Minimax Theory and its Applications

Volume 11 (2026), No. 5s, 58-73

A Distributionally Robust Minimax Optimization Framework with Transformer-Based Temporal Learning for Cross-Plant Photovoltaic Power Forecasting

Ganesh Karthikeyan Varadarajan[1*], Srikanth Lakumarapu[2], Sree Ranganayaki Vanamamaley[3], Malathi Subramanian[4], Koteeswaran Seerangan[5] ,Nalini Manogaran[6]

[1] School of Computing, Sastra Deemed University,Thanjavur-613401, Tamil Nadu, India.
[2] Department of CSE, CVR College of Engineering, Hyderabad - 501510, Telangana, India.
[3] Department of Computer Science and Engineering, University College of Engineering & Technology forWomen, Kakatiya University, Warangal-506009, Telangana, India.
[4] Department of AI&DS, Panimalar Engineering College, Poonamallee -600123, Tamil Nadu, India.
[5] Department of Computer Science and Engineering, R. M. K. Engineering College (Autonomous), Kavaraipettai - 601206, Tamil Nadu, India.
[6]Department of CSE (Cyber Security), S.A. Engineering College (Autonomous), Chennai – 600077, Tamil Nadu, India.
Email: ganeshkarthikeyan@cse.sastra.edu, dr.srikanthcse@cvr.ac.in, ranganayakisree36@gmail.com , Malathi.raghuram@gmail.com,skn.cse@rmkec.ac.in, drnalini@saec.ac.in

Abstract
Accurate photovoltaic (PV) power forecasting is essential for reliable renewable-energy integration, yet models trained by empirical risk minimization optimize average error and offer weak protection against distributional differences across geographically dispersed plants. This study addresses cross-plant PV forecasting by controlling the worst-case prediction risk over heterogeneous plants rather than the mean error. Existing PV forecasting studies, including transfer learning, graph-based, domain-adversarial and ensemble approaches, mainly improve average accuracy and provide no formal minimax formulation, worst-case guarantees or theoretical robustness analysis. This study therefore proposes Minimax-DRO-PVNet, a distributionally robust minimax optimization framework that couples Transformer-based temporal representation learning with a Kullback-Leibler (KL) divergence-constrained uncertainty set over plant-level risk weights. The BR-PVGen dataset is used, containing synchronized inverter and meteorological records at 15-minute intervals from 51 Brazilian PV plants, and all experiments are implemented in Python. The inner adversarial problem is solved through its KL dual representation, and the model is trained by a primal-dual algorithm. The theoretical analysis establishes the existence of minimax solutions, the dual formulation, worst-case risk bounds and monotonicity with respect to the uncertainty radius, while explicitly stating the limits imposed by non-convex neural parameterization. Performance is evaluated using MAE, RMSE, R^2 and nRMSE over 15-, 30-, 60- and 120-minute horizons, with comparison against conventional machine-learning and deep-learning baselines, ablation, and radius sensitivity analysis. Minimax-DRO-PVNet achieved R^2 between 0.982 (15 min) and 0.923 (120 min), and its best configuration at a KL radius of 0.05 gave a MAE of 31.84 kW and an RMSE of 48.92 kW. The study links minimax optimization theory with practical machine learning, providing a mathematically grounded and robust framework for reliable cross-plant PV forecasting.

Keywords*:* Minimax Optimization; Distributionally Robust Optimization; KL Divergence; Primal-Dual Algorithm; Transformer; Photovoltaic Power Forecasting.

1. Introduction

The increasing integration of renewable energy into modern power systems has intensified the need for accurate and reliable photovoltaic (PV) power forecasting under varying meteorological and operational conditions [1], [2], [3].The prediction tasks are subject to a considerable uncertainty due to solar power generation being influenced by changes in solar irradiance, ambient and panel temperatures, wind speed, precipitation and plant-specific characteristics [4], [5], [6].Traditional machine learning methods, such as ensemble learning and statistical forecasting models, typically aim to reduce the average prediction error without considering performance differences among diverse PV systems [7], [8], [9].While deep learning architectures, such as Transformer models, have been found to model complex temporal dependencies encountered in meteorological and electrical measurements, they tend to focus on the empirical risk minimization objective which offers weak

protection against distributional differences across plants located in geographically distributed areas.Moreover, current distributionally robust learning methods lack a solid mathematical foundation that can incorporate plant-wise uncertainty modelling, minimax optimisation and temporal representation learning for large-scale PV forecasting [10], [11], [12].In particular, optimizing average prediction risk does not necessarily control the worst-case prediction error associated with individual PV plants operating under different climatic conditions [13], [14], [15].These restrictions motivate a mathematically-based distributionally robust minimax learning framework that is able to handle the heterogeneity of plant-level distributions and meteorological uncertainty.Based on this, this study presents a Transformer-integrated minimax optimization framework with the BR-PVGen dataset, which includes synchronized measurements of the inverter and meteorological data from 51 PV plants in Brazil.The proposed approach calculates prediction risks within a distributional uncertainty set for the plants and weights the worst-case contribution of the prediction risk in the primal dual optimization process.The research objectives are to develop a plant-wise minimax formulation, incorporate Transformer-based temporal feature extraction, investigate the existence and duality properties of optimal solutions, establish appropriate convergence and worst-case risk guarantees, and experimentally evaluate forecasting accuracy, cross-plant generalization, and robustness under meteorological distribution shifts.The resulting framework is designed to link the minimax optimization theory and the actual PV power forecasting under uncertainty.

Recent studies improved PV forecasting using Transformers, transfer learning, graph networks and ensembles, but optimize average accuracy without worst-case guarantees.Piantadosi et al. [16] aimed to improve day-ahead PV production forecasting without relying on local irradiance measurements. The method takes an irradiance model and provider meteorological data as input, and uses a Transformer neural network to be trained. The model was tested with OpenMeteo data, where the MAE was 1.22 kW and the MAPE is 2.21%. The evidence is still limited to one-day-ahead time horizons, and a limited spatial and temporal data diversity. Y. Zhang et al. [17] re-examined architecture choices for PV forecasting combining historical generation with weather forecasts. The proposed Cross-Unet is designed to perform multi-scale temporal encoding, channel attention based on channel correlation, and decoding by cross-attention at different scales. Experiments on four northern Chinese plants and one Australian plant outperformed ten deep learning baselines across horizons from four hours to seven days. Performance based on forward looking weather inputs and no explicit distribution-shift handling.

Bak et al. [18] studied the cross-continental transfer learning for PV forecasting at data-scarce sites. A transformer-based model was pre-trained on a source dataset of over 3.5 million points from Australia, followed by zero-shot, linear probing, and fine-tuning on targets from Korea, Germany, the USA and the UK. MAPE fell by up to 38.8%. The drawbacks are the lack of humidity features and the skewed volume of data in the various regions. Yan et al. [19] tackled PV forecasting in the presence of severe data missingness in both power and auxiliary variables. The MDCTL-MCI framework consists of multivariate singular spectrum analysis, a multiscale covariate interaction module, and multi-source transfer learning and output-layer fine tuning. Experiments on four Chinese PV installations improved average accuracy by 10.5% with complete data and 15.3% under missing-data scenarios. Limited variables resulted in the use of few covariate types in the study.

Zhao et al. [20] focused on spatiotemporal distribution shifts to degrade distributed PV forecasting for the estimation of the baseline load of EVs. The graph attention network models the dependency between stations, the HDBSCAN model clusters features into the distribution domains and the temporal convolutional network with gradient reversal learns the invariant representation. Tests with real-world distributed PV datasets enhanced accuracy and generalization in transitional weather. It lacks worst-case risk formulation and formal robustness guarantees. J. Zhang et al. [21] aimed to enhance long-term PV forecasting with the aid of environmental data. The A-GNN is a graph neural network that models spatiotemporal coupling through an alternating architecture of temporal convolution layers and variable interaction layers, which are based on a directed graph. Evaluated on three regional sites, it lowered MSE and MAE by 2.19% and 1.57% at 384 steps, and 2.81% and 2.47% at 768 steps. The dynamic adjacency matrix is not very stable.

Song et al. [22] aimed to forecast PV power robustly for small distributed solar systems. The flexible hybrid ensemble framework uses a meta-model to select the suitable base model according to prediction error patterns. Four South Korean plants were assessed and MAPE increased by 30% compared with a support vector regression model. However, the accuracy was not investigated for smaller plants and transfer learning between plants was not explored. The reviewed studies have demonstrated that PV forecasting has evolved with transfer learning, multi-domain collaboration, domain-adversarial learning, graph networks, and ensemble selection, primarily to achieve better average accuracy. However, these methods optimize mean error, rely on specific datasets or weather inputs, and offer no formal worst-case guarantees under plant-wise distribution shifts. None of these studies integrates distributional uncertainty, minimax risk optimization and theoretical robustness analysis with Transformer-based temporal learning. This gap inspires the proposed distributionally robust minimax solution for cross-plant forecasting.

Existing PV power forecasting approaches predominantly employ conventional machine learning and deep learning models that minimize average prediction errors, often overlooking worst-case performance variations across heterogeneous PV installations [23]. Although Transformer-based architectures capture complex

temporal dependencies in meteorological and electrical measurements, their optimization generally relies on empirical risk minimization, which provides limited protection against distributional variations among geographically dispersed PV plants [24]. Furthermore, existing forecasting approaches provide insufficient integration of plant-wise distributional uncertainty, minimax risk optimization, and mathematically justified robustness guarantees within a unified learning framework. These limitations can result in inconsistent forecasting performance under changing solar irradiance, temperature, wind conditions, and plant-specific operating characteristics [25]. To address these challenges, this study proposes a distributionally robust minimax optimization framework integrated with Transformer-based temporal representation learning for cross-plant photovoltaic power forecasting [27]. The proposed framework formulates plant-wise prediction risk as a worst-case optimization problem over a mathematically defined uncertainty set and employs primal-dual optimization to identify robust model parameters. Additionally, theoretical analyses investigate solution existence, minimax duality, convergence properties, and worst-case risk bounds, while cross-plant experiments evaluate forecasting accuracy and robustness under meteorological distribution shifts. The main contributions of this study are summarized as follows:

- Develops a plant-specific minimax risk formulation that explicitly models heterogeneous photovoltaic distributions and prioritizes worst-case forecasting performance across multiple PV installations.
- Establishes a mathematically defined uncertainty set over plant-level risk weights, enabling controlled distributional robustness through divergence-constrained adversarial optimization.
- Integrates Transformer-based temporal representation learning with plant-wise risk estimation to capture complex dependencies in meteorological and photovoltaic operational measurements.
- Formulates a primal-dual optimization strategy and investigates solution existence, minimax duality, convergence properties, and worst-case forecasting risk bounds under appropriate assumptions.
- Provides a computational evaluation using the BR-PVGen dataset to assess forecasting accuracy, cross-plant generalization, uncertainty sensitivity, and robustness under meteorological distribution shifts.

