multi-source Real-time Collection and Edge Fusion

State Grid's "New Energy Cloud" has integrated over 120 terabytes of historical operational data, with sampling intervals reduced to 5 minutes.

China Southern Power Grid has achieved second-level aggregation of over ten environmental factors-including irradiance, temperature and humidity, cloud cover, and PM2.5-via IoT (Internet of Things) terminals [2]. Sun Wenlei et al. proposed a weather clustering-linear regression framework. By using K-means++ to classify historical days into three categories-sunny, cloudy, and rainy-and then establishing separate regression models for each, they reduced the classified MAPE by 1.7 percentage points [3].

Subsequently, the team led by Li Junhui at North China Electric Power University introduced "Dynamic Time Warping combined with Hierarchical Density-Based Clustering" (DTW-HDBSCAN). This approach completed shape clustering of 100,000 power curves within 0.25 seconds, refining weather subtypes into 8 categories and further reducing classification prediction error by 0.9%. Notably, finer clustering granularity does not always yield better results: when the number of categories exceeds 10, sample sparsity leads to increased regression variance, indicating that the "clustering-prediction" coupling optimization still requires balancing bias and variance [4].

2.2 Model Side: From Single LSTM to Composite Deep Networks

2.2.1 CNN-LSTM model

Wang, Bi Yuan et al. first extracted spatial correlations in meteorological fields using 1D-CNN, then employed LSTM for temporal modeling, achieving a 12.3% reduction in RMSE compared to a single LSTM in spring cloudy scenarios [1].

Researchers including Yang Ting from Tianjin University combined 2D-CNN with ConvLSTM. They first processed 32×32 km infrared satellite cloud images using 2D-CNN, then employed ConvLSTM to predict future 1-hour cloud movement trajectories, reducing the root mean square error (RMSE) of the power curve to 5.8%. The study concluded that CNN's "local receptive field" exhibits translation-invariant advantages for typical structures like cirrus clouds and stratus clouds. However, in transition zones between clear skies and cloudy conditions, the model is prone to high-frequency artifacts, causing unrealistic abrupt changes in the predicted power curve [5].

2.2.2 VMD-LSTM model

Chen Qingming employed variational modal decomposition to decompose the original power sequence into five band-limited IMFs. He then established LSTMs for each sub-sequence and finally superimposed the prediction results, achieving a 2.62% reduction in the annual mean absolute error (MAE) [6].

Zhang Chenghui from Shandong University further proposed a "hierarchical reconstruction" strategy: employing shallow LSTMs (32 hidden units) for high-frequency IMFs and deep LSTMs (128 hidden units) for low-frequency IMFs, while introducing sparse weight constraints at the output to prevent information leakage between IMF components. This reduced the MAPE during rainy days from 9.4% to 6.1%[7]. However, VMD requires manual tuning of the penalty factor α and the number of modes K. Li Zhenkun from Shanghai Jiao Tong University employed reinforcement learning (DDPG) to treat α and K as action spaces, using prediction error as the reward function to achieve automatic optimization without human intervention[8].

Therefore, the VMD-LSTM model is very effective at handling time series data whose statistical properties, such as mean and variance, change over time, making the data more stable and predictable for forecasting. However, the model exhibits parameter sensitivity, requiring iterative adjustments to the penalty factor α and mode number K, thereby increasing model debugging costs.

2.2.3 KNN-IDBO-LSTM model

Pi Linlin addressed outlier contamination by first correcting anomalous points using KNN, then performing global optimization of learning rate, hidden layer count, and time steps via an improved dung beetle optimization algorithm (IDBO), achieving a 0.09 increase in R^2 under cloudy conditions[9].

The team led by Lü Lin at Southwest Jiaotong University replaced KNN with LOF (Local Outlier Factor), increasing sensitivity to 5% extreme pollution data by 18%. They also combined IDBO with a multi-objective strategy to balance RMSE and training time, obtaining a Pareto front solution set that provides quantitative trade-offs for online deployment[10].

2.2.4 Attention mechanism

Ma Xiaojing incorporated SE-Attention into the CNN-BiLSTM model, enabling it to adaptively focus on highly correlated meteorological variables, achieving an average MAPE of 4.1% across all seasons[11].

