

Article

Multi-Model Stacking Ensemble with Multi-Perspective Interpretability Analysis for Solar Power Forecasting

Shijie Wu ¹, Dejing Lin ^{2,3,4,*} and Yushuai Zhang ^{5,*}¹ Department of Economics, Faculty of Social Sciences, Lingnan University, Hong Kong; wushijie25fz@163.com² College of Artificial Intelligence, Yango University, Fuzhou 350015, China³ Fujian Key Laboratory of Spatial Information Perception and Intelligent Processing, Yango University, Fuzhou 350015, China⁴ Fujian Jofee Technology Co., Ltd., Fuzhou 350015, China⁵ School of Electronic Information, Xijing University, Xi'an 710123, China

* Correspondence: dejinglin@ygu.edu.cn (D.L.); 20240241@xijing.edu.cn (Y.Z.)

Abstract

Accurate photovoltaic (PV) power forecasting is important for grid dispatch, energy-storage management, and electricity-market trading. This paper presents a PV power forecasting framework that combines multi-model ensemble learning with multi-perspective interpretability analysis. A 22-dimensional feature set was constructed from irradiance, meteorological, temporal, and lagged-power variables. Five supervised regressors, a zero-shot Chronos-Bolt-Small time-series foundation model, and six additional forecasting baselines were evaluated at two PV plants. The three ensemble schemes used only the five supervised regressors and were compared on the final 20% of the 2019 development data; the complete 2020 period was used only for final evaluation. The selected methods achieved normalized RMSE values of 0.0350 and 0.0376 at Sites 1 and 2, respectively. SHAP, PDP/ICE, LIME, and permutation importance were applied to the ensemble model at each site. Recent power-history features dominated this one-step-ahead forecasting task with a 15 min horizon, while irradiance features provided additional site-dependent information.

Keywords: photovoltaic power forecasting; stacking ensemble learning; explainable artificial intelligence; SHAP; partial dependence plot; LIME; multi-site

1. Introduction

The rapid deployment of PV has increased the operational importance of accurate short-term forecasting. PV output varies with irradiance, cloud evolution, and seasonal and diurnal cycles [1,2]. Forecasts support unit commitment, reserve allocation, energy-storage scheduling, and electricity-market bidding [3], and improved forecast accuracy can assist market operation. Accordingly, this study focuses on forecasting accuracy, a sound validation procedure, and reliable model interpretation.

According to differences in modeling mechanisms, existing PV power forecasting methods can be grouped into three categories: physical methods, statistical methods, and machine-learning methods [1,4]. Physical methods are based on numerical weather prediction (NWP) and the analytical modeling of PV modules; they have a clear physical meaning and strong interpretability, but are highly sensitive to the accuracy of meteorological inputs and have difficulty characterizing the rapid movement of cloud clusters at ultra-short-term scales. Statistical methods, represented by the autoregressive integrated moving average (ARIMA) model, are simple to build and computationally efficient, but



Academic Editor: Georg Frey

Received: 12 June 2026

Revised: 16 July 2026

Accepted: 21 July 2026

Published: 27 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

their linear assumptions struggle to represent the strongly nonlinear mapping between irradiance and power [2,5].

In recent years, data-driven machine-learning methods have demonstrated significant advantages in PV power forecasting owing to their powerful nonlinear modeling capability [4,6]. Common model families include gradient-boosted decision trees such as extreme gradient boosting (XGBoost) [7] and the light gradient boosting machine (LightGBM) [8], random forests (RF) [9], support vector regression (SVR), and long short-term memory (LSTM) networks [10]. Solar-forecasting reviews document applications of these methods across diverse horizons and datasets [4,6,11], while also emphasizing that comparative performance depends on the dataset, forecast horizon, input variables, and evaluation design [4,11]. However, a single model often possesses an inherent inductive bias toward a specific data distribution: tree models excel at capturing thresholds and interactions among features, neural networks are good at characterizing continuous smooth mappings, and kernel methods perform well in local neighborhoods or with small samples. Under different irradiance conditions and different plant characteristics, a single model struggles to simultaneously achieve high forecasting accuracy and robust generalization; this challenge is especially relevant in cross-site applications.

To overcome the performance bottleneck of a single model, ensemble learning, which integrates multiple heterogeneous base learners to optimize the bias–variance trade-off, has been widely shown to significantly improve forecasting robustness [12]. Among ensemble approaches, stacked generalization (stacking) [13] performs second-stage learning on the outputs of the base learners through a meta-learner and can theoretically achieve performance superior to any single base model; it has been widely applied in solar forecasting in recent years [14–16]. For example, Lateko et al. [14] used a recurrent neural network (RNN) as the meta-learner in a stacking ensemble for short-term PV power forecasting and reported improved accuracy; Nguyen et al. [15] used the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm with bounds (L-BFGS-B) to optimize the weights of base learners, achieving performance superior to deep-learning baselines on a large-scale plant dataset; Islam et al. [16] further built an interpretable stacking ensemble model for solar-irradiance forecasting. However, stacking faces unique challenges in time-series scenarios: if ordinary K-fold cross-validation is used, future information leakage occurs; when time-series cross-validation (TimeSeriesSplit) is used, the early folds have insufficient training samples, leading to noisy meta-features; meanwhile, nighttime PV power is commonly zero, and zero-filled unavailable meta-features cannot be distinguished from genuine nighttime zeros; in addition, an unconstrained linear meta-learner may yield negative weights, causing the base models to pull against one another and weakening the ensemble effect. How to construct a time-series ensemble scheme that both avoids information leakage and remains robust and reliable still leaves considerable room for improvement. Stacking, blending, and constrained weighted averaging are established methods; the present framework differs by combining expanding-window out-of-fold prediction, NaN masking of unavailable meta-features, non-negative simplex constraints, and validation-only selection in one time-ordered PV forecasting workflow.

Recent solar and PV studies cover machine-learning forecasting comparisons and post hoc interpretability analyses [17–22]. Chronos-Bolt can generate zero-shot forecasts from historical sequences and is therefore evaluated as a modern baseline. SHAP provides additive feature attribution; PDP and ICE summarize average and sample-level marginal responses [23,24]; and LIME fits a local interpretable surrogate around an individual prediction [25]. This study combines these complementary views, reports the consistency of feature rankings across methods, and reports local fidelity for LIME.

Although the above studies have made considerable progress, three shortcomings remain. First, most works adopt only a single XAI method and lack a systematic comparison of the consistency among different methods; relying on a single method may yield one-sided or even misleading conclusions. Second, few studies analyze the spatial differences in feature importance from a multi-site perspective, making it difficult to answer which factors remain stable under cross-site transfer and which vary significantly with climate conditions—a question crucial to unified regional modeling. Third, a unified framework that organically integrates a robust time-series ensemble with multi-perspective interpretability analysis is still rare.

To address these shortcomings, this paper proposes a PV power forecasting framework integrating a multi-model weighted ensemble with multi-perspective interpretability analysis and evaluates it on two PV plants with different installed capacities and observed data characteristics. The main contributions of this paper are summarized as follows:

- (1) A time-series ensemble-selection framework is proposed. It combines time-series cross-validation and NaN masking with a non-negative simplex-constrained weighted-averaging candidate to reduce nighttime-zero contamination and information leakage, while selecting among stacking, weighted averaging, and blending using validation data only.
- (2) A multi-perspective interpretability framework integrating SHAP, PDP/ICE, and LIME is developed. Kendall τ and Spearman ρ are used to quantify feature-ranking consistency across XAI methods, providing a quantitative basis for multi-method cross-validation.
- (3) A cross-site interpretability comparison is conducted, revealing the spatial stability of the importance of power-history features and the spatial non-stationarity of the importance of irradiance features, thereby providing a quantitative basis for unified regional modeling and model transfer.

The remainder of this paper is organized as follows. Section 2 describes the data, preprocessing, feature engineering, base models, ensemble scheme, and interpretability methods; Section 3 presents the results and discussion; and Section 4 concludes the paper and outlines future work.

2. Data and Methods

The overall workflow of the proposed framework is shown in Figure 1, comprising in sequence the five stages of data preprocessing, feature engineering, base-model training, ensemble construction, and multi-perspective interpretability analysis. The framework takes historical irradiance, meteorological, and power observations as input; after multidimensional features are constructed through feature engineering, five heterogeneous base learners are trained; their outputs are fused through the selected time-series ensemble scheme to obtain the final forecast; and three XAI methods are then jointly used to analyze the model's decision mechanism.

2.1. Data Sources and Descriptive Statistics

The experimental data were collected from two PV plants in actual operation, denoted Site 1 and Site 2, with installed capacities of 50 MW and 110 MW, respectively. The data span from 1 January 2019 to 31 December 2020 with a temporal resolution of 15 min, yielding 70,176 records per site. The original variables cover total solar irradiance (TSI), direct normal irradiance (DNI), global horizontal irradiance (GHI), ambient temperature (Temp), atmospheric pressure (Pressure), and measured PV power (Power). Neither site contained missing values in the raw data. Descriptive statistics for each variable are given in Table 1.

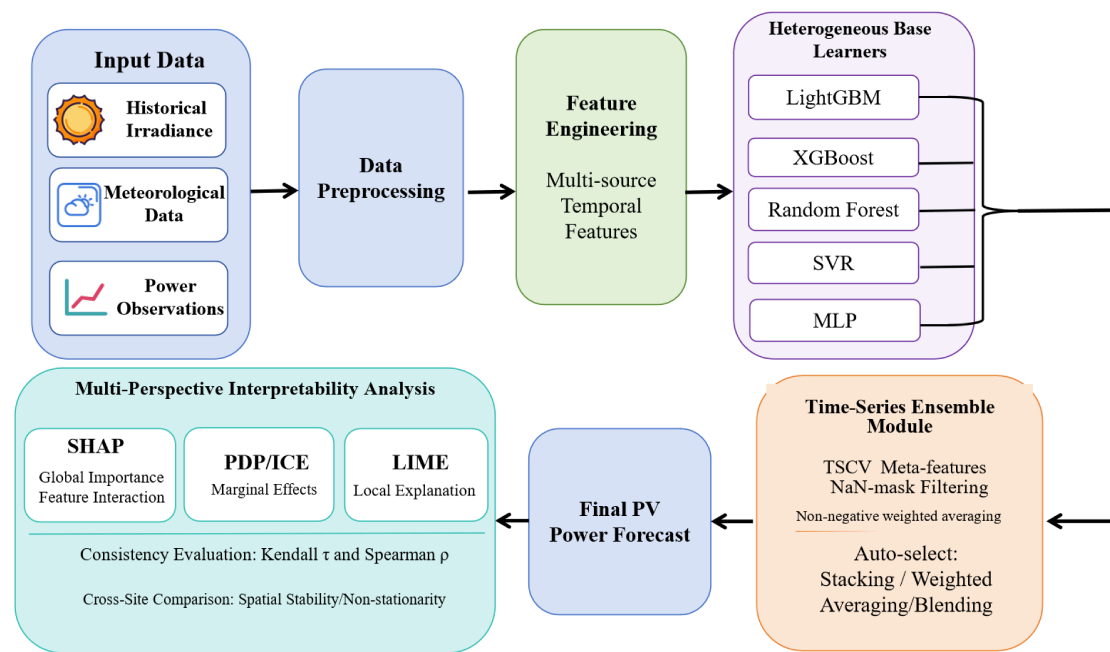


Figure 1. Overall framework of PV power forecasting and interpretability analysis.