2. Mathematical Preliminaries and Minimax Optimization Formulation

This section presents the mathematical basis of Minimax-DRO-PVNet by modeling the PV forecasting problem as a minimax distributionally robust optimization problem. Unlike traditional learning methods that minimize the average prediction error, the proposed formulation explicitly takes into consideration the variation in the performance of forecasts across different PV plants.

2.1 Dataset Description and Problem Formulation

The BR-PVGen dataset, proposed by Jesus et al. (2026), is employed to study PV prediction at geographically distributed PV plants[26]. The dataset consists of 51 Brazilian PV plants recorded from 26 March 2024 to 9 June 2025 at 15-minutes interval. The dataset contains inverter level electrical data such as active power, reactive power, dc power, and inverter temperature, together with meteorological data that includes POA, GHI, GRI radiation, ambient temperature, panel temperature, wind speed, and precipitation. Plant metadata include capacity, panel efficiency, bifaciality, and physical characteristics. These diverse measurements aid in analyzing risks associated with forecasting for a particular plant.

2.1.1 Data Synchronization and Preprocessing

Inverter and meteorological data are synchronized based on time stamps for creating 15-minute observations. Missing and abnormal data points are analyzed prior to model training. Outliers are detected using statistical principles and evaluated to determine whether they represent actual operating differences or measurement errors. Continuous variables are scaled using statistics from the training dataset to avoid scale differences from impacting optimization. Records are organized by plant identifiers to retain the plant-specific distribution and allow for chronological and inter-plant analysis.

The input vector combines meteorological variables, electrical measurements, plant characteristics, and historical active-power observations. The future active power is chosen as the prediction variable in order to make multi-step forecasts using historical data without including future measurements.Mathematical Formulation of the Forecasting Problem: For plant k, the data set is formulated as Eqn. (1).

$$\mathcal{D}_k = \{(x_{ki}, y_{ki})\}_{i=1}^{n_k} \quad (1)$$

Where x_{ki} denotes historical input features and y_{ki} denotes future active power. The objective is to learn $f_\theta(x_{ki})$ that minimizes forecasting errors across heterogeneous plants.

Assumptions and Notation: Let $\theta \in \Theta$ represent model parameters, P_k the data distribution of plant k, and ℓ the forecasting loss. The plant-wise empirical risk is $R_k(\theta) = \frac{1}{n_k}\sum_{i=1}^{n_k} \ell\,(f_\theta(x_{ki}), y_{ki})$.

To account for distributional uncertainty, plant weights are part of $\mathcal{Q} = \{q \in \Delta_K : D(q\|p) \le \rho\}$,where p denotes nominal weights, D a divergence measure, and ρ the uncertainty radius. This formulation supports worst-case risk minimization across plants.

2.2 Preliminary Definitions and Notation

Let Kdenote the total number of PV plants, and let $\mathcal{X}$and $\mathcal{Y}$represent the input and output spaces, respectively. The observations from plant $k$follow an unknown joint probability distribution $P_k$over $\mathcal{X} \times \mathcal{Y}$. The forecasting model is represented by $f_\theta : \mathcal{X} \to \mathcal{Y}$, where $\theta \in \Theta$denotes the trainable model parameters and $\Theta$is the parameter space. Expected prediction risk for each plant is defined as Eqn. (2) and Eqn. (3).

$$R_k(\theta) = \mathbb{E}_{(x,y)\sim P_k}[\ell(f_\theta,(x)y)] \quad (2)$$

$$R_k(\theta) = \mathbb{E}_{(x,y)\sim P_k}[\ell(f_\theta(x),y)] \quad (3)$$

Where $\ell$is a non-negative loss function, such as mean squared error (MSE). In finite cases, the expectation is approximated by the empirical risk function as given in Eqn. (4) and Eqn. (5).

$$R_k(\theta) = \frac{1}{n_k}\sum_{i=1}^{n_k} \ell\,(f_\theta,(x_{ki})y_{ki}) \quad (4)$$

$$R_k(\theta) = \frac{1}{n_k}\sum_{i=1}^{n_k} \ell\,(f_\theta(x_{ki}),y_{ki}) \quad (5)$$

Here, $n_k$denotes the number of observations from plant k. A probability vector $q = (q_1, \dots, q_K)$assigns weights to individual plants, satisfying $q_k \geq 0$and $\sum_{k=1}^{K} q_k = 1$. These weights determine the contribution of each plant to the aggregated forecasting risk.

2.3 Empirical Risk Minimization

In conventional empirical risk minimization (ERM), the estimation of model parameters occurs by minimizing the average prediction risk of all K PV plants, as shown mathematically in Eqn. (6).

$$\min_{\theta\in\Theta} \frac{1}{K}\sum_{k=1}^{K} R_k\,(\theta) \quad (6)$$

The formula places equal weight on each plant regardless of forecasting challenge or distributional features. Therefore, plants which have stable meteorological conditions could be dominant in the entire learning goal, whereas those which have fluctuations in irradiance or operation may have higher prediction errors. ERM therefore provides an average-performance objective but does not explicitly control the worst-performing plant.

2.4 Plant-Wise Minimax Risk Formulation

To overcome this drawback, the developed framework employs adversarial plant weighting and views photovoltaic power forecasting as a minimax risk optimization problem for each PV plant, where the model seeks to minimize the maximum weighted risk of prediction over all K PV plants, as defined by Eqn. (7).

$$\boxed{\min_{\theta\in\Theta}\max_{q\in Q} \sum_{k=1}^{K} q_k\, R_k(\theta)} \quad (7)$$

The inner maximization finds the distribution of plant weights which gives maximum prediction risks under the parameters of the existing model. The outer minimization updates $\theta$to reduce this worst-case risk. Here, $q_k$represents the adversarial weight assigned to plant k, while Qdefines the allowable weighting distributions. The formulation ensures that the model improves its performance on hard cases rather than just improving its average accuracy.

2.5 Distributionally Robust Uncertainty Set Construction

The uncertainty set is defined using a divergence constraint: $Q = \{q \in \Delta_K : D(q \parallel p) \leq \rho\}$,where $\Delta_K = \left\{q \in \mathbb{R}_+^K : \sum_{k=1}^{K} q_k = 1\right\}$. Here, $p = (p_1, \dots, p_K)$denotes the nominal plant-weight distribution, generally selected using equal weights or plant sample proportions. The divergence $D(q\|p)$measures the deviation between adversarial and nominal weights. One suitable measure is the Kullback-Leibler divergence as given in Eqn. (8)

$$D_{\mathrm{KL}}(q\|p) = \sum_{k=1}^{K} q_k \log\left(\frac{q_k}{p_k}\right) \quad (8)$$

The parameter $\rho \geq 0$controls the uncertainty radius. The smaller radius limits adversarial deviation, while the larger radius allows more redistribution to the high-risk plants. This constraint prevents unrestricted weighting and provides a controllable mechanism for distributional robustness.

2.6 Minimax Optimization Properties

The mathematical properties of the minimax formulation rely on the continuity of risk functions for plants, compactness of the uncertainty set, and properties of the model parameter space. The following properties define the requirements for analyzing optimal solutions and adversarial plant weighting.

2.6.1. **Existence of optimal solutions**: If the parameter space Θ is nonempty and compact, the uncertainty set Q is nonempty and compact, and each plant-wise risk function $R_k(\theta)$ is continuous on Θ, the minimax objective admits an optimal solution. The inner maximization also attains its maximum over Q.

2.6.2.Convexity of adversarial optimization: For fixed model parameters θ, the aggregated risk $\sum_{k=1}^{K} q_k\, R_k(\theta)$ is linear in the adversarial weight vector q. Since the KL-divergence is defined as a convex uncertainty set, the inner optimization problem can be considered a convex optimization problem because it involves finding the maximum of a concave function within a convex set.

2.6.3. Non-convexity of model optimization: The outer minimization depends on the neural-network parameters θ. While the Transformer can model complicated temporal dependencies, the risk associated with its predictions is non-convex with regard to its trainable parameters. Consequently, global optimality cannot be inferred solely from the convexity of the adversarial weight optimization.

2.6.4. Compactness of the uncertainty set: For strictly positive nominal weights p_k, the KL-divergence-constrained set is defined as a closed subset of the probability simplex. Since the probability simplex is compact in finite-dimensional space, the resulting uncertainty set is also compact for finite ρ.

2.6.5. Continuity of the robust objective: If all plant-wise risk functions are continuous, the weighted risk $F(\theta, q) = \sum_{k=1}^{K} q_k\, R_k(\theta)$ is jointly continuous. Maximization over the compact uncertainty set results in a continuous robust risk function, ensuring the achievement of the minimum on a compact parameter set. These properties set the mathematical basis for the proposed distributionally robust formulation. In particular, the inner

adversarial optimization is guaranteed to have a solution, while the global optimality of the full minimax problem for the Transformer model is subject to some convexity constraints that do not hold for neural networks.