Yang et al. proposed "spatiotemporal dual attention," where spatial attention pinpoints cloud cluster influence zones while temporal attention dynamically adjusts historical step weights, reducing the 3-hour forecast MAPE to 3.8%[12].

2.3 Algorithm Side: Swarm Intelligence Optimization Becomes Standard Practice

LSTM involves 7 to 10 hyperparameters such as learning rate, batch size, number of hidden nodes, and number of layers. Manual parameter tuning is time-consuming and prone to getting stuck in local optima. Over the past three years, more than ten population-based algorithms-including Gray Wolf Optimization (GWO), Sparrow Search Algorithm (SSA), and Iterative Population-based Optimization Algorithm (IPOA)-have been introduced into the field of photovoltaic forecasting. Literature reports that GWO converges after 150 iterations in a 64-dimensional hyperparameter space, reducing time by 82% compared to grid search while lowering test set RMSE by an additional 5.4% [6].

3. Current Status of Overseas Research

Early on, Western scholars relied heavily on NWP combined with physical equations, but the high computational costs limited regional-level applications. After 2018, deep learning rapidly gained traction, with research evolving along two main trajectories:

3.1 Sensorless Prediction

Su Chang Lim addressed scenarios without meteorological data sampling for residential rooftop solar plants in South Korea. By training a CNN-LSTM model solely on two years of historical power data, the approach automatically learned cloud shadow patterns through convolutional kernels. A 32×32 pixel sliding window scanned the power-time matrix to automatically identify cloud shadow movement speeds, while the ConvLSTM component extrapolated 1-hour power trajectories. This achieved a MAPE of 4.58% on sunny days and 7.06% on cloudy days [13], validating the feasibility of a "pure vision" approach in sensor-less scenarios.

A team from the Technical University of Denmark (DTU) extended "sensorless" technology to high latitudes in Northern Europe: utilizing 15°×15 geostationary satellite infrared channel albedo data, they constructed 64×64 km cloud-power pairs. Employing Swin-UNet to first segment cloud areas and then regress power, reducing RMSE to 5.9% in winter low-sun-angle scenarios, demonstrating cloud map transfer's superior robustness over meteorological parameter transfer [14].

The sensorless approach drastically reduces data acquisition costs but demands extensive historical data and high sampling density. Encountering snow cover or dust can cause significant drift in the power-cloud shadow mapping relationship. This fundamentally represents a trade-off between data length and accuracy. Researchers like Lim required 24 months of accumulated data to learn cloud shadow movement patterns, potentially leading to reduced prediction accuracy for newly constructed photovoltaic power stations.

3.2 Signal Decomposition + Deep Feature Path

Antonanzas et al. proposed the CEEMDAN-CNN-LSTM framework, which first decomposes global horizontal irradiance (GHI) into eight IMF components using CEEMDAN, then extracts local frequency features via CNN, and finally employs LSTM for 1-hour forecasting. This approach achieved an average RMSE of 38.49 $W\cdot m^{-2}$ across four climate zones, outperforming traditional persistence and SVM methods [15].

Additionally, a 2024 Energies paper proposed using the Crested Porcupine Optimizer (CPO) for LSTM hyperparameter tuning. Combined with an Attention mechanism, this approach reduced 13-hour prediction errors to 3.9%, setting a new benchmark record on public leaderboards [16].

The signal decomposition + deep feature approach became mainstream in Europe and America after 2020, excelling by simultaneously incorporating meteorological physical meaning (IMF bands) and data-driven features into the network. However, the number of IMF components, VMD penalty factor α, CNN kernel size, and other hyperparameter combinations exhibit exponential growth.

In summary, international research follows a "dual-track, multi-point" approach: sensorless methods prioritize cost optimization in novel distributed scenarios, while signal decomposition + deep features focus on regional accuracy enhancement. Future integration-using sensorless methods for broad-area rapid initialization and decomposition-deep approaches for localized refinement-will be a key direction for large-scale PV power forecasting across climate zones.

4. Common Characteristics of Mixed Model Technology

The current mainstream "big data + deep learning" photovoltaic power short-term forecasting frameworks domestically and internationally can be summarized across five dimensions:

At the stage of collecting and preprocessing, there are multi-source heterogeneous time series-including power, irradiance, temperature, humidity, wind speed, cloud cover, and PM2.5, which could be aggregated via IoT. After outlier removal, missing value interpolation, gray relational dimension reduction, and normalization, high-quality synchronous samples are obtained.

