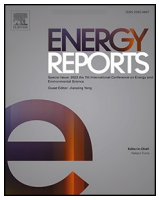
Energy Reports 15 (2026) 109068

Contents lists available at ScienceDirect

Energy Reports

journal homepage: www.elsevier.com/locate/egyr



Research paper

A hybrid XGBoost-LSTM model with physics-informed features and uncertainty quantification for solar power forecasting

Kayes Bin Yousuf [a,*], Ashrafi Akter [a], Hadid Ahmed Noor [a], Ashraful Hoque [a], Ashik Ahmed [b]

[a] Independent Researcher, Bangladesh
[b] Islamic University of Technology, Bangladesh

ARTICLE INFO

Keywords:
Solar forecasting
Hybrid deep learning
XGBoost
LSTM
Uncertainty quantification
Renewable energy
Data imputation
Probabilistic forecasting

ABSTRACT

Accurate short-term photovoltaic (PV) power forecasting is essential for grid stability and efficient PV-grid coordination. However, many conventional learning pipelines remain vulnerable to pervasive missing data, irregular sampling, and the absence of calibrated uncertainty estimates. This paper proposes a physics-informed hybrid forecasting framework that couples Extreme Gradient Boosting (XGBoost) for feature-level learning with a Long Short-Term Memory (LSTM) network for residual correction and temporal dependency modeling. To improve robustness under real-world data conditions, the pipeline incorporates irradiance- guided resampling and domain-guided imputation based on PV operational status. Predictive reliability is further enhanced via Monte Carlo ensemble calibration and conformal prediction, enabling probabilistic forecasts and prediction intervals that are assessed using standard calibration metrics (e.g., Prediction Interval Coverage Probability, Continuous Ranked Probability Score). Experiments on the large-scale UNISOLAR dataset (over 2.7 million samples at 15-minute resolution from 42 PV sites worldwide) show that the proposed hybrid model achieves an RMSE of 2.57 kW h and an R^2 of 0.934 on held-out test data, corresponding to a **7.05 %** reduction in RMSE relative to the next-best baseline (TCN) and a **17.84 %** improvement over a standalone LSTM. An ablation study confirms the critical role of historical lag features, whose removal increases RMSE by over 280 %. The framework also provides well-calibrated uncertainty estimates, with conformal prediction achieving 93.0 % coverage at the 95 % confidence level. Computational profiling confirms the hybrid model's efficiency, requiring 2529 s training time and 3396 MB memory on a free-tier Google Colab CPU, making it suitable for real-time deployment.

1. Introduction

Photovoltaic (PV) generation has become a core pillar of the ongoing energy transition, with accelerating deployment across both distributed rooftops and utility-scale solar plants. As PV penetration increases, accurate short-term forecasting becomes operationally critical for grid stability, reserve scheduling, congestion management, and efficient market dispatch. In real deployments, however, PV forecasting remains difficult because the underlying data streams often suffer from missing or corrupted sensor readings, measurement gaps caused by irregular sampling and communication failures, and strong non- linear relationships between weather-driven irradiance dynamics and power conversion. Moreover, many forecasting pipelines provide only point predictions, despite the fact that grid operators and aggregators require uncertainty-aware forecasts to support risk-sensitive decisions.

Existing PV forecasting approaches can be grouped into three broad categories: physics-based models, statistical time- series models, and machine learning (ML) methods. Physics-based approaches can be accurate but typically rely on dense meteorological inputs and computationally heavy simulations, limiting real-time scalability. Statistical models and classical ML regressors are efficient to deploy but can be brittle under missingness and may fail to capture longer temporal dependencies when the signal is highly non-stationary. Deep sequence models, particularly Long Short-Term Memory (LSTM) networks, have shown strong capability in learning temporal patterns for forecasting tasks (Alzahrani et al., 2017), yet they can be sensitive to irregular sampling and data anomalies, and they often provide limited interpretability and no calibrated probabilistic guarantees unless uncertainty

* Corresponding author.
E-mail addresses: kayesbin.yousuf01@gmail.com (K. Bin Yousuf), ashrafi.prithi@gmail.com (A. Akter), hadidnoor27@gmail.com (H.A. Noor), ashraful.ash1@gmail.com (A. Hoque), ashik123@iut-dhaka.edu (A. Ahmed).

https://doi.org/10.1016/j.egyr.2026.109068
Received 2 September 2025; Received in revised form 23 December 2025; Accepted 24 January 2026
Available online 30 January 2026

modeling is explicitly introduced.