2.7. Proposed Minimax-DRO-PVNet Framework

The proposed Minimax-DRO-PVNet combines the temporal representation learning using a Transformer with the distributionally robust minimax optimization for cross-plant PV power forecasting. It consists of six computational steps – data preprocessing and synchronization, temporal representation learning, plant-wise risk estimation, adversarial plant-weight optimization, primal-dual parameter updating, and future active-power prediction. The BR-PVGen dataset, which consists of inverter measurements, meteorological observations and static characteristics of 51 PV plants in Brazil, is first sorted by plant identifier and by time. Electric measurements are summarized at the plant site and synced up with the appropriate meteorological data. Static plant characteristics are incorporated using plant identifiers. Data is synchronized and then assessed for missing values, duplicate detection, flagging of interpolated values, outlier analysis and feature normalization. Historical data of irradiance, temperature, wind, electrical operating variables and active-power data are then grouped into fixed-length temporal sequences. The complete pipeline of the presented Minimax-DRO-PVNet framework, comprising the main processing phases, along with its optimization technique, is shown in Figure 1.

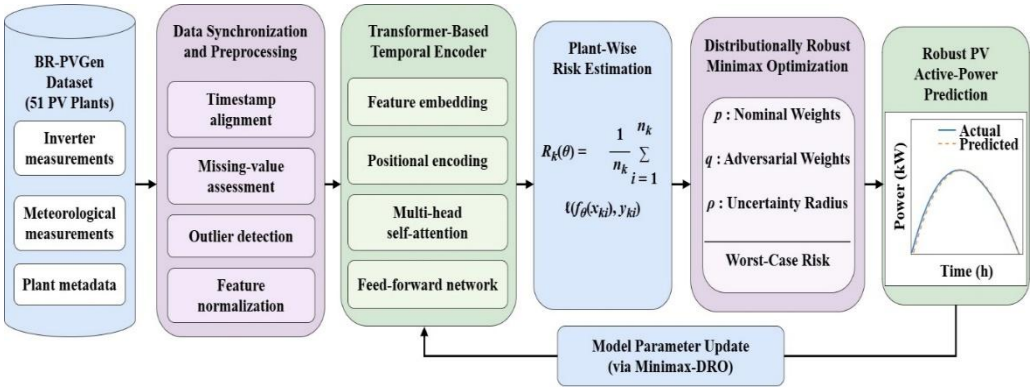


Figure 1. Overall Architecture of Minimax-DRO-PVNet

The Transformer encoder learns temporal relationships by feature embedding, positional embedding, multi-head self-attention and feed-forward layers. The forecasting head produces future plant-level active-power forecasts, and the mean squared error of the forecasts at each level (plant level) is computed to obtain the forecasting risk vector. These risks are incorporated into a minimax objective that maximizes weighted prediction error over a Kullback–Leibler (KL)-divergence-constrained uncertainty set. The primal-dual optimization algorithm cycles between estimating the plant weights and updating the weights of the neural networks, using an importance adjustment for the plants to manage the difficulty of forecasting and limiting the deviation from the nominal plant distribution. Theoretical analysis deals with the existence of solutions to the problems, dual representation, convergence conditions, worst case risk bounds and uncertainty-radius sensitivity. Chronological experiments and cross-plant experiments are used to test forecasting accuracy and robustness in heterogeneous meteorological conditions. Plant-specific forecasting errors are calculated independently to obtain the risk vector given in Eqn. (9).

$$\mathrm{R}(\theta) = [R_1(\theta) \quad R_2(\theta) \quad \cdots \quad R_{51}(\theta)]^T \tag{9}$$

The distributionally robust minimax module is used to estimate the weights of the plants in the uncertainty set defined by KL-divergence constraints, where the uncertainties are these risks. The higher-risk plants can be given heavier weights subject to the stated uncertainty radius. Using gradient-based optimization, the resulting weighted forecasting objective is used to update the parameters of the Transformer. These operations are iterated until the stopping criterion for training is met. The optimized model is subsequently evaluated using independent chronological and cross-plant test data.

2.7.1 Transformer-Based Temporal Representation Learning

Transformer-based temporal representation learning is capable of learning the temporal dependency from the historical meteorological and electrical measurements. It features embedding of features, position encoding, multiple attention heads, residual connections, layer normalization and feed-forward layers. Feature embedding converts heterogeneous inputs into a common dimensional space, while positional encoding preserves temporal order. Feed-forward layers help to improve the representations extracted and Self-attention captures the relationship between the historical observations. Temporal features are generated as a result, to facilitate active-power forecasting in the future. Transformer-based temporal representation learning is capable of learning the temporal dependency from the historical meteorological and electrical measurements. The historical input sequence of plant k is represented as Eqn. (10).

$$X_{k,t} = [x_{k,t-L+1}, \dots, x_{k,t}] \in \mathbb{R}^{L \times d} \quad (10)$$

Where L denotes the historical sequence length, d represents the number of input features, and $x_{k,t}$ is the feature vector at timestamp t. The input variables are data from the irradiance available, ambient temperature, panel temperature, wind data, data from the inverter electrical measurements, historical active power, time data and selected static plant characteristics. The forecasting model predicts future active power at horizon h. In Feature embedding the heterogeneous input variables are projected into a common representation space using a trainable linear embedding as presented in Eqn. (11).

$$H_k^{(0)} = X_{k,t} W_E + b_E \quad (11)$$

Where W_E and b_E denote the embedding weight matrix and bias vector, respectively. The resulting representation provides a common feature dimension for subsequent attention operations. Positional encoding: Temporal information is incorporated into the embedded representation through positional encoding as in Eqn. (12).

$$Z_k^{(0)} = H_k^{(0)} + PE \quad (12)$$

Where PE represents the positional encoding matrix. This operation preserves the sequential order of historical observations and enables the attention mechanism to distinguish different temporal positions. Multi-head self-attention: Temporal dependencies are extracted using scaled dot-product attention as shown in Eqn. (13) and (14).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_h}}\right) V \quad (13)$$

Where

$$Q = ZW_Q, \qquad K = ZW_K, \qquad V = ZW_V \quad (14)$$

Here, W_Q, W_K, and W_V are trainable projection matrices, while d_h denotes the dimension of each attention head. Multiple attention heads capture different temporal relationships among irradiance variations, temperature changes, inverter measurements, and historical power-generation patterns. Residual and feed-forward transformation: The attention representation is refined through residual connections, layer normalization, and feed-forward transformations. For Transformer block l, the operations are defined as Eqn. (15).

$$A_k^{(l)} = \text{MHA}\left(H_k^{(l)}\right), U_k^{(l)} = \text{LayerNorm}\left(H_k^{(l)} + A_k^{(l)}\right), H_k^{(l+1)} = \text{LayerNorm}\left(U_k^{(l)} + \text{FFN}\left(U_k^{(l)}\right)\right) \quad (15)$$

Here, $A_k^{(l)}$ represents the multi-head attention output, $U_k^{(l)}$ denotes the normalized residual representation, and $\text{FFN}(\cdot)$ represents the position-wise feed-forward network. The number of Transformer encoder blocks is denoted by L_T. Temporal representation and forecasting: After the final Transformer block, the temporal representation is obtained through pooling: $h_k = \text{Pool}\left(H_k^{(L_T)}\right)$.The resulting representation is supplied to a forecasting head as given in (16).

$$\hat{y}_{k,t+h} = g_\theta(h_k) \quad (16)$$

Where $g_\theta(\cdot)$ denotes the prediction head and $\hat{y}_{k,t+h}$ represents the predicted future active power for plant k. This representation enables the forecasting model to learn temporal relationships from multiple electrical and meteorological input variables.

2.7.2 Plant-Wise Prediction Risk Estimation

Plant-wise prediction risk estimation is used to quantify prediction errors for every photovoltaic installation separately. The forecasting head generates predicted active power from the extracted temporal representations. The Mean squared error is used to identify the difference in predicted and measured power to each observation. The errors are averaged in each plant to get the empirical risks for each plant. This stage detects differences in the difficulty of forecasting and offers personal risk values for the subsequent minimax optimization. For plant k, the forecasting output for observation i is expressed as in Eqn. (17).

$$\hat{y}_{k,i} = f_\theta\left(X_{k,i}\right) \quad (17)$$

Where $X_{k,i}$ represents the historical input sequence, $f_\theta(\cdot)$ includes the Transformer encoder and forecasting head, and $\hat{y}_{k,i}$ denotes the predicted future active power. The plant-specific mean squared error is calculated as in Eqn. (18).

$$\boxed{R_k(\theta) = \frac{1}{n_k} \sum_{i=1}^{n_k} \left(y_{k,i} - \hat{y}_{k,i}\right)^2} \quad (18)$$

Where $y_{k,i}$ denotes the measured future active power and n_k represents the number of valid forecasting observations for plant k. Capacity-normalised active power is used to calculate MSE between predicted and measured values, enabling comparison across plants with different installed capacities. Plant-wise risks are retained separately to identify higher-error plants and adjust their contributions to the minimax learning objective.

2.7.4 Distributionally Robust Minimax Learning

Distributionally robust minimax learning is a technique which determines the worst-case weighted risk of forecast over a set of diverse PV power plants. It employs nominal plant weights, adversarial weight estimation and an uncertainty set that is defined using the divergence. The inner maximization maximises the contributions of the higher risk plants, the outer minimisation updates the model parameters to decrease the worst case

aggregated risk. The uncertainty radius sets weight deviations, offering a systematic method to enhance robustness to plant level distributional variations. Let $p = (p_1, \dots, p_{51})$ denote the nominal plant-weight distribution. For equal nominal importance, the weights are defined as Eqn. (19).

$$p_k = \frac{1}{51}, \qquad k = 1, \dots, 51 \tag{19}$$

The adversarial plant-weight distribution is represented by $q = (q_1, \dots, q_{51})$, where q_k denotes the weight assigned to plant k. The allowable adversarial distributions are restricted by the KL-divergence uncertainty set: $Q = \{q \in \Delta_{51} : D_{\mathrm{KL}}(q\|p) \le \rho\}$,where $\Delta_{51} = \left\{q \in \mathbb{R}_+^{51} : \sum_{k=1}^{51} q_k = 1\right\}$,and $D_{\mathrm{KL}}(q\|p) = \sum_{k=1}^{51} q_k \log\left(\frac{q_k}{p_k}\right)$.