Table 1. Descriptive statistics of the main variables for the two sites.

Site	Variable	Mean	Std.	Min.	Max.
Site 1	TSI (W/m ²)	266.21	367.89	0.00	1359.00
Site 1	DNI (W/m ²)	93.26	200.78	0.00	980.00
Site 1	GHI (W/m ²)	67.69	111.20	0.00	989.00
Site 1	Temp (°C)	13.15	14.34	−18.20	41.20
Site 1	Pressure (hPa)	913.37	8.74	894.00	936.30
Site 1	Power (MW)	9.67	13.71	0.00	48.32
Site 2	TSI (W/m ²)	169.30	248.08	0.00	1041.93
Site 2	DNI (W/m ²)	122.15	178.99	0.00	751.75
Site 2	GHI (W/m ²)	78.30	117.59	0.00	561.80
Site 2	Temp (°C)	13.70	12.04	−13.92	40.47
Site 2	Pressure (hPa)	861.04	6.15	844.51	881.67
Site 2	Power (MW)	19.57	27.94	0.00	109.36

As shown in Table 1, the two sites exhibit different observed irradiance and meteorological distributions. Site 1 has a higher mean TSI, whereas Site 2 has a higher mean DNI and a lower mean atmospheric pressure. These measured differences provide the basis for the descriptive cross-site comparison below.

2.2. Data Preprocessing

The raw records were standardized by variable name and sorted by timestamp. Negative values of physically non-negative quantities were clipped to zero, and power was normalized by installed capacity C . Because neither source file contained missing values, no interpolation, forward filling, or backward filling was applied.

$$Power_{\text{norm}}(t) = Power(t)/C, Power_{\text{norm}} \in [0, 1] \quad (1)$$

Normalization places the power outputs of plants with different capacities on a comparable scale, facilitating cross-site modeling and error comparison. After preprocessing, neither site had any missing values.

2.3. Feature Engineering

The forecasting task is a strict one-step-ahead prediction at a 15 min horizon. For a target time t , the model predicts normalized PV power at t using measurements available at or before $t - 15$ min and deterministic calendar variables for t . The resulting feature vector contains 22 dimensions in four groups:

- The first category contains deterministic calendar features for target time t : hour, month, day of year, and sine–cosine encodings of hour and month;
- The second category contains TSI, DNI, GHI, temperature, and pressure observed at $t - 15$ min, together with DNI/TSI and GHI/TSI ratios computed from those observations;
- The third category contains normalized-power lags at 1, 4, and 96 steps and TSI lags at 4 and 96 steps, all defined relative to target time t ;
- The fourth category contains the four-step mean and standard deviation of TSI and the four-step mean of power, all computed after shifting the corresponding series by one interval so that the target value is excluded.
- The periodicity encoding takes the following form:

$$\text{Hour}_{\sin}(t) = \sin\left(\frac{2\pi h_t}{24}\right), \text{Hour}_{\cos}(t) = \cos\left(\frac{2\pi h_t}{24}\right) \quad (2)$$

$$\text{Month}_{\sin}(t) = \sin\left(\frac{2\pi m_t}{12}\right), \text{Month}_{\cos}(t) = \cos\left(\frac{2\pi m_t}{12}\right) \quad (3)$$

After removal of the first 96 rows required by lag construction, 70,080 target-aligned samples remained at each site: 34,944 in 2019 and 35,136 in 2020.

2.4. Dataset Partitioning

The 34,944 target-aligned 2019 records formed the development set: the first 27,955 records were the chronological fitting block and the final 6989 records were the validation and selection block. Five-fold expanding-window TimeSeriesSplit with a one-step gap generated out-of-fold predictions within 2019. The complete 2020 period (35,136 records) did not participate in fitting, hyperparameter setting, or scheme selection, and was used only for final evaluation.

2.5. Base Learners

Five regression models with complementary inductive biases were selected as base learners to maximize ensemble diversity and to fully exploit the structural advantages of different models. The gradient-boosting tree methods (LightGBM, XGBoost) minimize a regularized objective through forward stage-wise additive modeling. Taking XGBoost as an example, its objective function at the t -th round is

$$L^{(t)} = \sum_i l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t), \Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (4)$$

where l is the loss function, f_t is the t -th tree, Ω is the regularization term, T is the number of leaves, w is the leaf weight, and γ and λ are regularization coefficients [6]. LightGBM further adopts histogram-based splitting and a leaf-wise growth strategy to improve efficiency [7]. The key hyperparameter settings of each base learner are given in Table 2. The listed base-model settings were manually specified and frozen; no grid or Bayesian search was performed, and no 2020 label was used to tune a base learner or select an ensemble scheme.

Chronos-Bolt-Small was evaluated in zero-shot mode without plant-specific fine-tuning. Each forecast used only the preceding normalized-power sequence ending at $t - 15$ min. The fixed configuration used amazon/Chronos-Bolt-Small, context length 512,

prediction length 1, and the median forecast. Its predictions were evaluated only as a standalone zero-shot baseline and were not included in the ensemble candidates.

Table 2. Fixed model and inference settings. Base-learner values were manually specified and frozen; no grid or Bayesian search was performed, and no test-set information was used.

Model	Key Hyperparameters	Scaling
LightGBM	500 trees; learning rate 0.05; max depth 8; 63 leaves; column subsampling 0.8; L1/L2 = 0.1	No
XGBoost	500 trees; learning rate 0.05; max depth 8; row/column subsampling 0.8; L1/L2 = 0.1	No
Random Forest	300 trees; max depth 15; min split 10; min leaf 5	No
SVR	RBF kernel; C = 10; $\epsilon = 0.01$; $\gamma = \text{scale}$; latest 12,000 permitted training samples	Yes
MLP	Hidden layers 128–64–32; ReLU; Adam; max_iter = 500; early_stopping = false	Yes
Chronos-Bolt-Small	amazon/chronos-bolt-small; zero-shot; no fine-tuning; context 512; horizon 1; median forecast	No additional scaler

2.6. Ensemble Construction and Selection

Let $M = 5$ denote the five supervised learning models. Based on their five prediction streams, stacking, chronological blending, and simplex-constrained weighted averaging were constructed. The best scheme was selected using only the 2019 validation block; the 2020 test set did not participate in scheme selection.

2.6.1. Stacking

Five-fold expanding-window time-series cross-validation with a one-step gap was used to generate the supervised out-of-fold meta-features. Positions not covered by a fold were stored as NaN rather than zero, allowing unavailable meta-feature rows to be distinguished from genuine nighttime zero predictions. A RidgeCV meta-learner was fitted on valid out-of-fold rows preceding the final validation block for candidate comparison and on all valid 2019 out-of-fold rows for final refitting.

$$\min_{\beta} \sum_{i \in \Omega} \left(y_i - \beta_0 - \sum_m \beta_m Z_{im} \right)^2 + \alpha \|\beta\|^2 \quad (5)$$

The regularization strength α was selected from $\{0.01, 0.1, 1, 10, 100\}$ by time-series cross-validation within the 2019 meta-training data. Five-fold splitting produced valid out-of-fold predictions for approximately 83.3% of the 2019 records. In Equation (5), β_0 is the intercept and β contains the five slope coefficients; the intercept is not penalized.

2.6.2. Simplex-Constrained Weighted Averaging

On the valid fitting rows Ω_{fit} , the weight vector w was estimated by minimizing RMSE under non-negativity and unit-sum constraints:

$$\hat{w} = \underset{w \in \mathbb{R}^M}{\operatorname{argmin}} \sqrt{\frac{1}{|\Omega_{\text{fit}}|} \sum_{i \in \Omega_{\text{fit}}} \left(y_i - \sum_{m=1}^M w_m \hat{y}_{im} \right)^2}, w_m \geq 0, \sum_{m=1}^M w_m = 1 \quad (6)$$

In Equation (6), Ω_{fit} denotes the valid out-of-fold fitting rows and w contains the five model weights. SLSQP was used. Candidate weights were fitted before validation and, if selected, refitted on all valid 2019 out-of-fold rows; 2020 labels were never used.

2.6.3. Blending

Chronological blending used separate 2019 blocks for base-model fitting, combiner fitting, and final candidate comparison. RidgeCV was fitted to the five prediction streams without using any 2020 observations for parameter estimation or selection.

The three ensemble candidates were compared on the same final 20% of 2019, and the ensemble with the lowest validation RMSE was selected for each site. Separately, the best of the five supervised models on that block was selected as the statistical comparator. The 2020 test period was evaluated only after both choices were finalized.

2.7. Evaluation Metrics and Baselines

The root-mean-square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) are used to evaluate model accuracy; all three are computed on the normalized-power scale:

$$\text{RMSE} = \sqrt{\left(\frac{1}{n}\right) \sum_i (y_i - \hat{y}_i)^2} \quad (7)$$

$$\text{MAE} = \frac{1}{n} \sum_i |y_i - \hat{y}_i| \quad (8)$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (9)$$

where y_i and $\hat{y}_i$ are the normalized measured and predicted power of the i -th sample, $\bar{y}$ is the measured mean, and n is the number of samples.

The comparison also included 15 min persistence, 24 h seasonal persistence, irradiance-corrected persistence, empirical smart persistence, seasonal autoregressive Ridge, and a one-layer GRU. These baselines cover simple persistence, statistical autoregression, and recurrent neural forecasting. Chronos-Bolt-Small was reported separately as a zero-shot time-series foundation-model baseline. The empirical smart-persistence profile used the 95th percentiles for each month and 15 min time slot in 2019; the seasonal Ridge used lagged power and cyclic calendar terms, and the GRU used the preceding 16 multivariate steps with 32 hidden units and chronological 2019 epoch selection.

The validation-selected ensemble was compared with the validation-selected single model by a paired day-cluster bootstrap with 5000 resamples, a Diebold–Mariano test on squared-error loss with heteroskedasticity- and autocorrelation-consistent variance, and a paired Wilcoxon test on daily RMSE. Holm adjustment was applied across the two sites, and 95% bootstrap confidence intervals (CIs) are reported.

2.8. Multi-Perspective Interpretability Analysis Methods

2.8.1. SHAP

SHAP, based on the Shapley value from cooperative game theory, decomposes an individual prediction into the sum of feature contributions. For feature j in the feature set F , its Shapley value is defined as the weighted average of the feature's marginal contributions over all possible subsets S :

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \left[\frac{|S|!(|F|-|S|-1)!}{|F|!} \right] \cdot (f(S \cup \{j\}) - f(S)) \quad (10)$$

The SHAP explanation model satisfies additivity, local accuracy, missingness, and consistency. Because the ensemble contains heterogeneous base learners, model-agnostic permutation SHAP was used. For each site, 12 background samples from 2019 and 100 evenly spaced samples from the 2020 test period were used, with five permutations per sample. The resulting values were used for global importance, beeswarm, and dependence plots.