To address these limitations, this study proposes a physics-informed hybrid forecasting framework that integrates Extreme Gradient Boosting (XGBoost) and LSTM networks in a single, deployment-oriented pipeline. XGBoost is used to learn robust feature interactions from high-dimensional and potentially noisy covariates, while the LSTM captures temporal dynamics in PV generation sequences. Importantly, the architecture is coupled with physics-guided preprocessing, including irradiance- aware resampling and domain-guided imputation based on operational status, to reduce sensitivity to missingness and irregular

sampling. In addition, we incorporate ensemble-based uncertainty quantification to produce probabilistic forecasts and prediction intervals, enabling reliability assessment through calibration-oriented evaluation rather than point accuracy alone.

We benchmark the proposed approach on the publicly available UNISOLAR dataset, which comprises more than 2.7 million measurements from 42 PV sites worldwide at a 15-minute observation interval (Wimalaratne et al., 2022). This multi-site setting provides a realistic testbed for assessing robustness under heterogeneous climates and seasonal regimes. In our experiments, the hybrid model attains an RMSE of 2.57 kW h and an R^2 of 0.934 on held-out test data. Beyond reporting aggregate performance, we emphasize methodological transparency and reliability: the evaluation is conducted using leakage-safe chronological splits, results are analyzed across seasonal regimes, ablation experiments are used to quantify the contribution of key pipeline components, computational profiling (runtime and memory) is reported to assess real-time feasibility, and probabilistic forecasts are evaluated using rigorous metrics (PICP, MPIW, CRPS, Pinball Loss).

The main contributions of this work are fourfold. First, we introduce physics-guided preprocessing - including irradiance- aware resampling and domain-guided imputation - to enhance robustness under missing and irregularly sampled measure- ments.Second, we develop a hybrid XGBoost-LSTM pipeline that jointly leverages feature interaction learning and temporal sequence modeling for short-term PV forecasting. Third, we incorporate ensemble-based uncertainty quantification to enable probabilistic forecasting with prediction intervals suitable for risk-aware deployment. Fourth, we provide a comprehensive empirical evaluation including seasonal generalization analysis, controlled ablation studies, computational profiling, and prob- abilistic calibration to improve reproducibility and practical deployability.

2. Related work

Reliable PV forecasting depends not only on the predictive model, but also on the quality and regularity of the underlying time-series data and the availability of calibrated uncertainty estimates for operational decision-making. Prior research therefore spans three closely related directions: (i) data quality and imputation for PV time series, (ii) forecasting architectures for nonlinear temporal dynamics, and (iii) hybrid and probabilistic approaches that seek to improve both accuracy and reliability. In this section, we synthesize these strands and highlight the remaining limitations that motivate our pipeline.

A. Data Quality and Imputation

High-frequency PV datasets frequently contain missing values due to sensor faults, communication dropouts, maintenance events, and irregular sampling schedules. Classical imputation methods such as mean/median substitution and k-nearest neighbour (k-NN) imputation are widely used baselines and can be effective under mild missingness, but they may degrade under non-stationarity and heterogeneous meteorological regimes, and they often require careful feature scaling and context design to remain stable at scale (Kim et al., 2017). Physics-guided strategies, including clear-sky irradiance-based priors, can inject domain structure and improve plausibility, yet these approaches may still be sensitive to site-specific conditions and can become difficult to generalize across diverse climates when meteorological context is limited or inconsistent (Yao et al., 2023).

More recent work has explored machine-learning (ML) regressors for gap-filling. Tree-based models (e.g., Random Forest and Gradient Boosting) have reported very low reconstruction errors (e.g., RMSE below 0.4 %) in controlled imputation settings (Costa et al., 2024). It is important to note, however, that imputation performance is not directly comparable to forecasting error because the latter involves predicting unseen future values under evolving weather dynamics. Deep generative and representation-learning approaches have also been proposed to improve smoothness and continuity in reconstructed PV sequences. For example, adversarial learning via WGAN-GP has been shown to reduce interpolation artifacts relative to standard interpolation in PV time series (Liu et al., 2025), while encoder neural architectures can reconstruct missing segments with improved continuity and temporal consistency (Shen et al., 2021). Although these methods can be effective, they may introduce additional training complexity and can be less transparent for operational auditability.