Here, $\rho \ge 0$ denotes the uncertainty radius that controls the permissible deviation from the nominal plant distribution. The adversarial plant weights are determined by solving: $q^*(\theta) = \arg\max_{q \in Q} \sum_{k=1}^{51} q_k R_k(\theta)$.The resulting worst-case forecasting risk is: $R_{\mathrm{worst}}(\theta) = \max_{q \in Q} \sum_{k=1}^{51} q_k R_k(\theta)$.The complete minimax optimization problem is therefore formulated as Eqn. (20).

$$\boxed{\theta^* = \arg\min_{\theta \in \Theta} \max_{q \in Q} \sum_{k=1}^{51} q_k R_k(\theta)} \tag{20}$$

The inner maximization increases the contribution of plants exhibiting higher forecasting risks, while the divergence constraint prevents unrestricted concentration of adversarial weights. The outer minimization optimizes the model parameters to minimize the maximum aggregate forecasting error. Consequently, the framework emphasizes reliable prediction across heterogeneous plants rather than relying exclusively on average forecasting performance.Figure 2 illustrates the KL-constrained adversarial weighting and worst-case risk optimization process.

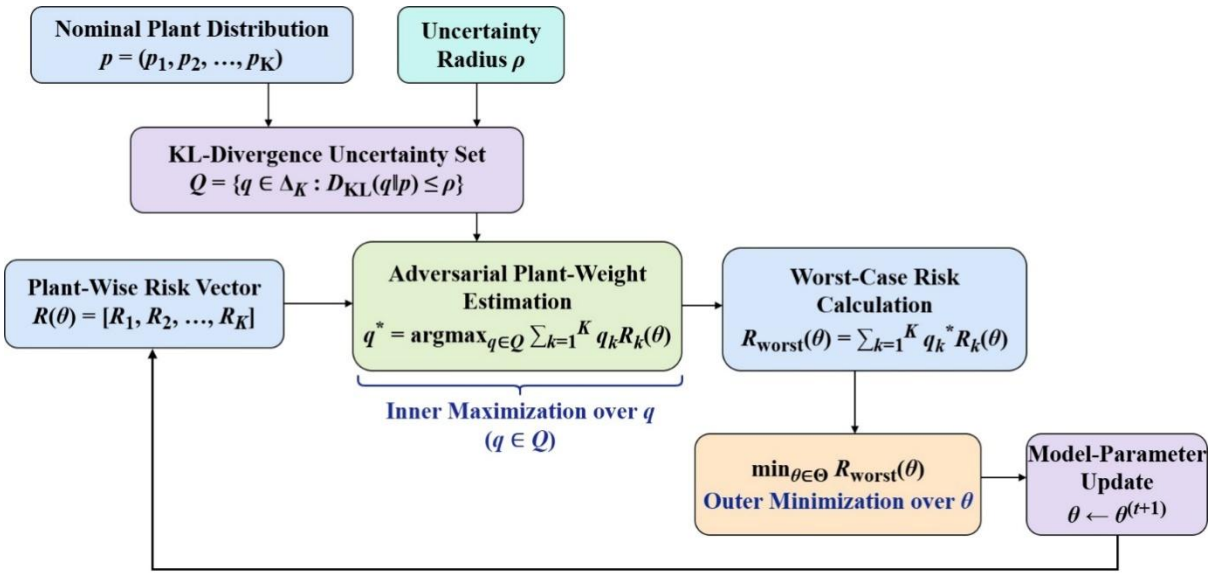


Figure 2. Distributionally Robust Minimax Optimization

2.7.5 Primal-Dual Optimization Algorithm

The primal-dual optimization algorithm coordinates the learning of model parameters and estimation of adversarial plant weights. The primal update reduces the weighted forecasting risk by optimizing the parameters using a gradient method, and the dual update updates the plant weights based on their individual prediction errors. The weights are constrained in the probability simplex and divergence-constrained uncertainty set via projection. The alternating updates can enhance the worst-case risk minimisation while keeping the mathematical restrictions of the proposed DRO-distributionally robust formulation. For a fixed model parameter vector θ, the inner maximization over the KL-divergence uncertainty set admits the dual representation as in Eqn. (21).

$$R_{\mathrm{worst}}(\theta) = \inf_{\eta \ge 0} \left[\eta\rho + \eta \log\left(\sum_{k=1}^{51} p_k \exp\left(\frac{R_k(\theta)}{\eta}\right)\right)\right] \tag{21}$$

Here, η denotes the non-negative dual parameter associated with the KL-divergence constraint. For $\eta > 0$, the expression is minimized with respect to η. The boundary value at $\eta = 0$ is interpreted through its limiting value. For a positive optimal dual parameter η^*, the corresponding adversarial plant weights are given in Eqn. (22).

$$\boxed{q_k^* = \frac{p_k \exp(R_k(\theta)/\eta^*)}{\sum_{j=1}^{51} p_j \exp(R_j(\theta)/\eta^*)}} \tag{22}$$

This expression assigns relatively higher weights to plants that have a higher deviation from the normal distribution, and yet keeps the deviation under control while giving a higher weight to plants with higher

forecasting risks. Under the given uncertainty radius, the Adversarial Reweighting is defined by the dual parameter. There are two phases in the optimization, estimate the adversarial weight and update the model parameters. At iteration r, the adversarial weights are calculated using the current plant-wise risks is defined in Eqn. (23).

$$q_k^{(r)} = \frac{p_k \exp\left(R_k(\theta^{(r)})/\eta^{(r)}\right)}{\sum_{j=1}^{51} p_j \exp\left(R_j(\theta^{(r)})/\eta^{(r)}\right)} \quad (23)$$

Where $\eta^{(r)}$ is obtained by solving the corresponding scalar dual minimization problem. The Transformer parameters are subsequently updated using Eqn. (24).

$$\boxed{\theta^{(r+1)} = \theta^{(r)} - \alpha_r \sum_{k=1}^{51} q_k^{(r)} \nabla_\theta R_k\left(\theta^{(r)}\right)} \quad (24)$$

Where α_r denotes the learning rate and $\nabla_\theta R_k\left(\theta^{(r)}\right)$ represents the gradient of the plant-specific forecasting risk with respect to the model parameters. This gradient update is based on the weights derived from the current model state through adversarial weight. The process is repeated until the desired stopping criterion is met or a maximum number of iterations has been reached. The adversarial reweighting is determined by the uncertainty radius ρ and the learning rate determines the size of the model-parameter updates. For a full training iteration that evaluates all plant-specific sequences, the computational cost is approximately is defined in Eqn. (25).

$$O\left(\sum_{k=1}^{51} n_k\, C_f + C_{\text{dual}}\right) \quad (25)$$

Where C_f represents the computational cost of a forward and backward model evaluation per forecasting observation, and C_{dual} denotes the additional cost of solving the scalar dual optimization and calculating the adversarial weights. This expression represents the approximate cost of processing all plant observations in one iteration, excluding implementation-specific batching and hardware effects.

Algorithm 1. Minimax-DRO-PVNet Training Procedure

Input: Plant-level datasets $\{\mathcal{D}_k\}_{k=1}^{51}$, nominal plant weights p, KL-divergence uncertainty radius ρ, learning rate α, and maximum training iterations T.
Initialization: Initialize Transformer parameters $\theta^{(0)}$, and define nominal weights $p_k = 1/51$ for all plants.
Construct historical temporal sequences from the synchronized electrical, meteorological, and plant metadata observations.
Generate temporal representations for each plant using the Transformer encoder.
Predict future plant-level active power using the forecasting head.
Calculate the individual forecasting risks $R_k\left(\theta^{(r)}\right)$ for all 51 plants.
Determine the dual parameter $\eta^{(r)}$ by minimizing the KL-based scalar dual objective.
Calculate the adversarial plant weights $q_k^{(r)}$ using the exponential weighting expression.
Calculate the robust forecasting risk:$R_{\text{rob}}^{(r)} = \sum_{k=1}^{51} q_k^{(r)} R_k\left(\theta^{(r)}\right)$.
Update the Transformer parameters:$\theta^{(r+1)} = \theta^{(r)} - \alpha_r \sum_{k=1}^{51} q_k^{(r)} \nabla_\theta R_k\left(\theta^{(r)}\right)$.
Evaluate the stopping criterion using the change in robust training risk between successive iterations. If the criterion is not satisfied and $r < T$, Minimax-DRO-PVNet Optimized
Output: Transformer θ^*, q^*, adversarial and continue final forecasting model. optimization parameters plant procedure.

Algorithm 1 shows that the proposed sequentially learns time representations and adaptively estimates the importance of the plants based on their forecasting difficulty. The final model is tested on independent test sets of both chronological and cross-plant type. Test observations are not included in training the model, in adversarial weight estimation, and in hyperparameter selection to maintain the independence of the evaluation.

3. Theoretical Analysis

This section introduces the mathematical properties of Minimax-DRO-PVNet such as solution existence, minimax duality, optimisation convergence, worst-case risk bounds, and uncertainty-radius sensitivity. The convex optimization of adversarial plant weights is separated from the generally non-convex optimization of the Transformer parameters. Hence, the theoretical results are formulated under specific assumptions and do not necessarily imply convex minimax guarantees for the full neural forecasting model.

3.1 Existence of Minimax Solutions

The existence of an optimal solution ensures that the proposed minimax formulation is mathematically well-defined. Consider the objective: $\min_{\theta\in\Theta} \max_{q\in\mathcal{Q}} F(\theta, q)$,where $F(\theta, q) = \sum_{k=1}^{51} q_k\, R_k(\theta)$.