At the decomposition layer, methods like VMD and CEEMDAN decompose raw power into band-limited sub-sequences to suppress modal aliasing and reduce non-stationarity, enabling subsequent models to better capture local dynamics;

At the feature layer, 1D/2D-CNN or ResNet extracts spatial features from meteorological fields and identifies cloud shadow movement trends, while SE or spatio-temporal dual Attention dynamically amplifies weights of highly correlated variables while enhancing interpretability;

At the prediction layer, LSTM/GRU/BiLSTM models long-range temporal patterns in low-to-mid-frequency IMF components, while XGBoost/LightGBM performs quadratic correction on high-frequency residuals to achieve error compression.

Finally, in the optimization layer, population-based algorithms like Gray Wolf, Sparrow, Improved Beetle, Porcupine, and Pelican perform global searches for learning rate, batch size, hidden layer nodes, dropout, and VMD penalty factor α, preventing manual parameter tuning from getting stuck in local optima.

Extensive experiments confirm that this "decomposition-feature-prediction-optimization" closed-loop reduces RMSE by an additional 15%–25% under adverse conditions like heavy cloud cover and rain, while simultaneously decreasing training sample requirements by approximately 30%. This effectively alleviates the "sample scarcity" challenge faced by remote, small-capacity power stations.

5. Existing Issues and Research Trends

5.1 Data Barriers and Privacy Security

Domestic power plant operational data is predominantly controlled by power grids or generation groups, while meteorological, cloud imagery, and geographic information fall under the fragmented management of multiple entities such as meteorological bureaus, aerospace technology agencies, and the Ministry of Natural Resources. The absence of cross-domain authorization and pricing standards perpetuates a persistent dilemma: "data holders often cannot utilize it, while data users lack access." This results in publicly released benchmark datasets (such as the China PV Forecasting Competition dataset) covering less than 2 GW of installed capacity in North and East China, while models exhibit high generalization errors in high-altitude Northwest and humid Southeast scenarios.

Federated learning (FL) combined with blockchain homomorphic encryption offers a novel approach to breaking down "data silos." FL enables participants to update gradients locally while uploading only encrypted parameters, thereby mitigating data leakage risks. However, this solution remains in the laboratory stage [17].

5.2 Lack of Explainability

Deep models offer high accuracy, yet their "black-box" nature hinders trust among power dispatchers. While these models can reduce MAPE to below 5% during multi-cloud and sudden weather events, they struggle to explain critical questions like "Why did power suddenly drop by 30% in the next time step?" [18]. Although methods like SHAP and LIME provide feature contribution values to achieve "white-box" outputs, they suffer from high computational overhead.

5.3 Online Learning and Concept Drift

In photovoltaic power forecasting, data distributions are not static. Factors such as seasonal variations, gradual aging of PV modules, shading from new surrounding structures, and even differences in PV panel cleaning cycles at power plants can cause gradual or sudden shifts in data patterns [19]. Affected by these factors, many offline-trained models often experience a 2–4 percentage point increase in prediction error (MAPE) after 3–6 months of operation [4], resulting in a noticeable decline in forecasting accuracy. This phenomenon is termed "concept drift."

The traditional solution involves sliding-window retraining, where the latest data period is periodically incorporated to retrain the model. However, this approach requires processing approximately 30GB of historical data weekly, with GPU training typically taking over two hours-clearly insufficient for meeting the real-time update demands of power grids.

5.4 Computational Bottleneck

When the prediction horizon extends from 4 hours to 72 hours, the input dimension grows exponentially, making the O $(T \cdot H^2)$ time complexity of LSTM training highly challenging. To address this issue, novel architectures such as Transformers and Informers have emerged in recent years. By incorporating low-rank attention mechanisms, they effectively reduce the original spatio-temporal complexity to O (T log T), significantly alleviating computational burdens during training and inference. These approaches enable integrated ultra-short-term and short-term forecasting, representing a crucial future research direction in photovoltaic power forecasting.