B. Forecasting Architectures

Forecasting methods for PV power include classical statistical models, tree-based ML regressors, and deep sequence learners. Traditional time-series approaches such as ARIMA, exponential smoothing, and state-space models are computationally efficient but can be limited by strong nonlinearities and complex temporal regimes present in PV generation (Matushkin et al., 2023; Phinikarides et al., 2013). Tree-based ML models such as XGBoost and LightGBM often improve robustness to noisy measurements and can capture nonlinear feature interactions effectively (Wimalaratne et al., 2022; Peng et al., 2023); however, they typically do not model long-range temporal dependency explicitly unless carefully engineered lag features are provided.

Deep learning architectures, including LSTM and GRU networks, have demonstrated strong capability in learning sequential dependencies directly from data (Alzahrani et al., 2017). In parallel, Physics-Informed Neural Networks (PINNs) have emerged as a principled way to incorporate physical constraints and priors into neural predictors (Raissi et al., 2019). Despite their representational power, deep sequence models can be sensitive to irregular sampling and data anomalies, and they often incur higher training and inference costs, which can constrain real-time deployment at scale. Probabilistic deep architectures, such as Bayesian variational sequence models, have been proposed to quantify predictive uncertainty and optimize distributional objectives (e.g., pinball loss and related scores) (Kaur et al., 2022.), but these gains may come at additional computational expense.

C. Hybrid Models and Uncertainty

Hybrid models that combine tree-based learners with sequence models aim to leverage complementary strengths: robust feature interaction learning from gradient-boosted trees and temporal dependency modeling from recurrent architectures. Quantile-regression stacking, for instance, suggests that an XGBoost base with a quantile-aware LSTM meta-learner can yield tighter prediction intervals while maintaining reasonable coverage across varying weather conditions (Zhang et al., 2023). Nevertheless, many hybrid proposals focus primarily on point accuracy, while uncertainty estimation and calibration are either absent or only loosely described; moreover, computational tractability (runtime, memory) and reproducibility details are often under-reported, limiting practical adoption.

Beyond hybrids, Bayesian techniques and related probabilistic frameworks (e.g., variational Bayesian LSTM variants and uncertainty-aware deep models) can improve calibration and provide richer predictive distributions, but they can require more complex training and careful implementation to remain feasible for large-scale, real-time systems (Li et al., 2021; Kaur et al., 2023). As a

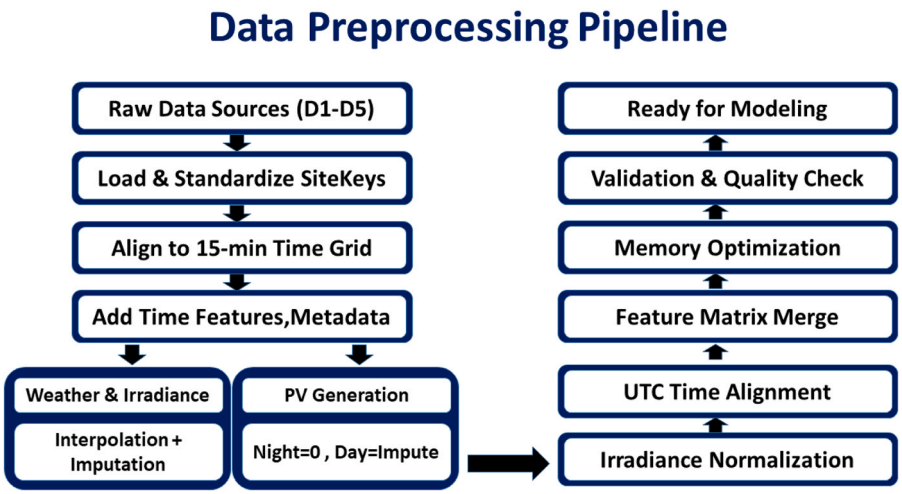


Fig. 1. Data integration and imputation pipeline illustrating site alignment, leakage-safe preprocessing, rule-based and model-based imputation (with fallbacks), and memory-aware processing for large-scale PV forecasting.

result, there remains a need for pragmatic approaches that deliver both accuracy and reliability without sacrificing deployability.