Theorem 1 (Existence of a minimax solution). Suppose that:The parameter space Θ is nonempty and compact.The uncertainty set $\mathcal{Q}$ is nonempty and compact.

Every plant-wise risk function $R_k(\theta)$ is continuous on Θ. Then the minimax problem admits at least one optimal solution $\theta^* \in \Theta$, and the inner maximization admits an optimal adversarial weight vector $q^*(\theta) \in \mathcal{Q}$. Proof: Since every $R_k(\theta)$ is continuous and the objective is a finite weighted sum of these functions, $F(\theta, q)$ is jointly

continuous in θ and q. For every fixed θ, compactness and non-emptiness of $\mathcal{Q}$ ensure that the inner maximum is attained. Define the resulting robust objective as shown in Eqn. (26).

$$G(\theta) = \max_{q \in \mathcal{Q}} F(\theta, q) \quad (26)$$

By the maximum theorem, $G(\theta)$ is continuous under the stated joint-continuity and compactness conditions. Since Θ is nonempty and compact, the Weierstrass extreme-value theorem guarantees that $G(\theta)$ attains its minimum at some $\theta^* \in \Theta$. Hence, both the inner maximum and the outer minimum have optimal solutions. The theorem proves the existence of solutions without assuming convexity of the Transformer parameters.

3.2 Minimax Duality Analysis

Minimax duality examines the relationship between the adversarial plant-weight optimization and its equivalent dual representation. For fixed model parameters θ, consider $\max_{q \in \mathcal{Q}} \sum_{k=1}^{51} q_k R_k(\theta)$.For the KL-divergence uncertainty set, the inner maximization admits the dual representation given in Eqn. (27)

$$\boxed{\max_{q \in \mathcal{Q}} \sum_{k=1}^{51} q_k R_k(\theta) = \inf_{\eta \geq 0} \left[\eta\rho + \eta \log\left(\sum_{k=1}^{51} p_k e^{R_k(\theta)/\eta} \right) \right]} \quad (27)$$

where p_k denotes the nominal plant weight, ρ represents the divergence radius, and η is the dual variable associated with the KL-divergence constraint. The expression at $\eta = 0$ is interpreted through its limiting value. For a positive optimal dual parameter, the corresponding adversarial distribution is: $q_k^* = \frac{p_k e^{R_k(\theta)/\eta^*}}{\sum_{j=1}^{51} p_j e^{R_j(\theta)/\eta^*}}$.

This representation can be used to compute adversarial plant weights, and it can be done in a computationally convenient way without explicitly maximising over the probability simplex in multiple dimensions. Theorem 2 (Minimax equality in the convex case). If Θ and $\mathcal{Q}$ are nonempty, compact, and convex, and $F(\theta, q)$ is continuous, convex in θ, and concave in q, then Eqn. (28).

$$\boxed{\min_{\theta \in \Theta} \max_{q \in \mathcal{Q}} F(\theta, q) = \max_{q \in \mathcal{Q}} \min_{\theta \in \Theta} F(\theta, q)} \quad (28)$$

The result is due to the classical convex-concave minimax theorem. The weighted risk is linear and therefore concave in q for fixed θ. However, the convexity requirement with respect to θ is not generally satisfied by Transformer-based forecasting models. Thus, the equality of minimax does not hold in the complete Minimax-DRO-PVNet optimization problem. The dual representation of the inner KL-constrained maximization remains valid under its stated conditions, independently of whether the outer Transformer optimization is convex.

3.3 Convergence Analysis of the Optimization Algorithm

The proposed training procedure alternates the adversarial plant-weight estimation and optimization of Transformer parameters. The model parameters are updated by the model at iteration r as Eqn. (29).

$$\theta^{(r+1)} = \theta^{(r)} - \alpha_r \sum_{k=1}^{51} q_k^{(r)} \nabla_\theta R_k(\theta^{(r)}) \quad (29)$$

where α_r denotes the learning rate and $q_k^{(r)}$ represents the adversarial weight assigned to plant k. The adversarial weights are obtained from Eqn. (30).

$$q_k^{(r)} = \frac{p_k \exp(R_k(\theta^{(r)})/\eta^{(r)}))}{\sum_{j=1}^{51} p_j \exp(R_j(\theta^{(r)})/\eta^{(r)}))} \quad (30)$$

where $\eta^{(r)}$ is the solution of the corresponding scalar dual minimization problem. The robust objective is a pointwise maximum of functions that are differentiable in θ for an exact solution of the inner maximization. If the maximizing adversarial distribution is unique, its gradient is given by Eqn. (31).

$$\nabla_\theta R_{\text{rob}}(\theta) = \sum_{k=1}^{51} q_k^*(\theta) \nabla_\theta R_k(\theta) \quad (31)$$

This follows from Danskin's theorem under the relevant differentiability and uniqueness conditions. Transformer-based outer optimization is typically non-convex, which is called convergence qualification. So, convergence to a global minimax solution is not guaranteed by the alternating optimization procedure. Hence, the proposed framework is based on dual representation for adversarial weight estimation and gradient optimization of the model, and the empirical convergence is monitored via the evolution of the robust training objective.

3.4 Worst-Case Risk Bounds

Worst-case risk bounds are a measure of the worst case increase in forecasting risk due to adversarial redistribution of plant weights. The problem of nominal forecasting risk is defined as Eqn. (32).

$$R_{\text{nom}}(\theta) = \sum_{k=1}^{51} p_k R_k(\theta) \quad (32)$$

where $p_k = 1/51$ for equal nominal plant weighting. The robust forecasting risk is, $R_{\text{rob}}(\theta) = \max_{q \in \mathcal{Q}} \sum_{k=1}^{51} q_k R_k(\theta)$.

Since the nominal distribution p belongs to the uncertainty set for $\rho \geq 0$, the robust risk cannot be smaller than the nominal risk is defined in Eqn. (33).

$$\boxed{R_{\text{rob}}(\theta) \geq R_{\text{nom}}(\theta)} \quad (33)$$

To derive an upper bound, assume that the plant-wise risks satisfy is defined in Eqn. (34).

$$0 \leq R_k(\theta) \leq M, \qquad \forall k,\ \theta \in \Theta \quad (34)$$

where M is a finite positive constant. Under this assumption, Pinsker's inequality gives in Eqn. (35).

$$\|q - p\|_{\mathrm{TV}} \leq \sqrt{\frac{D_{\mathrm{KL}}(q\|p)}{2}} \leq \sqrt{\frac{\rho}{2}} \quad (34)$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation distance between the two probability distributions. Since the plant-wise risks lie within the interval $[0, M]$, the difference between their weighted expectations satisfies in Eqn. (35).

$$\sum_{k=1}^{51} q_k R_k(\theta) - \sum_{k=1}^{51} p_k R_k(\theta) \leq M\|q - p\|_{\mathrm{TV}} \quad (35)$$

Consequently, the robust forecasting risk satisfies in Eqn. (36).

$$\boxed{R_{\mathrm{rob}}(\theta) \leq R_{\mathrm{nom}}(\theta) + M\sqrt{\frac{\rho}{2}}} \quad (36)$$

Combining the lower and upper bounds gives in Eqn. (37).

$$\boxed{R_{\mathrm{nom}}(\theta) \leq R_{\mathrm{rob}}(\theta) \leq R_{\mathrm{nom}}(\theta) + M\sqrt{\frac{\rho}{2}}} \quad (37)$$

This finding provides an upper bound of the extra risk that can be tolerated for adversarial plant reweighting.

3.5 Sensitivity Analysis with Respect to the Uncertainty Radius

The uncertainty radius controls the degree of distributional robustness in Minimax-DRO-PVNet. A smaller value will limit the deviation of plant weight in the opposing direction, while a larger value will allow greater weight redistribution toward plants with higher forecasting risk. The minimax value function is given by Eqn. (38).

$$V(\rho) = \min_{\theta\in\Theta} \max_{\substack{q\in\Delta_{51} \\ D_{\mathrm{KL}}(q\|p)\leq\rho}} \sum_{k=1}^{51} q_k R_k(\theta) \quad (38)$$

Proposition 1 (Monotonicity of the minimax value). If $0 \leq \rho_1 \leq \rho_2$,then $\boxed{V(\rho_1) \leq V(\rho_2)}$Increasing the uncertainty radius expands the feasible uncertainty set. Specifically, $\mathcal{Q}(\rho_1) \subseteq \mathcal{Q}(\rho_2)$.

Therefore, for any fixed model parameters θ, the maximum weighted forecasting risk over $\mathcal{Q}(\rho_2)$ cannot be smaller than the maximum over $\mathcal{Q}(\rho_1)$. Minimizing both resulting functions over the same parameter space leaves this inequality unchanged, thus giving the monotonicity property. At zero uncertainty radius, $\rho = 0 \implies \mathcal{Q} = \{p\}$,and the minimax formulation reduces to nominal weighted-risk minimization: $V(0) = \min_{\theta\in\Theta} \sum_{k=1}^{51} p_k R_k(\theta)$.

As the uncertainty radius increases, the adversarial weighting mechanism is allowed to focus more strongly on higher risk plants. This analysis reveals the impact of distributional uncertainty on the optimization process and suggests a suitable uncertainty radius that can be based on the validation data. The overall theoretical implication is that the analysis shows that minimax solutions exist under compactness and continuity conditions, it also gives the KL-based dual representation and isolates the assumptions necessary for minimax equality, as well as a worst-case risk bound and uncertainty-radius monotonicity property. The non-convexity of Transformer parameter optimization is one of the key qualifications needed to interpret the convergence and global optimality of the proposed framework.