Pairwise interaction strength among the eight highest-ranked features was estimated using Friedman's H statistic based on two-dimensional partial dependence. SHAP, PDP/ICE, LIME, and permutation importance were all calculated for the same ensemble model.

2.8.2. PDP and ICE

The partial dependence plot (PDP) characterizes the average effect of a target feature on the prediction by marginalizing over the remaining features [23,24]:

$$PD_S(x_S) = E_{x_C}[f(x_S, x_C)] \approx \frac{1}{n} \sum_i f(x_S, x_{C,i}) \quad (11)$$

The individual conditional expectation (ICE) shows the marginal response curve of an individual sample, which can reveal the heterogeneity masked by averaging in the PDP [24]. In this paper, the range of the PDP curve is used as the PDP-importance measure of a feature.

2.8.3. LIME

LIME generates perturbed samples in the neighborhood of the instance x to be explained and fits an interpretable local linear model g , giving the local contribution of each feature by solving the following optimization problem [25]:

$$\xi(x) = \operatorname{argmin}_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (12)$$

Here, π_x is the neighborhood weight, L is the local-fidelity loss, and $\Omega(g)$ is the complexity penalty. LIME was applied to the same ensemble model used for SHAP and PDP/ICE. Three cases were selected deterministically: a median case from 11:00 to 13:59 above the site-specific 90th percentile of last-observed TSI; a median case from 14:00 to 16:59 within the 40–60th TSI percentiles; and a median case from 06:00 to 08:59 with last-observed TSI above 10 W/m². Weighted local R^2 and the absolute gap between the ensemble-model prediction and the local surrogate are reported.

2.8.4. XAI Consistency Measures

To compare feature rankings from different XAI methods, Kendall's τ -b and Spearman's ρ were calculated for each pair of rankings. Kendall's τ -b, which accounts for ties, is defined as

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_c + n_d + t_x)(n_c + n_d + t_y)}} \quad (13)$$

where n_c and n_d are the numbers of concordant and discordant pairs, and t_x and t_y are the numbers of pairs tied only in the first and second rankings, respectively. Values closer to 1 indicate stronger agreement.

3. Results and Discussion

3.1. Exploratory Data Analysis

Figure 2 shows the measured PV power and TSI time series for both sites. Both variables exhibit clear diurnal and seasonal patterns, with higher peaks in summer than in winter. Figure 3 shows the correlations among the measured variables. Power was strongly correlated with TSI ($r = 0.95$ at Site 1 and $r = 0.91$ at Site 2) and positively correlated with GHI, whereas its correlations with temperature and pressure were weaker. Figure 4 compares the distributions of key variables. After capacity normalization, the mean power was approximately 0.193 at Site 1 and 0.178 at Site 2, as calculated from the values in Table 1.

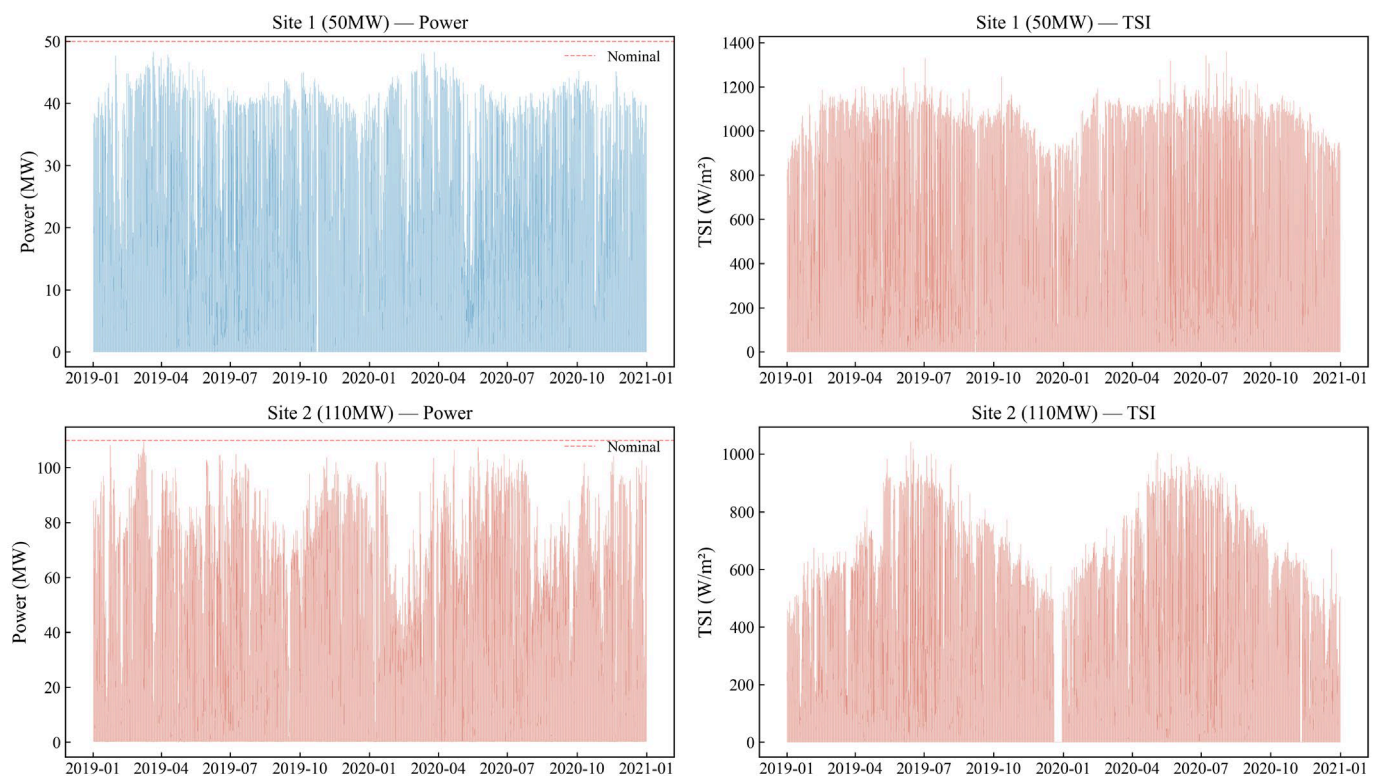


Figure 2. Time series of measured PV power and total solar irradiance for the two sites; dashed lines in the power panels indicate installed capacity. Colors are used only for visual differentiation and have no quantitative meaning.

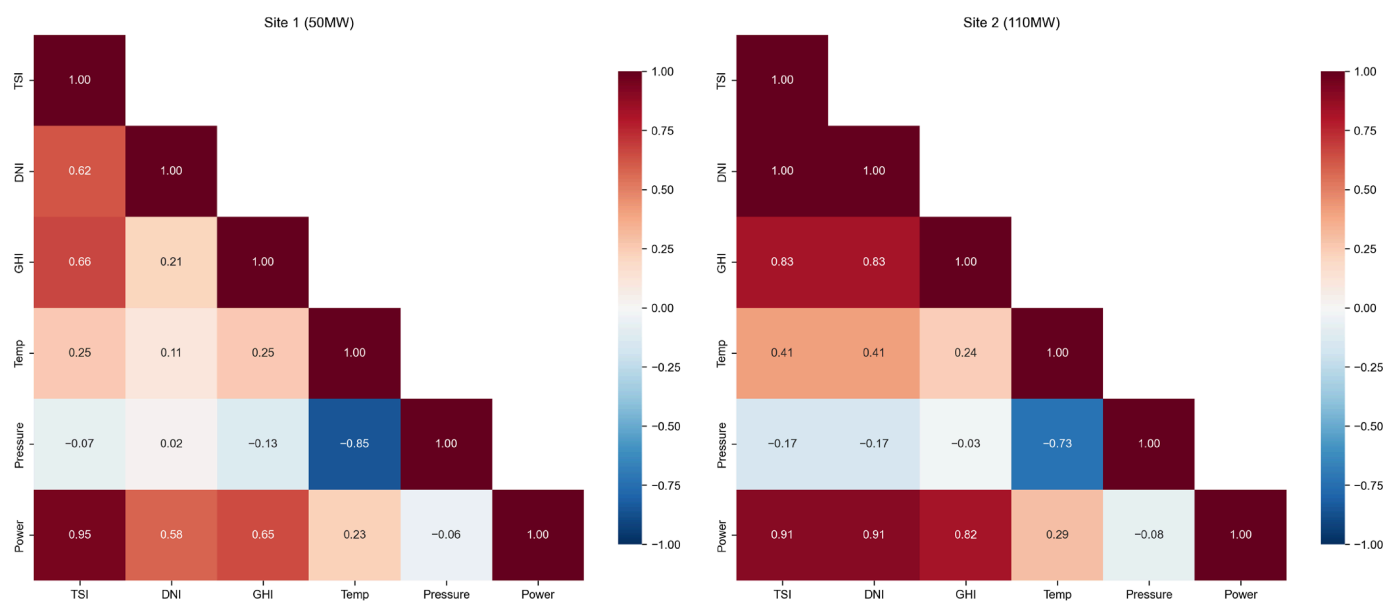


Figure 3. Feature correlation heatmap.

3.2. Comparison of Forecasting Accuracy

The 2020 test performance of the models selected using the 2019 validation data is shown in Table 3. The validation procedure selected simplex-constrained weighted averaging for Site 1 and stacking for Site 2. Both selected ensembles achieved the lowest RMSE and highest R^2 at their respective sites: the RMSE/ R^2 values were 0.03497/0.9835 at Site 1 and 0.03758/0.9772 at Site 2. Persistence, statistical, neural, and zero-shot time-series foundation-model baselines are also included for comparison.