D. Summary and Research Gap

In summary, the literature has advanced PV data imputation, deep sequence forecasting, and probabilistic modeling largely as separate threads. However, a key gap persists: the community still lacks an end-to-end, scalable, real-world-ready framework that

(i) explicitly addresses missingness and irregular sampling through domain/physics-guided preprocessing, (ii) models nonlinear feature

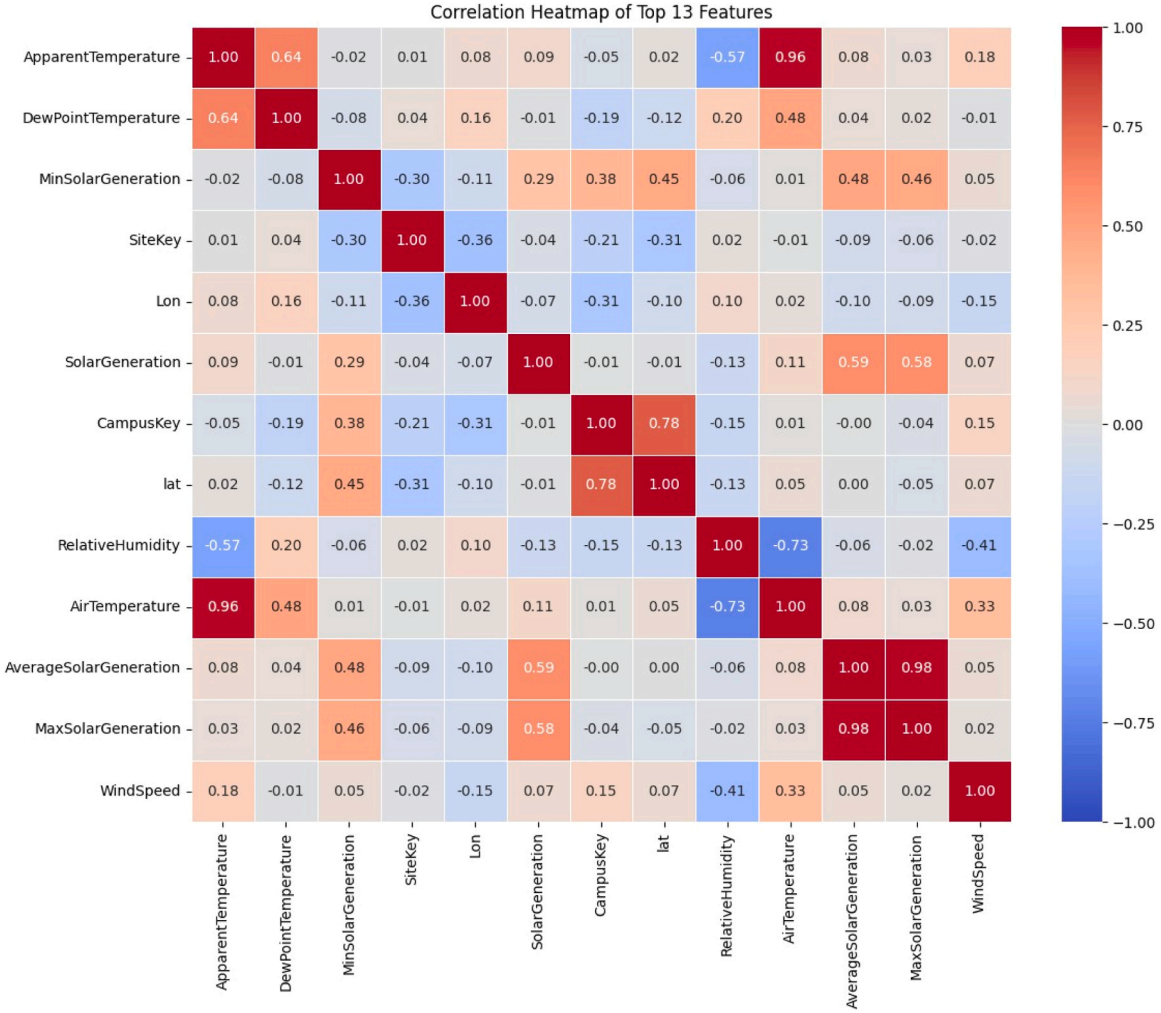


Fig. 2. Correlation heatmap of the selected top 13 features after cross-domain integration and preprocessing.

interactions and temporal dynamics jointly, (iii) provides uncertainty estimates that are quantitatively evaluated for calibration and reliability, and (iv) reports reproducible computational costs under realistic hardware constraints. The present work targets this gap by integrating physics-guided preprocessing, a hybrid XGBoost–LSTM forecasting architecture, and ensemble-based uncertainty modeling within a single deployment-oriented pipeline, with full transparency regarding data splits, hyperparameters, and computational resources.