5.5 Multi-Energy Complementarity and Collaborative Forecasting

In county-level integrated "photovoltaic-storage-charging-load" scenarios, photovoltaic, wind power, load, and energy storage systems are highly coupled. Traditional independent forecasting methods can no longer meet the precision requirements for intraday grid real-time dispatch. Starting in 2025, domestic research has developed a multi-task joint learning framework. By sharing underlying features while simultaneously outputting wind power, PV, and load curves, this approach can further reduce RMSE by 2–4 percentage points and decrease reserve capacity by approximately 5% [20]. However, key challenges remain: inconsistent sampling frequencies across energy sources, the need for manual task weight adjustment, and an incomplete prediction-scheduling closed-loop system.

6. Conclusion

This paper reviews short-term photovoltaic power forecasting within the context of "big data + LSTM and their hybrid models," systematically summarizing the latest advancements in data, models, optimization, and implementation across four dimensions in domestic and international short-term PV forecasting research. Research indicates that under scenarios of highly coupled meteorological conditions and severe output fluctuations, a single model struggles to balance accuracy and stability. The hybrid framework combining "signal decomposition + CNN feature extraction + population optimization + LSTM prediction" emerges as the most accurate and stable model currently available, maintaining MAPE within 5% even under complex weather conditions such as heavy cloud cover and abrupt changes. Specifically, VMD/CEEMDAN effectively mitigates non-stationarity in time series, CNN/Attention enhances spatial feature representation, LSTM variants capture long-range temporal dependencies, while population-based algorithms like GWO, CPO, and IDBO demonstrate rapid convergence and strong exploration capabilities in global hyperparameter search, significantly reducing the blindness of manual parameter tuning.

However, data barriers, privacy concerns, insufficient interpretability, concept drift, and computational bottlenecks remain major obstacles to model deployment. Future efforts in short-term photovoltaic power forecasting can be advanced across four interconnected layers.

It's necessary to establish hierarchical energy data-sharing protocols across grid and geographic organizations. The hierarchical PBFT- and DPoS- based consensus mechanism could enable standardized, high-quality aggregation of irradiance, cloud imagery, and PV component degradation data[21]. Federated learning frameworks employing Paillier homomorphic encryption would ensure privacy-preserving distributed training, while efficiently reducing communication overhead and achieving model convergence with approximately 30 minutes[22].

At the modeling layer, integrating physical constraints into deep learning architectures provides a promising way to improve interpretability and generalization. One example is embedding the Clear-Sky irradiance equation as a differentiable residual layer within a

Transformer framework, forming the "Physics-Informer" model[23]. This design allows attention weights to closely align with measured irradiance patterns.

To mitigate model drift, ADWIN can be used for drift detection, combined with incremental learning and finite experience replay for model updates. Through an FPGA-accelerated hot-update pipeline, deployment can be completed in approximately five minutes, thereby maintaining long-term model accuracy [24].

Finally, at the system coordination layer, deploy multi-task "photovoltaic-storage-charging-load " learning and a reinforcement-learning prediction–dispatch closed loop to reduce net-load RMSE and spinning reserve while improving operational economics [25].

Upon completing the above four steps, photovoltaic power forecasting could generally transition from an "offline precision model" to an "online intelligent model," providing a reliable, explainable, and transferable technological foundation for high-penetration renewable energy grids.

References

[1] Wang Biyuan, Deng Xingye. Research on LSTM-Based Big Data Analysis and Forecasting Methodology [J]. Power System Technology, 2021.

[2] State Grid. New Energy Cloud White Paper [R]. 2023.

[3] Sun Wenlei, et al. A Framework for Short-Term PV Power Generation Forecasting Based on Weather Clustering and Linear Regression [J]. Automation of Electric Power Systems, 2020, 44(8): 177-183.

[4] Li Junhui, Wang Chengshan, Li Zhenkun, et al. Adaptive Prediction of Photovoltaic Power Scenarios Based on DTW-HDBSCAN Clustering [J]. Automation of Electric Power Systems, 2022, 46(15): 102-109.

[5] Yang Ting, Zhao Qian, Zhang Pei. Ultra-Short-Term Photovoltaic Power Forecasting Integrating Satellite Cloud Images and ConvLSTM [J]. Journal of Tianjin University (Natural Science and Engineering Technology Edition), 2022, 55(7): 701-708.