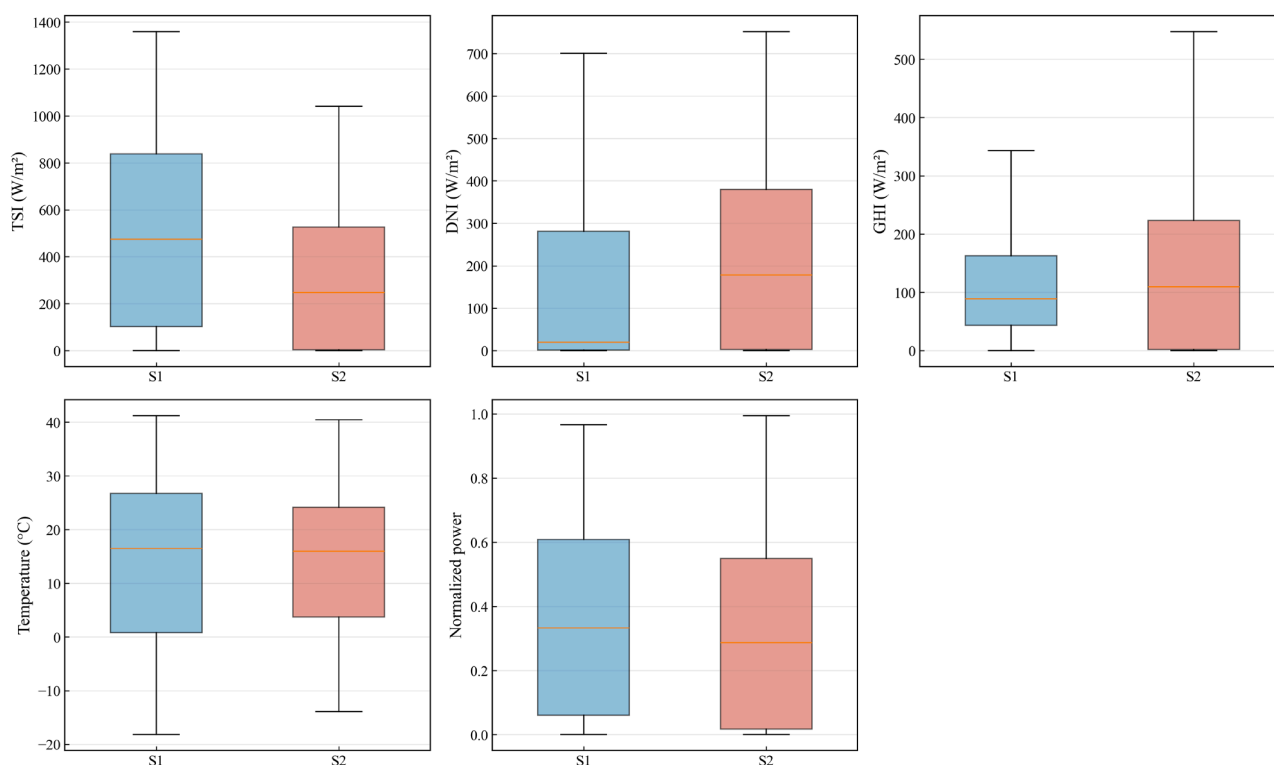


Figure 4. Boxplot comparison of the distribution of key variables for the two sites; orange lines denote medians, and blue and salmon boxes represent Sites 1 and 2, respectively.

Table 3. Forecasting performance on the 2020 test period after selection in 2019.

Model	S1 MAE	S1 RMSE	S1 R ²	S2 MAE	S2 RMSE	S2 R ²
Persistence	0.0213	0.04332	0.9747	0.0202	0.04312	0.9700
Seasonal persistence	0.0605	0.13842	0.7420	0.0539	0.12075	0.7646
Irradiance-corrected persistence	0.0216	0.06342	0.9459	0.0265	0.06687	0.9278
Empirical smart persistence	0.0136	0.03635	0.9822	0.0157	0.03831	0.9763
Seasonal autoregressive Ridge	0.0188	0.03846	0.9801	0.0190	0.03881	0.9757
GRU	0.0475	0.07307	0.9281	0.0390	0.06765	0.9261
LightGBM	0.0149	0.03622	0.9823	0.0170	0.03913	0.9753
XGBoost	0.0152	0.03641	0.9821	0.0174	0.04003	0.9741
Random Forest	0.0137	0.03511	0.9834	0.0158	0.03763	0.9771
SVR	0.1387	0.19498	0.4881	0.1468	0.21268	0.2696
MLP	0.0187	0.03857	0.9800	0.0206	0.04076	0.9732
Chronos-Bolt-Small	0.0167	0.03814	0.9804	0.0179	0.03888	0.9756
Five-model ensemble	0.0138	0.03497	0.9835	0.0167	0.03758	0.9772

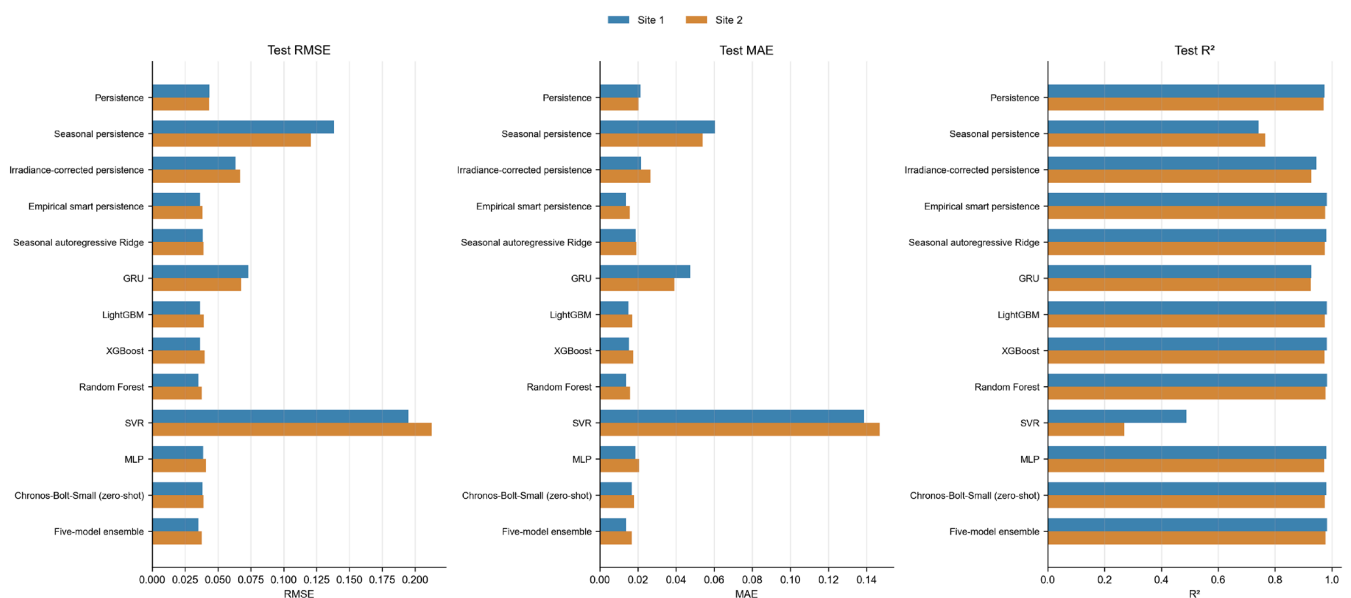
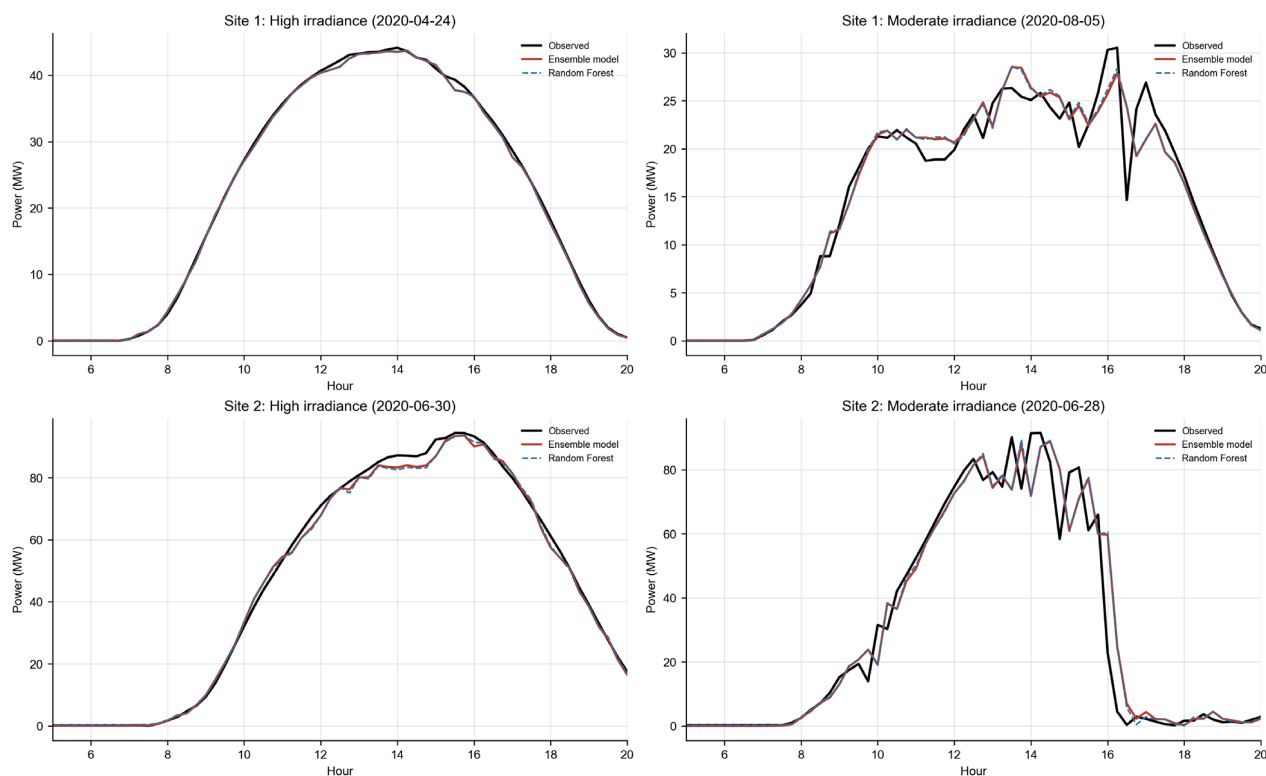
Relative to the validation-selected single supervised models (Random Forest at both sites), the RMSE reductions were 0.41% and 0.14%, respectively, indicating modest gains. The selected ensembles had MAE values of 0.0138 and 0.0167, slightly higher than those of empirical smart persistence (0.0136 and 0.0157). Thus, the selected ensembles were superior in RMSE and R², but not in MAE.

The coefficients of the ensemble models are shown in Table 4. The selected schemes were simplex-constrained weighted averaging at Site 1 and stacking at Site 2. Only simplex-constrained weighted averaging requires non-negative coefficients that sum to one.

The RMSE, MAE, and R² values for the forecasting baselines, standalone models, and five-model ensembles are compared in Figure 5. Measured and predicted power for the deterministically selected high- and moderate-irradiance representative days are shown in Figure 6. These examples were not used for model selection.

Table 4. Coefficients of the ensemble models.

Model/Term	Site 1 Coefficient	Site 2 Coefficient	Selected Schemes
LightGBM	0.149979	0.225612	S1: simplex-constrained weighted averaging; S2: stacking
XGBoost	0.000000	0.036253	S1: simplex-constrained weighted averaging; S2: stacking
Random Forest	0.827024	0.728456	S1: simplex-constrained weighted averaging; S2: stacking
SVR	0.000000	−0.001228	S1: simplex-constrained weighted averaging; S2: stacking
MLP	0.022997	0.015224	S1: simplex-constrained weighted averaging; S2: stacking
Intercept	0.000000	−0.001738	S1: simplex-constrained weighted averaging; S2: stacking

**Figure 5.** Comparison of the forecasting performance of each model.**Figure 6.** Comparison of predicted and measured power on deterministically selected high- and moderate-irradiance representative days.