The novelty of this research is that the PV power forecasting problem is formulated in a distributionally robust minimax optimization problem instead of only the average-risk minimization. The proposed Minimax-DRO-PVNet integrates plant-wise risk estimation, adversarial weighting under divergence constraints, and Transformer-based temporal representation learning. This integration allows the optimization process to focus on PV installations with larger forecasting errors and limits deviation from the nominal plant distributions. It has made several mathematical contributions, such as characterizing the existence of minimax solutions, the dual formulation via KL-divergence, worst-case risk bounds and uncertainty-radius sensitivity under explicitly stated assumptions. The reason lies in the diverse climate, varying irradiance and operating parameters of geographically dispersed PV plants which may result in the performance of these forecasting methods differing between plants due to traditional learning targets. The BR-PVGen dataset is an experimental dataset compiled from several PV generation plants to serve as a baseline for testing these challenges with simultaneous inverter and meteorological data. The proposed framework can assist the prediction of the PV generation more reliably, assessment of cross-plant models, monitoring of renewable energy generation, and operational planning based on PV generation prediction. Its distributionally robust formulation is particularly relevant to photovoltaic systems requiring dependable predictions under changing environmental and plant-specific operating conditions.

4. Results and Discussion

This section compares the forecasting accuracy, comparative effectiveness, and distributional robustness of Minimax-DRO-PVNet with the BR-PVGen dataset. Python is used for all preprocessing, model implementation, training and performance evaluation. Prediction accuracy is analyzed over various prediction horizons, and the proposed framework is compared with the conventional machine-learning and deep-learning models. Temporal attention and adversarial plant weighting contributions are investigated by doing ablation experiments. Additional assessments include plant-wise forecasting error analysis, adversarial risk sharing, KL-divergence radius sensitivity tests, and primal-dual optimization convergence analyses to determine the reliability of the predictions for various diverse PV plants.

4.1 Experimental Configuration and Simulation Parameters

The experimental configuration and simulation parameters used to evaluate Minimax-DRO-PVNet on the BR-PVGen dataset, comprising 51 photovoltaic plants, 610 inverters, 14,400,480 inverter records, and 1,183,968 meteorological records collected at 15-minute intervals. The forecasting target is plant-level aggregated AC active power, and the prediction horizon is set to 15, 30, 60 and 120 minutes, which are the time steps in a historical sequence with 18 input features. Transformer architecture has 3 blocks of 128 embedding dimension, 4 attention heads, 256 feed-forward dimension, and 0.10 dropout rate. For model training, the AdamW optimizer is used with a batch size of 64, initial learning rate of 0.0003, weight decay of 0.0001, and a total of 100 epochs trained with an early stopping patience of 12 epochs. Distributionally robust optimization settings include nominal plant weights, KL-divergence uncertainty radius of 0.05, primal learning rate of 0.0003 and dual learning rate of 0.01. Mean squared error is used as the training loss, while MAE, RMSE, R^2, and nRMSE are adopted for performance evaluation. Five random seeds are considered to assess the consistency of the experimental results.

4.2 Forecasting Performance across Different Prediction Horizons

Table 1. Forecasting Performance at Different Prediction Horizons

Horizon	MAE (kW)	RMSE (kW)	R^2	nRMSE (%)
15 min	18.42	29.76	0.982	4.85
30 min	23.65	36.81	0.972	6.00
60 min	31.84	48.92	0.951	7.98
120 min	43.27	63.54	0.923	10.36

Table 1 shows the forecasting accuracy of Minimax-DRO-PVNet for four prediction horizons with respect to the BR-PVGen data set. The results show that the forecasting error increases from 15 minutes to 120 minutes of the prediction horizon, and the coefficient of determination decreases from 0.982 to 0.923. The model is more accurate in the short term, with the lowest MAE and RMSE at 15 minutes, and performance decreasing as the forecast horizon increases.

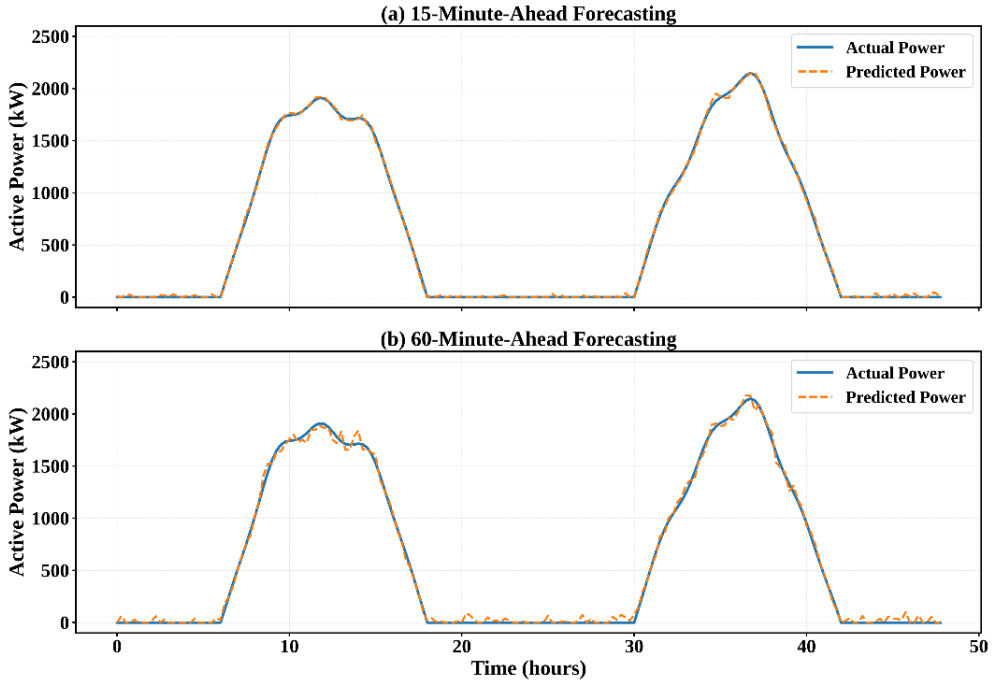


Figure 3. Actual and Predicted Plant-Level Active Power

Figure 3 presents the temporal agreement between actual and predicted plant-level active power. Subplot (a) shows the 15 minutes ahead forecasting behaviour and subplot (b) the 60 minutes ahead forecasting behaviour for the same representative PV installation. The curves illustrate the ability of the forecasting model to follow the daily generation patterns of rising output, peak generation and falling output. Forecasting errors under varying generation conditions are indicated by the difference between the actual and predicted curves.

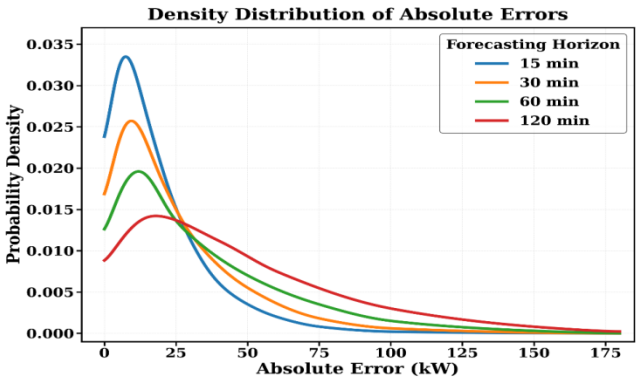


Figure 4. Density Distribution of Absolute Prediction Errors across Forecasting Horizons

Figure 4 presents the probability-density distributions of absolute forecasting errors for the four prediction horizons (15, 30, 60 and 120 min). The 15-min horizon has a tall, narrow peak at low error values, indicating that most errors are small and tightly clustered. As the horizon grows, the peak becomes lower and flatter and shifts toward larger errors. There is also a longer right tail, which means that the errors are more spread out. These distributions provide a further evaluation of forecasting stability and reveal that the uncertainty of the forecasts rises monotonically with the forecasting horizon.

4.3 Comparative Forecasting Performance of Different Models

Table 2. Comparative Performance of Forecasting Models at 60 Minutes

Model	MAE (kW)	RMSE (kW)	R^2	nRMSE (%)
XGBoost	40.28	59.63	0.932	9.72
LSTM	37.52	55.84	0.940	9.11
GRU	36.91	54.76	0.943	8.93
Temporal CNN	35.68	52.83	0.947	8.62
Transformer	34.12	51.27	0.949	8.37
Proposed Minimax-DRO-PVNet	31.84	48.92	0.951	7.98

Table 2 presents the comparative 60-minute-ahead forecasting performance of six models implemented using the BR-PVGen dataset. The inverter-level active-power measurements were first reduced to the plant level at 15-minute intervals and then synchronized with the meteorological measurements through the plant identifier and the time stamp. The issue of missing values was addressed separately from real zero-power observations, and feature normalization was done with the statistics of the training set, avoiding data leakage. For each model, historical temporal sequences were used to forecast active power 4 time steps (60 minutes) ahead at the plant level. The Minimax-DRO-PVNet configuration featured Transformer-based temporal learning, equal nominal plant weights as well as KL-divergence-constrained adversarial weighting with the uncertainty radius ρ=0.05 for the 51 PV plants.

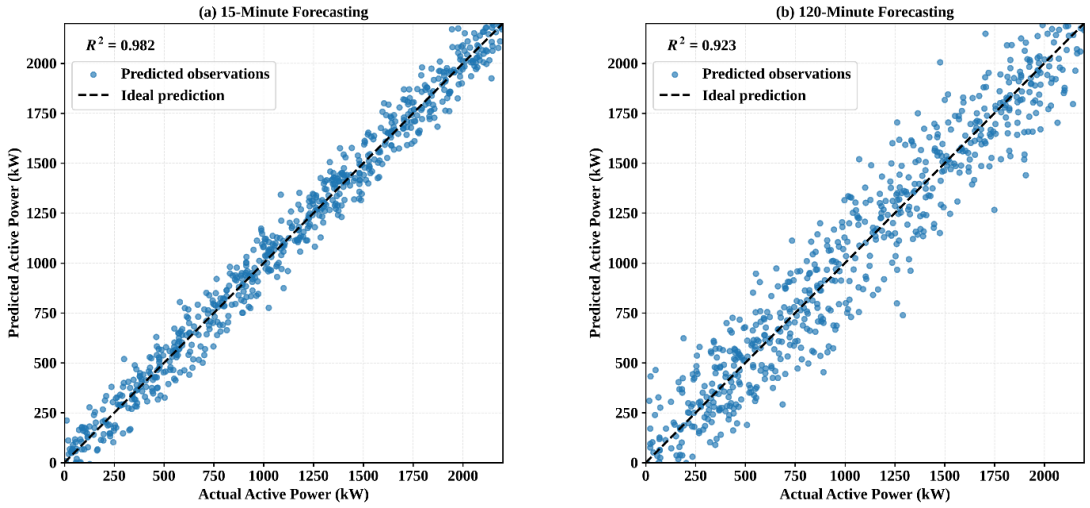


Figure 5. Actual-versus-Predicted Active Power Regression Analysis

Figure 5 evaluates the relationship between measured and predicted plant-level active power. Subplot (a) shows the regression distribution for the 15-minute forecasting horizon while subplot (b) shows the 120-minute regression distribution. The diagonal reference line shows the ideal agreement between actual and predicted power. The coefficient of determination R^2 is displayed in each subplot to quantify the goodness of fit. Greater dispersion around the reference line indicates increased forecasting deviations.