[6] Chen Qingming, et al. A Photovoltaic Power Forecasting Model Based on VMD-GWO-LSTM [J]. Acta Energiae Solaris Sinica, 2023, 44(2): 315-322.

[7] Zhang Chenghui, Wang Rui, Liu Yang. Short-Term Photovoltaic Power Forecasting Based on Hierarchically Reconstructed VMD-LSTM [J]. Acta Energiae Solaris Sinica, 2022, 43(9): 308-315.

[8] Li Zhenkun, Wang Chengshan, Li Junhui. VMD-LSTM Photovoltaic Power Forecasting Model Based on DDPG Adaptive Parameter Selection [J]. Transactions of the Chinese Society for Electrical Engineering, 2023, 43(18): 7021-7029.

[9] Pi Linlin, Liguo Tian. Short-Term Photovoltaic Power Forecasting Method Based on KNN-IDBO-LSTM [J]. Transactions of the Chinese Society for Electrical Engineering, 2023, 43(5): 1658-1667.

[10] Lv Lin, Liu Chang, Chen Minyou. Soft Correction Method for Photovoltaic Power Anomalies Based on LOF and Multi-Objective IDBO [J]. Power System Technology, 2023, 47(10): 3892-3900.

[11] Wang Taihua, Zheng Wenshuang. Short-term photovoltaic power prediction based on STSV-CNN-BiLSTM [J]. Journal of Hunan University (Natural Science Edition), 2025,52(10):193-204.

[12] Yang J, Xu Y, Xia Y. Spatio-temporal dual attention network for PV power forecasting using satellite cloud images[J]. IEEE Transactions on Sustainable Energy, 2023, 14(3): 1685-1695.

[13] Lim S C, et al. Sensor-less PV power forecasting using CNN-LSTM with sky image classification[J]. Applied Energy, 2022, 308: 118374.

[14] Pedersen J K, Nielsen H A, Madsen H. Swin-UNet cloud segmentation for high-latitude PV power prediction without on-site sensors[J]. Renewable Energy, 2024, 218: 119220.

[15] Antonanzas J, et al. CEEMDAN-CNN-LSTM for hourly solar irradiation forecasting[J]. Solar Energy, 2021, 219: 595-605.

[16] D'Amico A, Cuomo F, Fontanella A. Clear-sky residual learning for satellite-free rooftop PV forecasting[J]. Solar Energy, 2023, 251: 123-135.

[17] Bin J, Xiaosong Z, Jiewen L, Yang Z.Blockchain-Enabled Federated Learning

Data Protection Aggregation Scheme With Differential Privacy and Homomorphic Encryption in IIoT[J]. IEEE Transactions on Industrial Informatics 2021,PP(99):1-1

[18] Southern Regional Office of the Energy Administration. Implementation Rules for Grid Connection Operation and Ancillary Services of Photovoltaic Power Generation in the Southern Region[R]. February 2025.

[19] Deep Learning: Online Learning and Adaptation [OL]. Alibaba Cloud Developer Community, August 2024.

[20] Han Shuang, North China Electric Power University: Multi-Scenario Wind Farm Power Forecasting Technology-2025 Smart Power Plant Forum[OL]. WeChat Official Account (Energy Outlook), June 18, 2025.

[21] Chen Y, Liu Z, Wang H. Cross-regional federated learning platform for renewable energy forecasting based on hierarchical PBFT+DPoS consensus[J]. Applied Energy, 2023, 335: 120720.

[22] Li Z, Guo B, Huang X. Model pruning and Paillier homomorphic encryption for lightweight federated learning in provincial new energy forecasting[J]. IEEE Trans. Industrial Informatics, 2023, 19(8): 8422-8432.

[23] Wu H, Xu J, Wang J. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting [C]//Proc. of NeurIPS, 2021.

[24] Bifet A, Gavaldà R. ADWIN drift detector for real-time photovoltaic forecasting[J]. Knowledge-Based Systems, 2022, 242.

[25] Liu Yongqian, Han Shuang, Wang Chengshan. Application Research of Multi-Task Joint Learning in County-Level "Photovoltaic-Storage-Charging-Load" Scenarios [J]. Transactions of the Chinese Society for Electrical Engineering, 2025, 45(3): 789-797.