3.3. Global Interpretability: SHAP Analysis

The SHAP global importance and beeswarm results for the ensemble model are shown in Figures 7 and 8, respectively. The three highest-ranked features were Power_lag1, Hour_cos, and Hour_sin at Site 1, and Power_lag1, Power_roll4_mean, and Hour at Site 2. The dominance of Power_lag1 indicates persistence-enhanced very-short-term forecasting rather than only a general irradiance-to-power relationship.

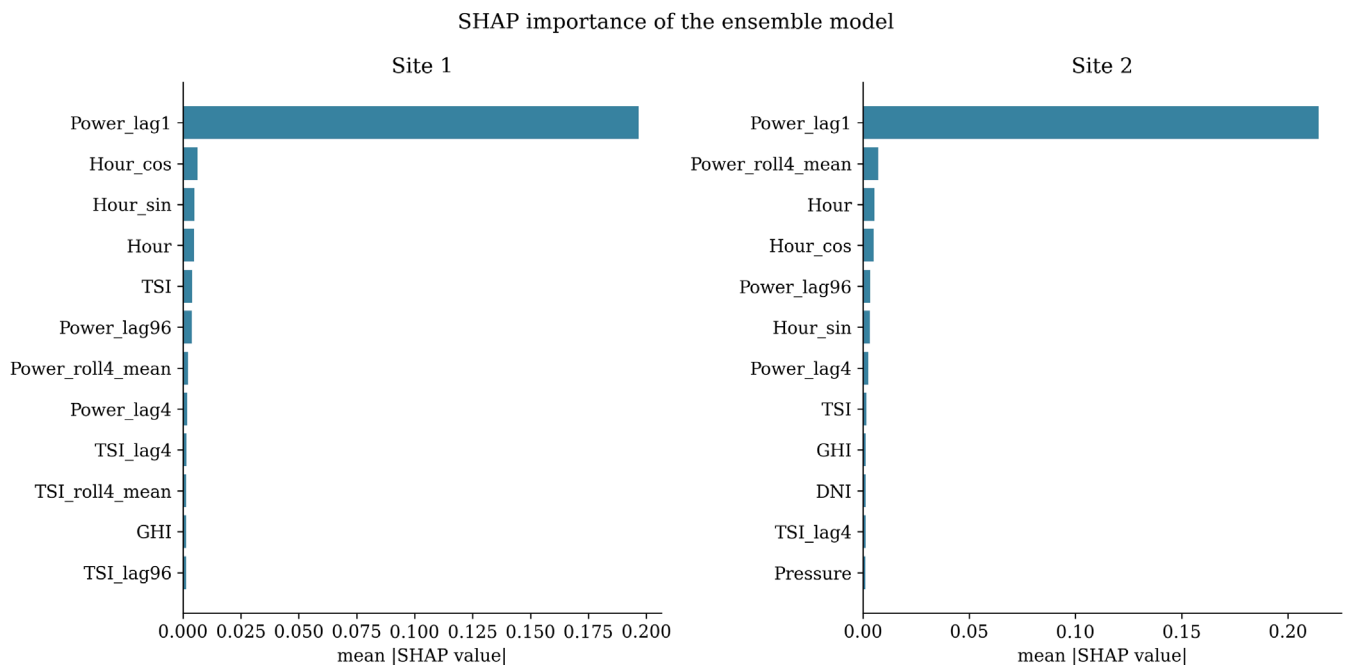


Figure 7. SHAP global feature importance.

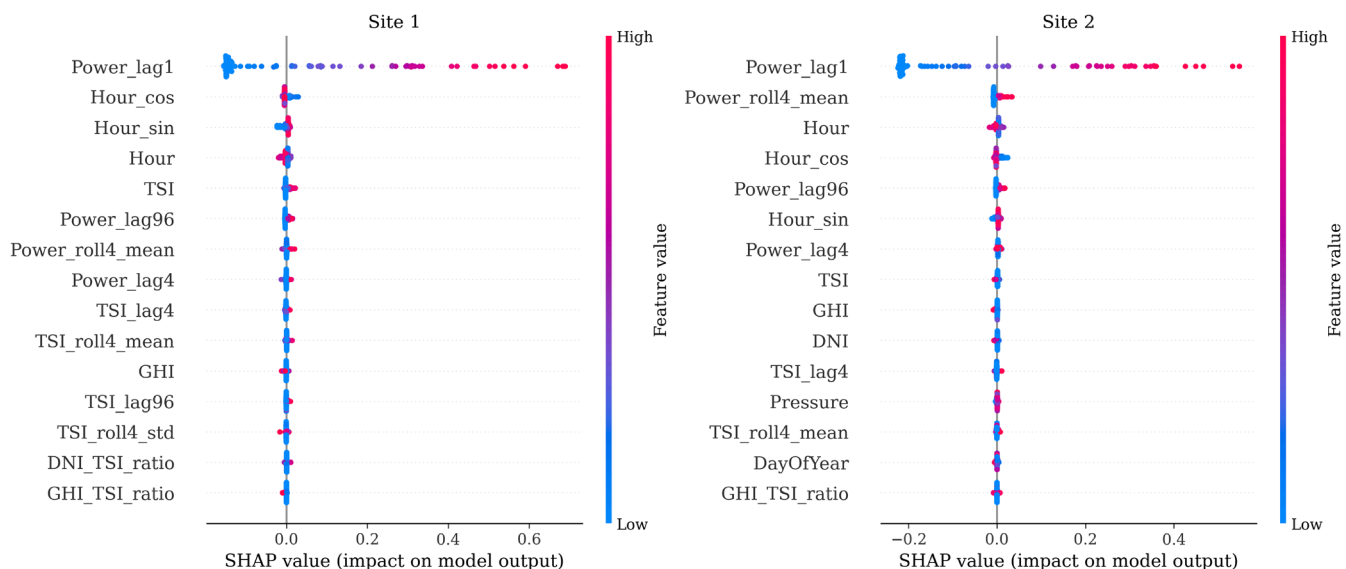


Figure 8. SHAP beeswarm plot.

Power_lag1 ranked first at both sites. Time-of-day and power-history features occupied most leading ranks, while irradiance features provided smaller, site-dependent contributions. The comparison is based on the observed feature distributions at the two PV plants.

The SHAP dependence and Friedman H interaction results for the ensemble model are shown in Figures 9 and 10, respectively. They describe fitted associations be-

tween prior-interval observations and the forecast and are not interpreted as causal physical mechanisms.

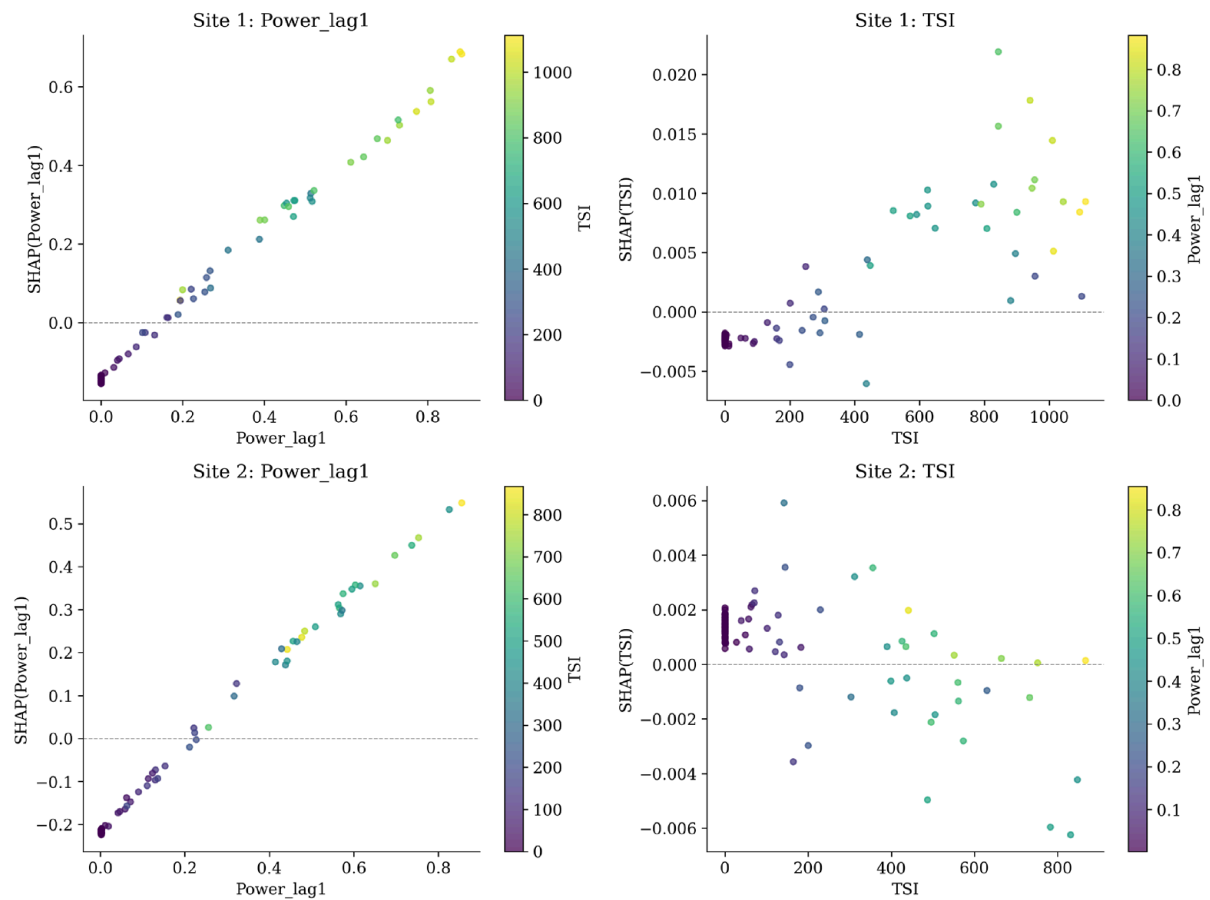


Figure 9. SHAP dependence plots of key features.