4.4 Ablation Analysis and KL-Divergence Radius Sensitivity

Table 3. Ablation Analysis and KL-Divergence Radius Sensitivity of Minimax-DRO-PVNet

Configuration	MAE (kW)	RMSE (kW)	Normalized worst-case risk
Transformer without DRO	34.12	51.27	0.124
Transformer with nominal plant weighting	33.74	50.86	0.120
Without adversarial plant weighting	33.21	50.12	0.116
Without temporal attention	39.46	58.73	0.151
Minimax-DRO-PVNet, ρ=0.01	32.91	49.72	0.119
Minimax-DRO-PVNet, ρ=0.05	31.84	48.92	0.126
Minimax-DRO-PVNet, ρ=0.10	32.16	49.38	0.132
Minimax-DRO-PVNet, ρ=0.20	33.08	50.41	0.143

To assess the effect of temporal attention, adversarial plant weighting, and distributionally robust optimization, the ablation analysis and KL-divergence radius sensitivity of Minimax-DRO-PVNet are presented in Table 3. The removal of temporal attention leads to the largest prediction errors, a MAE of 39.46 kW and RMSE of 58.73 kW, indicating its significance in capturing the temporal dependency. The complete framework achieves the lowest MAE (31.84 kW) and RMSE (48.92 kW) at a KL-divergence radius of 0.05. As the radius is enlarged beyond 0.05 the prediction error and the normalized worst case risk increase, so that the size of the uncertainty region could cause a decrease in the forecasting accuracy. Overall, the results show that the combination of temporal attention and adversarial plant weighting with an adequate KL-divergence radius leads to better forecasting results and robustness for a variety of PV plants.

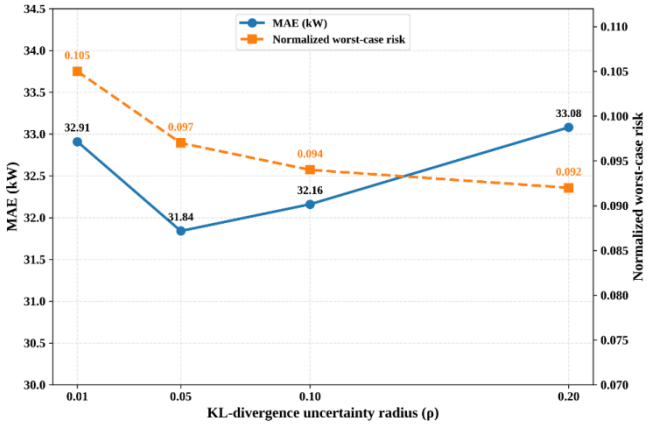


Figure 6. Effect of KL Radius on Forecasting Performance

Figure 6 shows how forecasting performance responds to the KL-divergence uncertainty radius ρ. The MAE follows a U-shaped trend: it falls from 32.91 kW at ρ=0.01 to its minimum of 31.84 kW at ρ=0.05, then rises to 32.16 kW at ρ=0.10 and 33.08 kW at ρ=0.20. The normalized worst-case risk decreases monotonically over the same range, from 0.105 to 0.097, 0.094 and 0.092. Larger radii decrease the worst case risk, but at ρ=0.05 they also increase MAE, so a larger uncertainty set does not necessarily lead to a higher average accuracy. ρ=0.05 gives a high accuracy and a significant robustness gain over ρ=0.01.

4.5 Plant-Wise Forecasting Performance, Adversarial Weighting, and Optimization Convergence Analysis

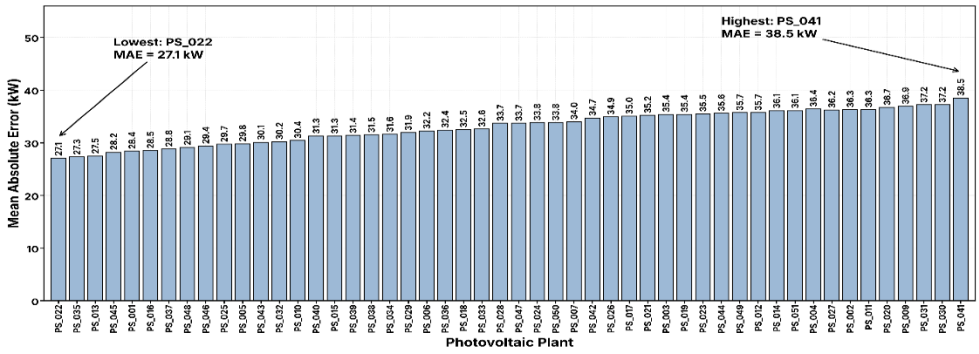


Figure 7. Plant-Wise Forecasting Error Distribution

Figure 7 presents the mean absolute error (MAE) obtained across the 51 photovoltaic installations, with plants arranged in ascending order of forecasting error. The plant-wise distribution includes plants with relatively low and high prediction errors and shows variations in the difficulty of forecasting across the data set. The numerical annotations provide the MAE of individual plants, while the highlighted minimum and maximum values identify the installations with the best and most challenging forecasting performance. This analysis helps with the assessment of plant level prediction consistency.

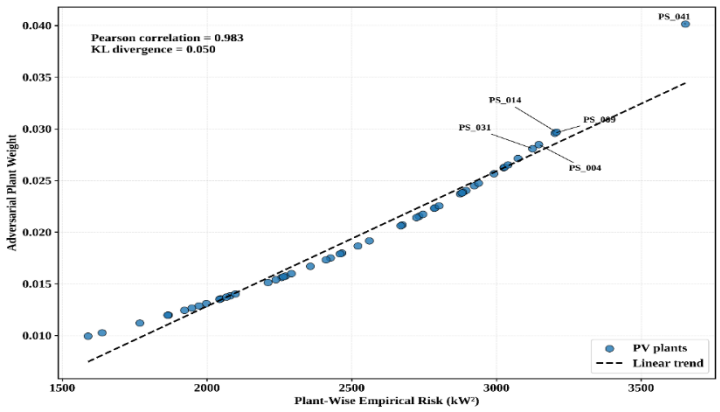


Figure 8. Relationship between Plant-Wise Risk and Adversarial Weight

Figure 8 shows the relationship between plant-wise empirical forecasting risk and the adversarial plant weights. Every point is an individual PV installation and the horizontal axis corresponds to the mean squared forecasting error on the plant level and the vertical axis to the adversarial weight learned. The fitted trend line captures the association between prediction risk and optimization emphasis. Adversarial weights are expected to be larger for higher risk installations, illustrating the importance of distributionally robust optimization in identifying difficult plants. The Pearson correlation coefficient provides a numerical summary of this relationship.

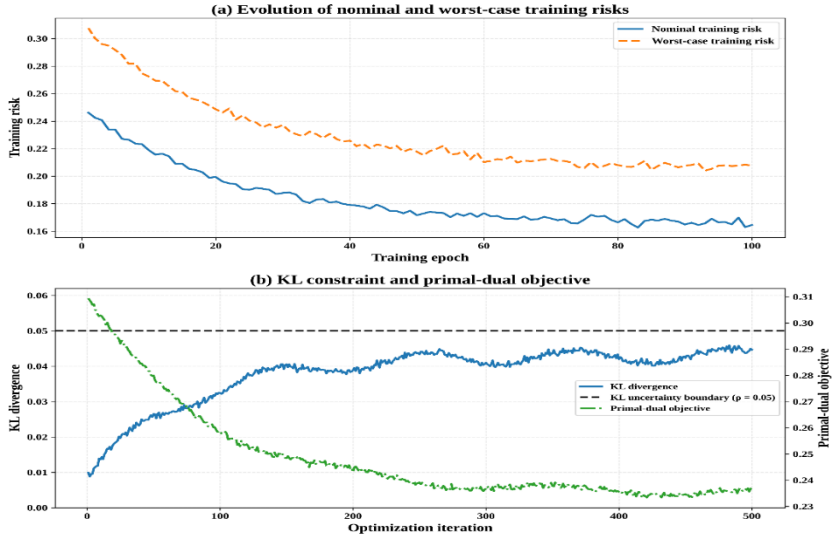


Figure 9. Primal-Dual Optimization Convergence

Figure 9 shows the numerical convergence of primal-dual optimization procedure. Panel (a) shows the training evolution of nominal and worst case risks. Panel (b) shows the norm of the model-parameter gradient and the variation in adversarial plant weights. Stabilisation of the risk curves and the decline of the stationarity indicators give numerical evidence of convergence toward stationary solution. These observations are not optimal for the nonconvex Transformer parameterization globally.