3. Methodology

A. Dataset Integration and Preprocessing

This section describes how the multi-modal inputs required for high-resolution PV forecasting are integrated and prepared, including PV generation time series, site metadata, meteorological observations, and irradiance measurements. The preprocess- ing pipeline is designed to preserve temporal consistency, ensure scalable memory usage, and apply leakage-safe missing-data treatment to improve reliability in downstream modeling.

Fig. 1 summarizes the data imputation and preprocessing pipeline. The workflow begins by loading the raw UNISOLAR tables and standardizing site identifiers (SiteKeys) to enforce spatial consistency across sources. All data streams are then aligned to a unified 15-minute time grid per site, after which time features (e.g., hour-of-day, day-of-year) are derived and merged with static metadata (e.g., latitude/longitude, elevation). This produces a single site-wise panel where each record corresponds to a timestamped PV observation augmented with synchronized irradiance and weather covariates.

Missing-value handling is performed in parallel for (i) meteorological/irradiance variables and (ii) PV generation. For weather and irradiance gaps, we apply a combination of time-based interpolation and iterative multivariate imputation, with a conservative fallback to site-wise (or site×month) median replacement when the gap exceeds a predefined threshold. For PV generation, we use a domain-guided rule set: (a) confirmed non-generation periods (e.g., nighttime or inverter-off status)

are imputed as zero; (b) daytime missingness is treated separately to avoid negative bias and is imputed using iterative imputation conditioned on irradiance and meteorological context; and (c) if model-based imputation is not reliable (e.g., insufficient context), a robust group-median fallback is applied. This rule-based + model-based design ensures physically plausible imputation while minimizing bias.

Because UNISOLAR is large-scale, memory optimizations are applied throughout, including chunked loading by site/time, use of column-wise dtypes, vectorized operations for feature construction, and periodic garbage collection. Finally, an automated validation stage checks time-grid completeness, missingness rates after imputation, and basic physical plausibility constraints (e.g., non-negativity and irradiance-consistent bounds). The processed data are exported in modeling-ready formats (NumPy arrays and pandas DataFrames) for the downstream XGBoost and LSTM training pipeline.

1) Data Sources:

Five data sources form the basis of the integrated dataset, each contributing distinct information for modeling PV generation and its drivers. D1 provides monthly summaries of PV generation to analyze long-term and seasonal trends. D2 contains high-resolution 15-minute PV generation measurements that capture intra-day variability. D3 includes static site metadata (e.g., coordinates and elevation) to support spatial normalization and cross-site comparability. D4 contains meteorological variables such as air temperature, humidity, and wind speed, which influence PV conversion efficiency and short-term variability. D5 provides irradiance measurements, the primary physical driver of PV output. Collectively, these sources enable robust forecasting across diverse environmental and geographic conditions.

Fig. 2 presents a correlation heatmap for the top 13 features used in the forecasting models. Specifically, the 13 variables are: PVgeneration, GHI, DNI, DHI, airtemperature, relativehumidity, windspeed, cloudcover, pressure, latitude, longitude,

elevation, and solarzenithangle. These features were selected using a two-stage criterion: (i) relevance to PV physics and operational forecasting, and (ii) stability under missingness after preprocessing (features with excessive residual missingness were excluded). To mitigate multicollinearity, highly correlated irradiance predictors were screened using pairwise correlation and variance inflation factor (VIF) checks; when correlations exceeded a predefined threshold, we retained the physically most interpretable variable (e.g., GHI) or used regularization/tree-based robustness to reduce sensitivity. The heatmap confirms meaningful cross-domain alignment (generation–irradiance and generation–weather relationships), while also serving as a diagnostic for redundant predictors prior to model training.