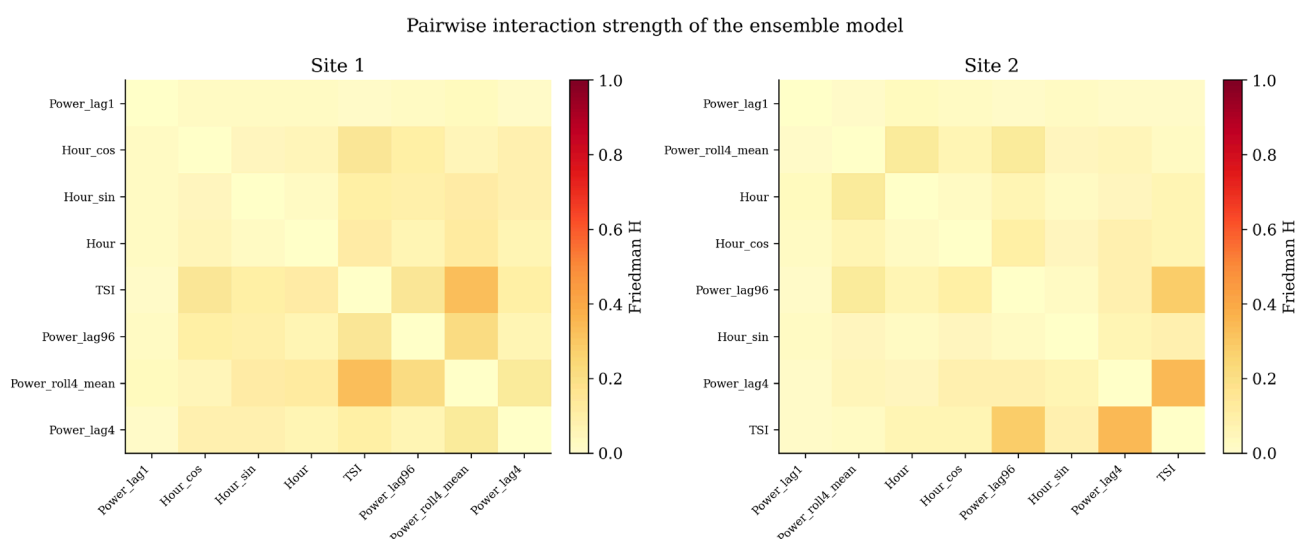


Figure 10. Feature-interaction heatmap of the ensemble model based on Friedman's H-statistic.

3.4. Marginal Effects: PDP and ICE Analysis

The PDP and ICE results for the ensemble model are shown in Figures 11 and 12, respectively. The curves show the average response and sample-level heterogeneity of the fitted prediction function, complementing the SHAP analysis.

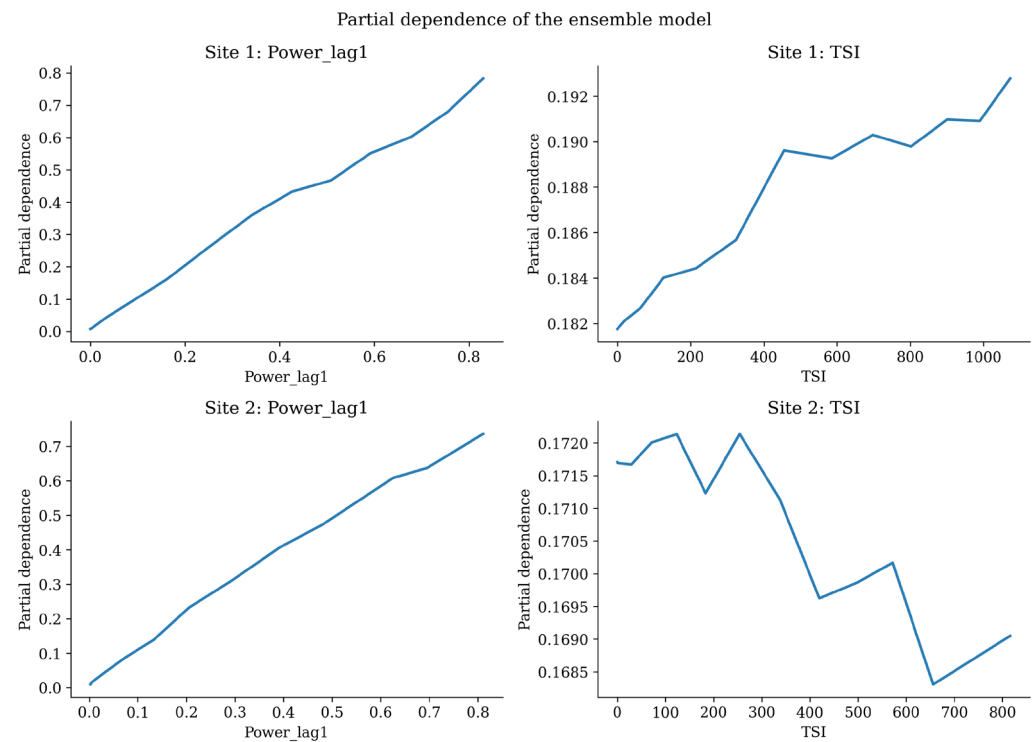


Figure 11. Partial dependence plots of the ensemble model for Power_lag1 and last-observed TSI.

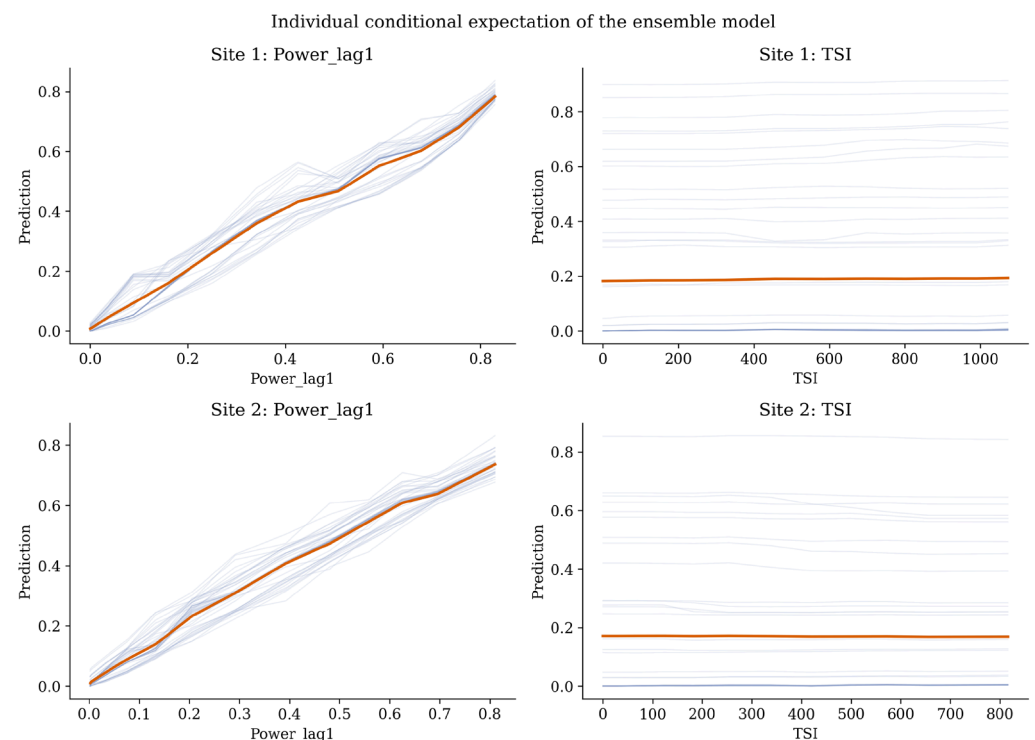


Figure 12. ICE curves of the ensemble model for Power_lag1 and last-observed TSI; thin light-blue curves show individual conditional responses, and orange curves show their mean response.

3.5. Local Interpretability: LIME Analysis

The LIME explanations of the ensemble model under the three deterministically defined conditions are shown in Figure 13. These conditions are high-irradiance noon, moderate-irradiance afternoon, and sunrise transition. Each panel includes the local weighted R^2 .

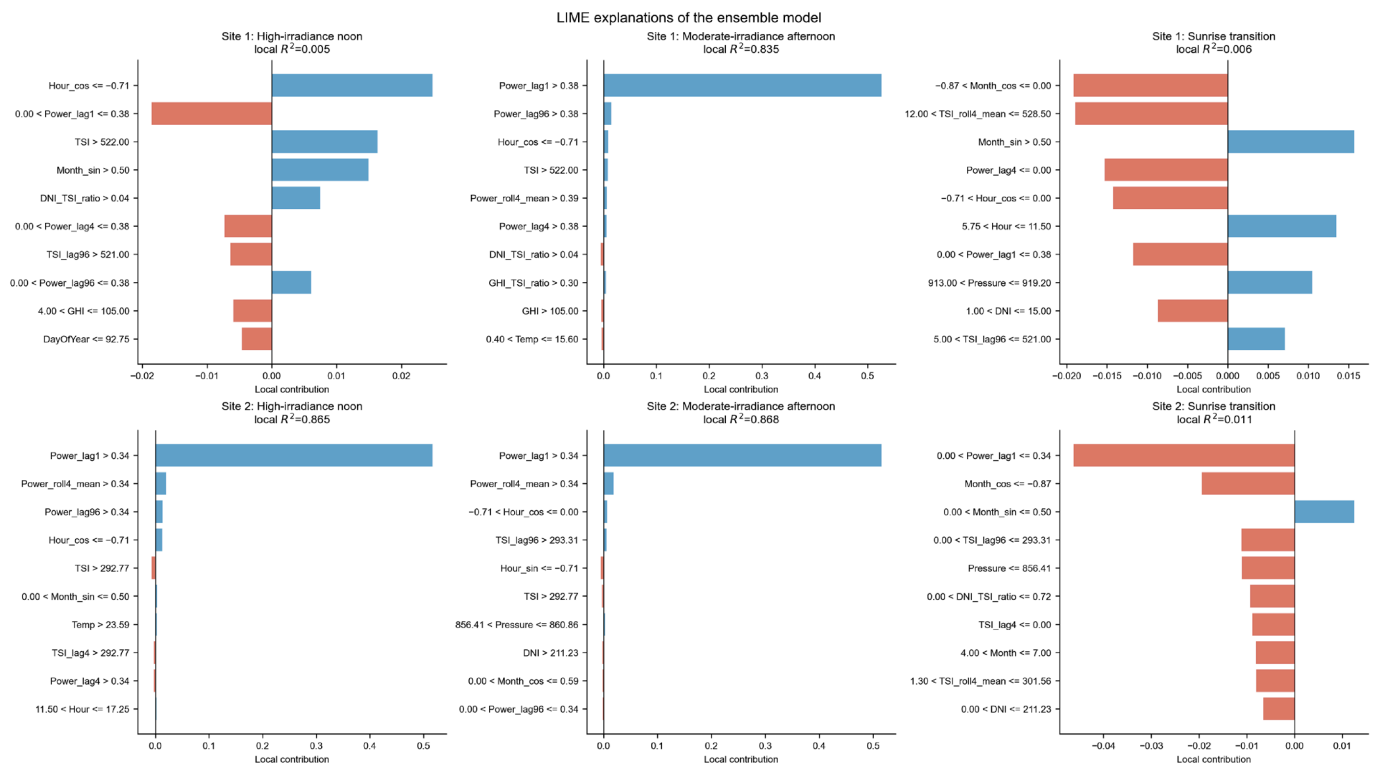


Figure 13. LIME explanations of the ensemble model under three deterministically selected operating conditions; local weighted R^2 is reported in each panel. Blue and red bars indicate positive and negative local contributions, respectively.