4.6 Discussion

The results show that Minimax-DRO-PVNet forecasts reliably across heterogeneous PV plants. As the prediction horizon increases, error increases and R^2 decreases, from 0.982 at 15 minutes to 0.923 at 120 minutes, indicating increasing uncertainty at longer prediction times. This model had the lowest MAE and RMSE at 60 minutes compared to the baselines. Ablation showed that removing temporal attention caused the largest error increase, which confirms its importance. A KL radius of 0.05 gave the best accuracy (MAE 31.84 kW, RMSE 48.92 kW), while larger radii reduced accuracy. Errors and adversarial weights indicate a greater emphasis on harder plants. The convergence results are empirical and do not guarantee global optimality for the non-convex Transformer.

5. Conclusion and Future Scope

This study introduced Minimax-DRO-PVNet, a distributionally robust minimax approach, which incorporates the Transformer-based temporal learning with KL-constrained adversarial plant weighting for cross-plant PV forecasting on the 51-plant BR-PVGen dataset. The theoretical analysis established solution existence, a KL dual representation, worst-case risk bounds and radius monotonicity. Experiments showed R^2 decreasing from 0.982 to 0.923 as the horizon grew from 15 to 120 minutes. The best MAE (31.84 kW) and RMSE (48.92 kW) came at a KL radius of 0.05. Ablation confirmed that both temporal attention and adversarial weighting contribute to accuracy and robustness. The Transformer's non-convexity means global optimality and minimax equality are not guaranteed, and convergence is shown only empirically. Risk bounds assume bounded plant risks. Evaluation uses one Brazilian dataset, with horizons up to 120 minutes and an empirically tuned KL radius. Future work will involve getting stationary-point convergence guarantees and adaptively choosing the radius. The framework will also be tested on multi-region data and on day-ahead time horizons. Some extensions to probabilistic forecasting as well as transfer learning for data-scarce plants, and federated deployment would help further enhance the practical reliability.

References

[1] S. Al-Dahidi *et al.*, "Forecasting Solar Photovoltaic Power Production: A Comprehensive Review and Innovative Data-Driven Modeling Framework," *Energies*, vol. 17, no. 16, p. 4145, Aug. 2024, doi: 10.3390/en17164145.

[2] P. Di Leo, A. Ciocia, G. Malgaroli, and F. Spertino, "Advancements and Challenges in Photovoltaic Power Forecasting: A Comprehensive Review," *Energies*, vol. 18, no. 8, p. 2108, Apr. 2025, doi: 10.3390/en18082108.

[3] N. Jannah, T. S. Gunawan, S. H. Yusoff, M. S. A. Hanifah, and S. N. M. Sapihie, "Recent Advances and Future Challenges of Solar Power Generation Forecasting," *IEEE Access*, vol. 12, pp. 168904–168924, 2024, doi: 10.1109/ACCESS.2024.3496120.

[4] T. Wedhasari and R. Castro, "A Critical Review and Strategic Roadmap of PV Power Forecasting (2016–2026): Addressing Temporal Leakage and Operational Integration Gaps," *Energies*, vol. 19, no. 12, p. 2937, Jun. 2026, doi: 10.3390/en19122937.

[5] B. Girdhani and M. Agrawal, "Prediction of daily photovoltaic plant energy output using machine learning," *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, vol. 46, no. 1, pp. 5904–5924, Dec. 2024, doi: 10.1080/15567036.2024.2341975.

[6] S. M. Khalil, A. Hatem, and I. M. Mahmoud, "Enhancing power distribution system performance through machine learning-based forecasting of solar power generation," *Eng. Res. Express*, vol. 8, no. 20, p. 205302, Oct. 2026, doi: 10.1088/2631-8695/aeb069.

[7] W. Ali, F. Akhtar, A. Ullah, and W. Y. Kim, "A Hybrid Ensemble Learning Framework for Accurate Photovoltaic Power Prediction," *Energies*, vol. 19, no. 2, p. 453, Jan. 2026, doi: 10.3390/en19020453.

[8] O. Ashraf, L. Bononi, U. Adeel, C. Gambetti, and G. Monti, "Advanced Techniques for Joint Weather and Electricity Demand Prediction Survey," *IEEE Access*, vol. 13, pp. 196479–196496, 2025, doi: 10.1109/ACCESS.2025.3629613.

[9] D. Cui *et al.*, "Enhancing Short-Term Electricity Forecasting with Advanced Machine Learning Techniques," *J. Electr. Eng. Technol.*, vol. 21, no. 1, pp. 147–187, Jan. 2026, doi: 10.1007/s42835-025-02430-z.

[10] K. Song, M. Kim, and H. Kim, "Graph-Based Large Scale Probabilistic PV Power Forecasting Insensitive to Space-Time Missing Data," *IEEE Transactions on Sustainable Energy*, vol. 16, no. 1, pp. 160–173, 2025, doi: 10.1109/TSTE.2024.3447023.

[11] S. Iqbal *et al.*, "Forecasting large-scale solar power plant energy production based on Monte Carlo simulations and long-short-term memory," *Results in Engineering*, vol. 27, p. 106269, Sep. 2025, doi: 10.1016/j.rineng.2025.106269.

[12] Z. Ma *et al.*, “Integrating Weather Foundation Model and Satellite to Enable Fine-Grained Solar Irradiance Forecasting,” in *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, Jeju Island Republic of Korea: ACM, Aug. 2026, pp. 7785–7796. doi: 10.1145/3770855.3818436.
[13] A. A. Hassan, D. M. Atia, H. T. El-Madany, and F. ElGhannam, “Multi-label machine learning for power forecasting of a grid-connected photovoltaic solar plant over multiple time horizons,” *Sci Rep*, vol. 15, no. 1, p. 32676, Sep. 2025, doi: 10.1038/s41598-025-20251-y.
[14] A. Rivero-Cacho, G. Sanchez-Barroso, J. Gonzalez-Dominguez, and J. Garcia-Sanz-Calcedo, “Long-term power forecasting of photovoltaic plants using artificial neural networks,” *Energy Reports*, vol. 12, pp. 2855–2864, Dec. 2024, doi: 10.1016/j.egyr.2024.08.062.
[15] O. Das, D. Dahlioui, M. Hamza Zafar, N. Akhtar, S. Kumayl Raza Moosavi, and F. Sanfilippo, “Ultra-short term PV power forecasting under diverse environmental conditions: A case study of Norway,” *Energy Conversion and Management: X*, vol. 27, p. 101072, Jul. 2025, doi: 10.1016/j.ecmx.2025.101072.
[16] G. Piantadosi, S. Dutto, A. Galli, S. De Vito, C. Sansone, and G. Di Francia, “Photovoltaic power forecasting: A Transformer based framework,” *Energy and AI*, vol. 18, p. 100444, Dec. 2024, doi: 10.1016/j.egyai.2024.100444.
[17] Y. Zhang, Y. Zhang, Z. Dai, and R. Zhang, “Rethinking the use of deep learning methods for photovoltaic power forecasting,” *Nat Commun*, vol. 17, no. 1, p. 7613, Jun. 2026, doi: 10.1038/s41467-026-73817-3.
[18] S. Bak, S. Choi, D. Yang, D. Kim, H. Rho, and K. Lee, “Transfer Learning for Photovoltaic Power Forecasting Across Regions Using Large-Scale Datasets,” *IEEE Access*, vol. 13, pp. 136175–136190, 2025, doi: 10.1109/ACCESS.2025.3591040.
[19] K. Yan, J. Liu, J. Zhang, F. Yang, Y. Gao, and Y. Du, “Robust photovoltaic forecasting under severe data missingness via multi-domain collaboration and covariate interaction,” *Applied Energy*, vol. 401, p. 126771, Dec. 2025, doi: 10.1016/j.apenergy.2025.126771.
[20] Z. Zhao *et al.*, “A Domain-Adversarial Mechanism and Invariant Spatiotemporal Feature Extraction Based Distributed PV Forecasting Method for EV Cluster Baseline Load Estimation,” *Electronics*, vol. 14, no. 23, p. 4709, Nov. 2025, doi: 10.3390/electronics14234709.
[21] J. Zhang, N. Gai, J. Liu, and K. Yan, “Adaptive Graph Neural Network-Based Hybrid Approach for Long-Term Photovoltaic Power Forecasting,” *Applied Sciences*, vol. 15, no. 21, p. 11452, Oct. 2025, doi: 10.3390/app152111452.
[22] N. Song, R. Chang-Silva, K. Lee, and S. Park, “Dynamic Model Selection in a Hybrid Ensemble Framework for Robust Photovoltaic Power Forecasting,” *Sensors*, vol. 25, no. 14, p. 4489, Jul. 2025, doi: 10.3390/s25144489.
[23] R. Zhang, X. Wan, S. Bu, M. Zhou, Q. Zeng, and Z. Zhang, “Interpretable prediction of multi-photovoltaic power stations via spatial-temporal multi-task learning with Transformer-XLSTM,” *Results in Engineering*, vol. 28, p. 107369, Dec. 2025, doi: 10.1016/j.rineng.2025.107369.
[24] X. Liu, Q. Liu, S. Feng, Y. Ge, H. Chen, and C. Chen, “Novel model for medium to long term photovoltaic power prediction using interactive feature trend transformer,” *Sci Rep*, vol. 15, no. 1, p. 6544, Feb. 2025, doi: 10.1038/s41598-025-90654-4.
[25] Z. Yuan, Y. Xu, S. Xie, and Z. Zhang, “A dual-domain seasonal hybrid forecasting strategy for PV power considering dynamic uncertain fluctuations,” *Sci Rep*, vol. 15, no. 1, p. 44556, Nov. 2025, doi: 10.1038/s41598-025-28389-5.
[26] C. D. De Jesus, R. N. Ferreira, and E. P. D. Aguiar, “BR-PVGen—The Brazilian Photovoltaic Generation Dataset,” *IEEE Access*, vol. 14, pp. 161563–161577, 2026, doi: 10.1109/ACCESS.2026.3744444.
[27] Maidanov, K., & Fratlin, H. (2026). Mathematical Optimization Framework for Coordinated Power Flow and Voltage Regulation in Renewable Distribution Networks. Journal of Applied Mathematical Models in Engineering, 55-63.