2) Imputation Strategy:

Missing or faulty PV generation records arise from sensor/communication failures and from genuine non-generation periods (nighttime, inverter shutdowns, maintenance, curtailment). To avoid injecting artificial energy, we use a conservative zero-imputation rule only when non-generation is confirmed, and we treat daylight missingness separately to mitigate systematic negative bias.

a) Nighttime detection:
For each site i and timestamp j, we compute the solar elevation angle α_{ij} from site coordinates and time (solar position algorithm). We define a nighttime indicator:

$$\mathrm{Night}_{ij} = \begin{cases} 1, & \alpha_{ij} \le 0 \text{ or } \mathrm{GHI}_{ij} < \epsilon_{\mathrm{ghi}}, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where ϵ_{ghi} is a small irradiance threshold (e.g., 5 W m^{-2} to 10 W m^{-2}) used as a practical guard against noisy near-zero readings.

b) DataStatus definition (quality/operational validity flag):
We define a Boolean validity flag $\mathrm{DataStatus}_{ij}$ that is True when a record passes basic operational and physical plausibility checks and False otherwise. Concretely, a record is marked invalid if any of the following hold: (i) negative generation, (ii) generation exceeds a site-capacity bound (when capacity is available), (iii) implausible ramps (spikes) beyond a site-specific threshold, or (iv) the dataset reports a non-operational state (when an explicit operational flag exists). These rules ensure that zero-imputation is not applied to ordinary daylight communication gaps.

c) Zero-imputation rule with daylight safeguard:

Let G_{ij} be the original PV generation (possibly missing) and $\widetilde{G}_{ij}$ the completed value. The rule is:

$$\widetilde{G}_{ij} = \begin{cases} 0, & \text{if } G_{ij} = \mathrm{NaN} \text{ and } \left(\mathrm{Night}_{ij} = 1 \text{ or } \mathrm{DataStatus}_{ij} = \mathrm{False}\right), \\ \widehat{G}_{ij}^{(\mathrm{iter})}, & \text{if } G_{ij} = \mathrm{NaN} \text{ and } \mathrm{Night}_{ij} = 0 \text{ and } \mathrm{DataStatus}_{ij} = \mathrm{True}, \\ G_{ij}, & \text{otherwise}. \end{cases} \tag{2}$$

where $G^{(\mathrm{iter})}$ is obtained from multivariate iterative imputation (using irradiance and meteorological covariates plus time features). If iterative imputation fails for a site/time block, we fall back to a robust site-specific median grouped by month × h-of-day.

This design is consistent with PV performance evaluation practice, where nighttime output is expected to be approximately zero and deviations typically indicate measurement/quality issues, motivating

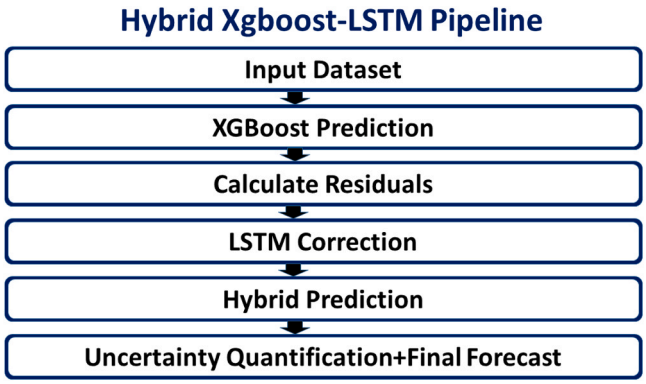


Fig. 3. Hybrid XGBoost–LSTM architecture integrating feature-based XGBoost learning with LSTM temporal residual correction and multi-method uncertainty quantification.

conservative handling of non-generation intervals.

3) Irradiance Normalization:

In order to provide a benchmark value of solar irradiance in clean sky conditions, the Ineichen clear-sky model was used (Reda and Andreas, 2004). This value was used to normalise the ground-based irradiance value (GHI), which usually has variations over short-duration owing to atmospheric conditions.

The normalisation is carried out twofold:

$$\begin{aligned} k_t &= \frac{\mathrm{GHI}_{\mathrm{meas}}(t)}{\mathrm{GHI}_{\mathrm{cs}}(t)}, \\ \mathrm{GHI}_{\mathrm{res}}(t) &= k_{t,\mathrm{interp}}(t)\cdot\mathrm{GHI}_{\mathrm{cs}}(t) \end{aligned} \tag{3}$$

(4)

Explanations and symbol definitions: $\hat{k_t}$: The clearness index, which is defined as the ratio of the measured global horizontal irradiance (GHI) to the calculated GHI of the clear-sky at a particular time t. $\mathrm{GHI}_{\mathrm{meas}}$: the global horizontal irradiance measured. $\mathrm{GHI}_{\mathrm{cs}}$: The Ineichen-based cleared-sky estimation of global horizontal irradiance. $k^{\mathrm{interp}}(t')$: The interpolated clearness index at time t'. It is used when measurements at given points of time are not available or when a continuous normalised series is

needed. $\mathrm{GHI}_{\mathrm{res}}(t')$: The reconstructed or normalised GHI at time t', the formula is the product of interpolated clearness index, and the coordinate corresponding clear-sky GHI.