For the high-irradiance noon cases, the local weighted R^2 values were 0.005 at Site 1 and 0.865 at Site 2; the absolute local surrogate gaps were 0.0542 and 0.1094, respectively. The Site 1 surrogate had very low fidelity and was not used for detailed local interpretation.

For the moderate-irradiance afternoon cases, the local weighted R^2 values were 0.835 at Site 1 and 0.868 at Site 2; the absolute local surrogate gaps were 0.0685 and 0.1618, respectively. These cases support qualitative local interpretation subject to the reported gaps.

For the sunrise-transition cases, the local weighted R^2 values were 0.006 at Site 1 and 0.011 at Site 2; the absolute local surrogate gaps were 0.1141 and 0.0779, respectively. Both surrogates had very low fidelity, so their feature contributions were not interpreted mechanistically.

The LIME cases complement the global SHAP and PDP/ICE results, but local feature contributions are interpreted only when the reported fidelity is adequate. This avoids treating a weak local linear surrogate as an exact description of the ensemble model.

3.6. Consistency Analysis of XAI Methods

The comparison of the ensemble-model feature rankings is shown in Table 5. At Site 1, SHAP agreed most strongly with permutation importance, whereas at Site 2, SHAP agreed most strongly with PDP.

Permutation importance, SHAP attribution, and PDP range measure different aspects of the same ensemble model. Their positive and statistically significant ranking associations provide a consistency check, while no ranking is treated as a ground-truth physical attribution.

The comparison of the eight highest-ranked displayed features at each site across the three methods is shown in Figure 14. Overall, SHAP, PDP, and permutation importance produced broadly consistent rankings. Power_lag1 ranked first at both sites under all three methods, indicating that short-term power persistence was the dominant driver of

the forecasts. The highest-ranked features were more stable at Site 2, whereas the rankings at Site 1 varied beyond Power_lag1, particularly for Power_lag96 and TSI. Thus, the three methods agreed on the core predictors but were more sensitive to method choice for features with weaker contributions or stronger correlations.

Table 5. Consistency of feature rankings for the ensemble model among XAI methods.

Site	Method Pair	Kendall τ -b	Kendall ρ	Spearman ρ	Spearman p
Site 1	PDP vs. Permutation	0.680	$p < 0.001$	0.871	$p < 0.001$
Site 1	SHAP vs. Permutation	0.758	$p < 0.001$	0.919	$p < 0.001$
Site 1	SHAP vs. PDP	0.732	$p < 0.001$	0.885	$p < 0.001$
Site 2	PDP vs. Permutation	0.576	$p < 0.001$	0.732	$p < 0.001$
Site 2	SHAP vs. Permutation	0.610	$p < 0.001$	0.747	$p < 0.001$
Site 2	SHAP vs. PDP	0.688	$p < 0.001$	0.831	$p < 0.001$

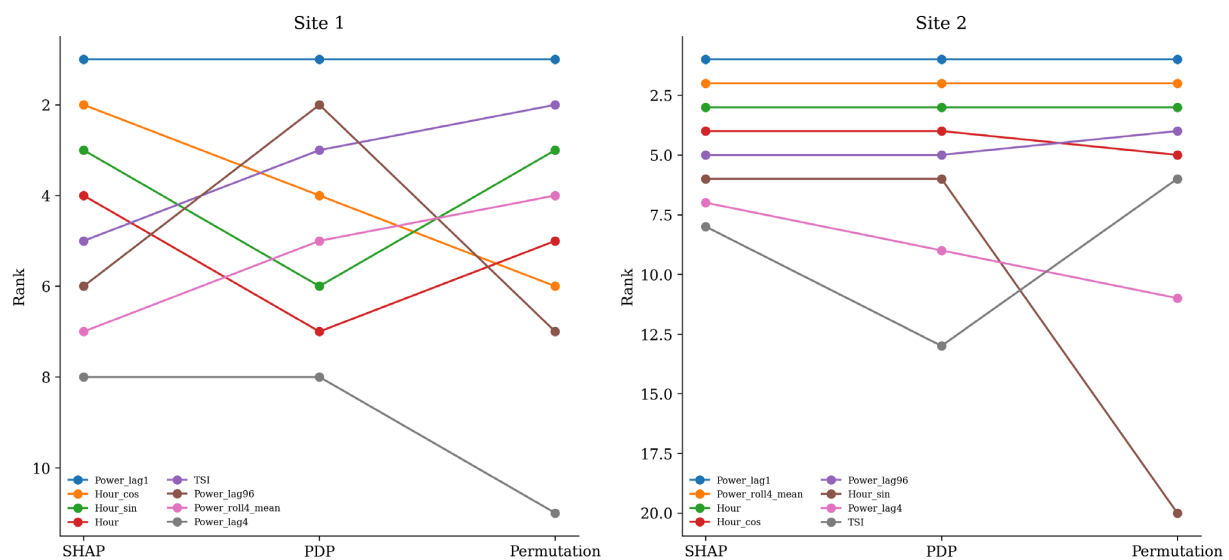


Figure 14. Comparison of the eight highest-ranked displayed features among the three methods (bump chart).

3.7. Cross-Site Interpretability Comparison

The cross-site comparison of normalized ensemble-model SHAP importance is shown in Figure 15, and the ranking differences are summarized in Table 6. Recent power-history features were important at both sites, while several irradiance features changed rank.

Table 6. Cross-site differences in ensemble-model SHAP rankings (excerpt in descending order of rank difference).

Feature	Site 1 SHAP	Site 1 Rank	Site 2 Rank	Rank Diff.
DNI	0.00027	21	10	11
Power_rol4_mean	0.00216	7	2	5
TSI_lag96	0.00138	12	17	5
Pressure	0.00047	17	12	5
DNI_TSI_ratio	0.00081	14	18	4
DayOfYear	0.00043	18	14	4
Hour_sin	0.00501	3	6	3
TSI	0.00402	5	8	3
TSI_rol4_mean	0.00142	10	13	3
TSI_rol4_std	0.00118	13	16	3
Temp	0.00050	16	19	3

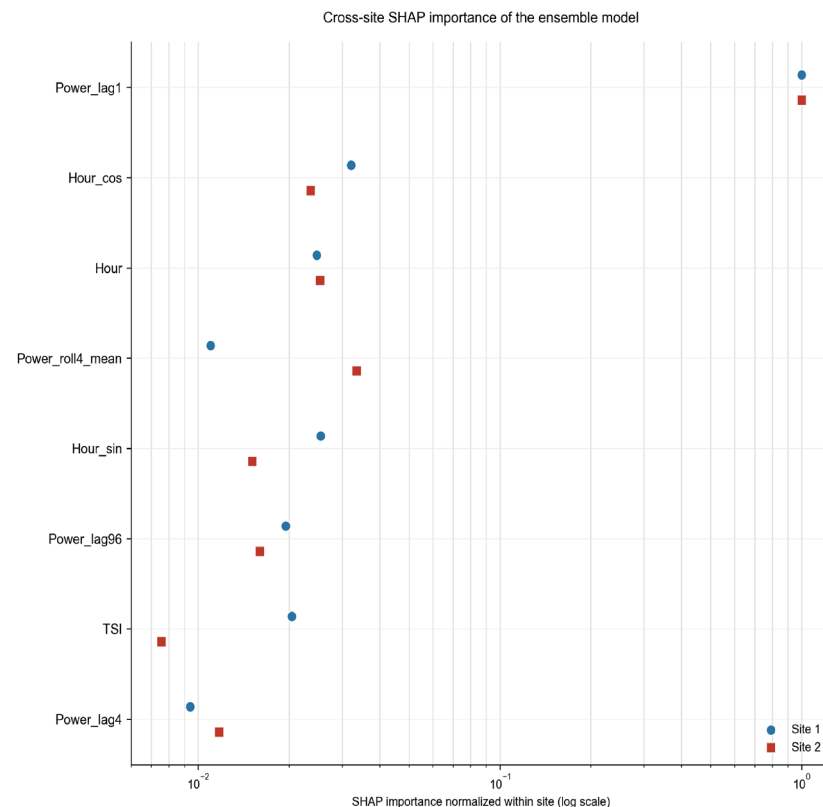


Figure 15. Cross-site comparison of normalized ensemble-model SHAP importance for the two sites; the logarithmic axis and vertically offset markers improve the visibility of secondary features.

Power_lag1 ranked first at both sites, whereas the supporting power-history, time, and irradiance features showed larger rank differences. This result supports site-aware calibration within the two evaluated PV plants.

3.8. Error Analysis

The time-of-day and monthly error distributions of the ensemble model are shown in Figure 16. Errors were concentrated mainly during daylight hours, when PV output was non-zero and more variable. This diagnostic was not used for model selection.

3.9. Comprehensive Discussion

The ensemble comparison used schemes and single-model comparators selected on the 2019 validation block. Relative to Random Forest, the Site 1 RMSE reduction was 0.41% (95% day-cluster bootstrap CI [0.24, 0.59]%). Relative to Random Forest, the Site 2 reduction was 0.14% (95% CI [−0.07, 0.35]%). The Holm-adjusted Diebold–Mariano and Wilcoxon p -values were 2.06×10^{-5} and 2.21×10^{-4} at Site 1, and 0.215 and 0.380 at Site 2, respectively. The Site 1 interval excluded zero, whereas the Site 2 interval included zero; statistical support was therefore obtained only at Site 1. These forecasts can support short-term dispatch and electricity-market operations.

SHAP, PDP/ICE, LIME, and permutation importance were all applied to the ensemble model. The global rankings agreed that recent power history was important. LIME was interpreted separately and only in proportion to the reported local fidelity of each case.

The two sites show similar importance for recent power history and larger differences for some irradiance features. This is evidence of site dependence within the present dataset and may inform site-aware model transfer and calibration.

The study is limited to two PV plants, a single 15 min horizon, and observation-based inputs without NWP forecasts. Model-agnostic SHAP and Friedman H were estimated

on fixed subsamples, and LIME was interpreted only when local fidelity was adequate. Chronos-Bolt-Small was evaluated as a fixed zero-shot standalone baseline without plant-specific fine-tuning.

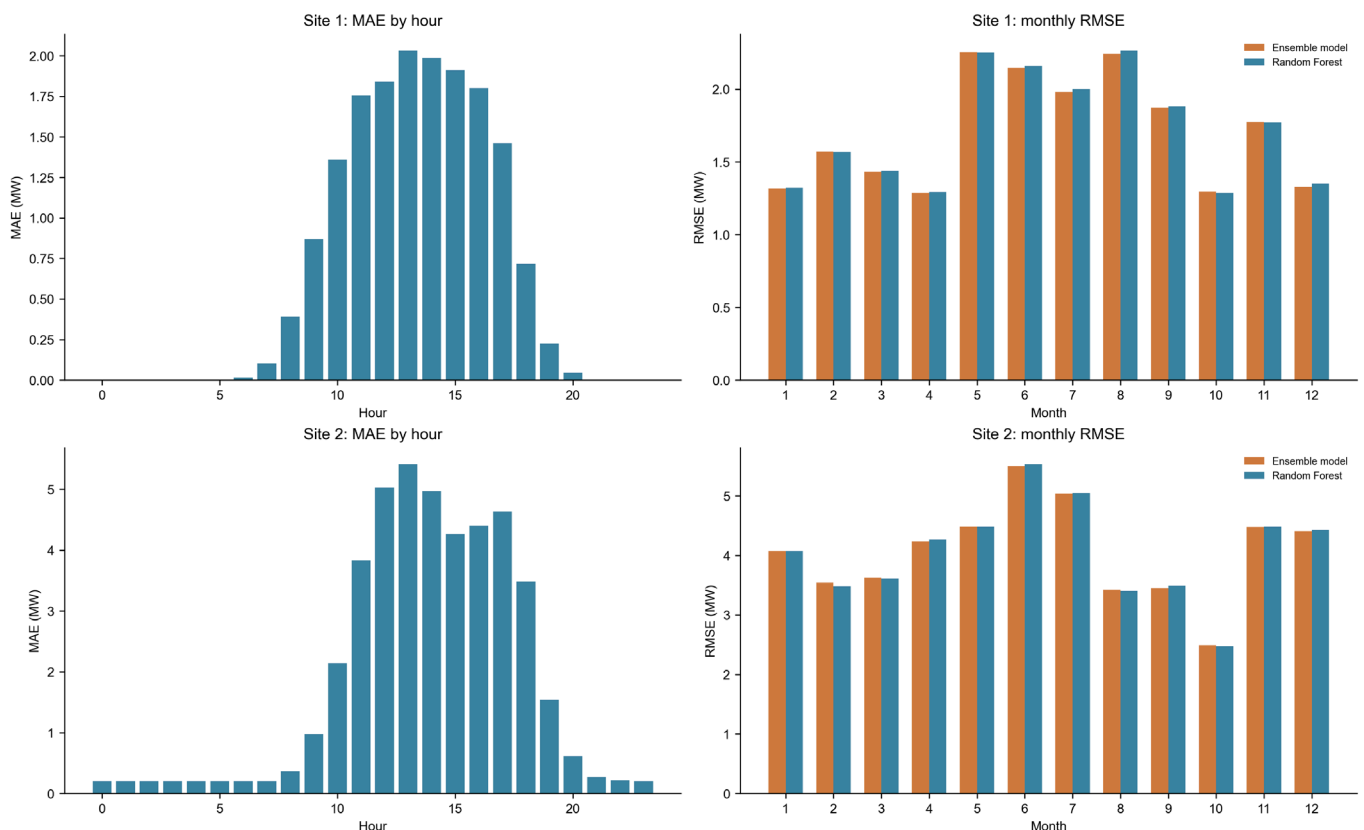


Figure 16. Hourly MAE and monthly RMSE of the ensemble model in MW.

4. Conclusions

This paper proposes and validates a PV power forecasting framework that integrates a multi-model weighted ensemble with multi-perspective interpretability analysis. The main conclusions are as follows:

- (1) The selected ensemble achieved the lowest RMSE among the evaluated methods at both sites. The improvement over the validation-selected Random Forest was modest and statistically supported only at Site 1.
- (2) The multi-perspective interpretation results show that the XAI methods generally agree on the core features but exhibit some method sensitivity for weakly contributing or highly correlated features.
- (3) Across the two sites, historical-power features showed more stable importance rankings, whereas irradiance features showed larger site-dependent differences, indicating that site-aware calibration may be needed for model transfer.

Future work will extend the analysis to additional plants and forecasting horizons, incorporate NWP inputs, evaluate weather-aware foundation models, and assess the stability and transferability of ensemble-model explanations.

Author Contributions: Conceptualization, S.W.; Methodology, S.W.; Software, S.W.; Validation, S.W., D.L. and Y.Z.; Formal analysis, S.W.; Writing—original draft preparation, S.W.; Writing—review and editing, D.L. and Y.Z.; Supervision, D.L. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by: 1. Fujian Provincial Science and Technology Special Commissioner Project for 2025–2026 (202535090182). 2. Research on the Spatial-Temporal Pattern of Fujian’s Carbon Footprint Based on Graph Spectrum and Its Impact Mechanism and Planning Strategy: Plan Document No.: MEPKFC [2024] No. 16, Project No.: 2025R016, Plan Name: Fujian Provincial Environmental Protection Technology Plan Project in 2025.

Data Availability Statement: The data supporting the conclusions of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: Author D.L. was employed by the company Fujian Jofee Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Ahmed, R.; Sreeram, V.; Mishra, Y.; Arif, M.D. A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. *Renew. Sustain. Energy Rev.* **2020**, *124*, 109792. [\[CrossRef\]](#)
2. Iheanetu, K.J. Solar Photovoltaic Power Forecasting: A Review. *Sustainability* **2022**, *14*, 17005. [\[CrossRef\]](#)
3. Di Leo, P.; Ciocia, A.; Malgaroli, G.; Spertino, F. Advancements and Challenges in Photovoltaic Power Forecasting: A Comprehensive Review. *Energies* **2025**, *18*, 2108. [\[CrossRef\]](#)
4. Mellit, A.; Massi Pavan, A.; Ogliari, E.; Leva, S.; Lughi, V. Advanced Methods for Photovoltaic Output Power Forecasting: A Review. *Appl. Sci.* **2020**, *10*, 487. [\[CrossRef\]](#)
5. Husein, M.; Gago, E.J.; Hasan, B.; Pegalajar, M. Towards energy efficiency: A comprehensive review of deep learning-based photovoltaic power forecasting strategies. *Heliyon* **2024**, *10*, e33419. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Gaboitaolelwe, J.; Zungeru, A.M.; Yahya, A.; Lebekwe, C.K.; Vinod, D.N.; Salau, A.O. Machine learning based solar photovoltaic power forecasting: A review and comparison. *IEEE Access* **2023**, *11*, 40820–40845. [\[CrossRef\]](#)
7. Zhu, C.; Tu, Y.; Wei, Q.; Zang, Y.; Zhou, P.; Yang, L.; Wang, J.; Yu, Y.; Lv, R.; Du, J.; et al. Short-term power prediction of photovoltaic power station based on LSTM-XGBoost model. *Sol. Energy* **2025**, *300*, 113819. [\[CrossRef\]](#)
8. Zheng, W.; Xiao, H.; Pei, W. Distributed-regional photovoltaic power generation prediction with limited data: A robust autoregressive transfer learning method. *Appl. Energy* **2025**, *380*, 125058. [\[CrossRef\]](#)
9. Scott, C.; Ahsan, M.; Albarbar, A. Machine learning for forecasting a photovoltaic (PV) generation system. *Energy* **2023**, *278*, 127807. [\[CrossRef\]](#)
10. Hu, Z.; Gao, Y.; Ji, S.; Mae, M.; Imaizumi, T. Improved multistep ahead photovoltaic power prediction model based on LSTM and self-attention with weather forecast data. *Appl. Energy* **2024**, *359*, 122709. [\[CrossRef\]](#)
11. Pandžić, F.; Capuder, T. Advances in Short-Term Solar Forecasting: A Review and Benchmark of Machine Learning Methods and Relevant Data Sources. *Energies* **2024**, *17*, 97.
12. Mienye, I.D.; Sun, Y. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access* **2022**, *10*, 99129–99149. [\[CrossRef\]](#)
13. Abdellatif, A.; Mubarak, H.; Ahmad, S.; Ahmed, T.; Shafiullah, G.M.; Hammoudeh, A.; Abdellatef, H.; Rahman, M.M.; Ghenni, H.M. Forecasting photovoltaic power generation with a stacking ensemble model. *Sustainability* **2022**, *14*, 11083. [\[CrossRef\]](#)
14. Lateko, A.A.H.; Yang, H.T.; Huang, C.M.; Aprillia, H.; Hsu, C.-Y.; Zhong, J.-L.; Phuong, N.H. Stacking ensemble method with the RNN meta-learner for short-term PV power forecasting. *Energies* **2021**, *14*, 4733. [\[CrossRef\]](#)
15. Nguyen, T.A.; Pham, M.H.; Vu, M.P.; Nguyen, N.T.; Nguyen, D.T.; Vu, T.A.T.; Tran, T.T.; Do, A.T. A study on novel solar power forecasting using an XGB-LiGBM-RF hybrid model and the L-BFGS-B optimization algorithm. *Eng. Technol. Appl. Sci. Res.* **2025**, *15*, 24516–24522. [\[CrossRef\]](#)
16. Islam, I.; Sristy, S.S.; Jahan, R.; Tareq, M.R.; Hasan, M.; Uddin, P. Advancing solar radiation prediction with explainable AI and ensemble learning techniques. *Telemat. Inform. Rep.* **2026**, *21*, 100292. [\[CrossRef\]](#)
17. Öter, A.; Ersöz, B. Artificial intelligence assisted solar energy forecasting by explainability approaches with LIME and SHAP. *El-Cezeri* **2025**, *12*, 205–212. [\[CrossRef\]](#)
18. Nguyen, H.N.; Tran, Q.T.; Ngo, C.T.; Nguyen, D.D.; Tran, V.Q. Solar energy prediction through machine learning models: A comparative analysis of regressor algorithms. *PLoS ONE* **2025**, *20*, e0315955. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Abdelghafar, S.; Alshater, H.; Abouelmagd, L.M.; Darwish, A.; Hassanien, A.E. Organic photovoltaic prediction model based on Bayesian optimization and explainable AI. *Sci. Rep.* **2025**, *15*, 32559. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Rizk-Allah, R.M.; Abouelmagd, L.M.; Darwish, A.; Snasel, V.; Hassanien, A.E. Explainable AI and optimized solar power generation forecasting model based on environmental conditions. *PLoS ONE* **2024**, *19*, e0308002. [\[CrossRef\]](#) [\[PubMed\]](#)

21. Petrosian, O.; Zhang, Y. Solar power generation forecasting in smart cities and explanation based on explainable AI. *Smart Cities* **2024**, *7*, 3388–3411. [[CrossRef](#)]
22. Varshitha, K.J.; Kumar, L.P.; Manellore, P.; Devanathan, B.; Lakshmanan, S.A.; Prakash, N.K. Explainable AI-enhanced energy forecasting using LightGBM with SHAP and LIME interpretability. In Proceedings of the 2025 IEEE PES GTD Grand International Conference and Exposition Asia (GTD Asia), Bangkok, Thailand, 26–29 November 2025; pp. 369–373.
23. Inglis, A.; Parnell, A.; Hurley, C.B. Visualizing variable importance and variable interaction effects in machine learning models. *J. Comput. Graph. Stat.* **2022**, *31*, 766–778. [[CrossRef](#)]
24. Goldstein, A.; Kapelner, A.; Bleich, J.; Pitkin, E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Stat.* **2015**, *24*, 44–65. [[CrossRef](#)]
25. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.

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