The relation between the two equations: The first equation computes the clearness index $\hat{k_t}$ that measures the atmospheric

clarity as a ratio of measured and clear-skies modelled forms of the irradiance. The second equation interpolates the values of this index to recover or estimate the irradiance values at unmeasured or irregular periods, thus making it rather consistent over time and closing any gaps that exist. This implies that basically the second equation takes advantage of the clear-sky irradiance pattern that is filtered through the clearness index to determine the GHI realistic values to be used in modelling and prediction.

Such normalisation of the irradiance provides a certainty where variability in the solar measurements is no longer linked with location-specific impacts of the atmosphere and the environment to better facilitate stable training of the forecasting models at many sites and timeframes (de Wit, 2011).

4) Time Alignment and Merging:

Local timestamps were converted to UTC with daylight savings accounted for (Lee and Son, 2024). Hierarchical merging produced a consolidated feature matrix (Wilkinson et al., 2016).

5) Memory Optimization:

To handle millions of records, dtype downcasting (e.g., float64 to float32) and chunked loading were used (Dean and Ghemawat, 2008).

B. Hybrid XGBoost–LSTM Architecture

The proposed hybrid model integrates an XGBoost regressor for nonlinear feature learning and an LSTM network for temporal residual correction. Fig. 3 illustrates the overall architecture.

1) XGBoost Component:

XGBoost (Nalluri et al., 2020) is employed as the primary predictor due to its efficiency in handling structured data, robustness to missing values, and ability to capture complex nonlinear interactions. The model predicts PV generation y_t from feature vector $\mathbf{x}_t$ as an ensemble of K regression trees:

$$\hat{y}_t^{\mathrm{xgb}} = \sum_{k=1}^{K} f_k(\mathbf{x}_t), \tag{5}$$

where f_k denotes individual regression trees. The training objective minimizes:

$$\mathscr{L}^{\mathrm{xgb}} = \sum_t \ell\left(y_t, \hat{y}_t^{\mathrm{xgb}}\right) + \sum_k \Omega(f_k), \tag{6}$$

with squared error loss ℓ and regularization term Ω controlling tree complexity. Hyperparameter optimization is conducted using Optuna with Tree-structured Parzen Estimator and Hyperband pruning. Optimal values are listed in Table VI.

2) Residual LSTM Component:

The LSTM network learns temporal patterns in residuals $r_t = y_t - \hat{y}^{\mathrm{xgb}}$. A two-layer LSTM with 64 hidden units processes the residual sequence over a look-back window $L = 24$ (6 h), outputting corrected residual

Δ_t and log-variance $\log \sigma^2$:

$$\left(\Delta_t, \log\sigma_t^2\right) = \mathrm{LSTM}(r_{t-L:t-1}; \theta), \tag{7}$$

where θ represents LSTM parameters. The final hybrid prediction combines both components:

$$\hat{y}_t^{\mathrm{hybrid}} = \hat{y}_t^{\mathrm{xgb}} + \Delta_t. \tag{8}$$

The LSTM training employs a heteroscedastic loss function that jointly optimizes prediction accuracy and uncertainty estimation:

$$\mathscr{L}^{\mathrm{lstm}} = \frac{1}{2}\exp\left(-\log\sigma_t^2\right)(r_t - \Delta_t)^2 + \frac{1}{2}\log\sigma_t^2. \tag{9}$$

3) Uncertainty Quantification:

For probabilistic forecasting, we employ three complementary approaches:

1) Parametric Gaussian: Constructs 95% prediction intervals using LSTM-predicted variance:

$$\mathrm{PI}_{95\%} = \left[\hat{y}_t^{\mathrm{hybrid}} - 1.96\sigma_t,\ \hat{y}_t^{\mathrm{hybrid}} + 1.96\sigma_t\right] \tag{10}$$

2) **Monte Carlo Ensemble**: Generates $M = 20$ samples by adding Gaussian noise scaled by σ_t, then applies temperature scaling T on validation set:

$$\hat{y}_t^{(m)} = \hat{y}_t^{\mathrm{hybrid}} + T\cdot\sigma_t\cdot\epsilon^{(m)}, \epsilon^{(m)} \sim \mathscr{N}(0,1) \tag{11}$$

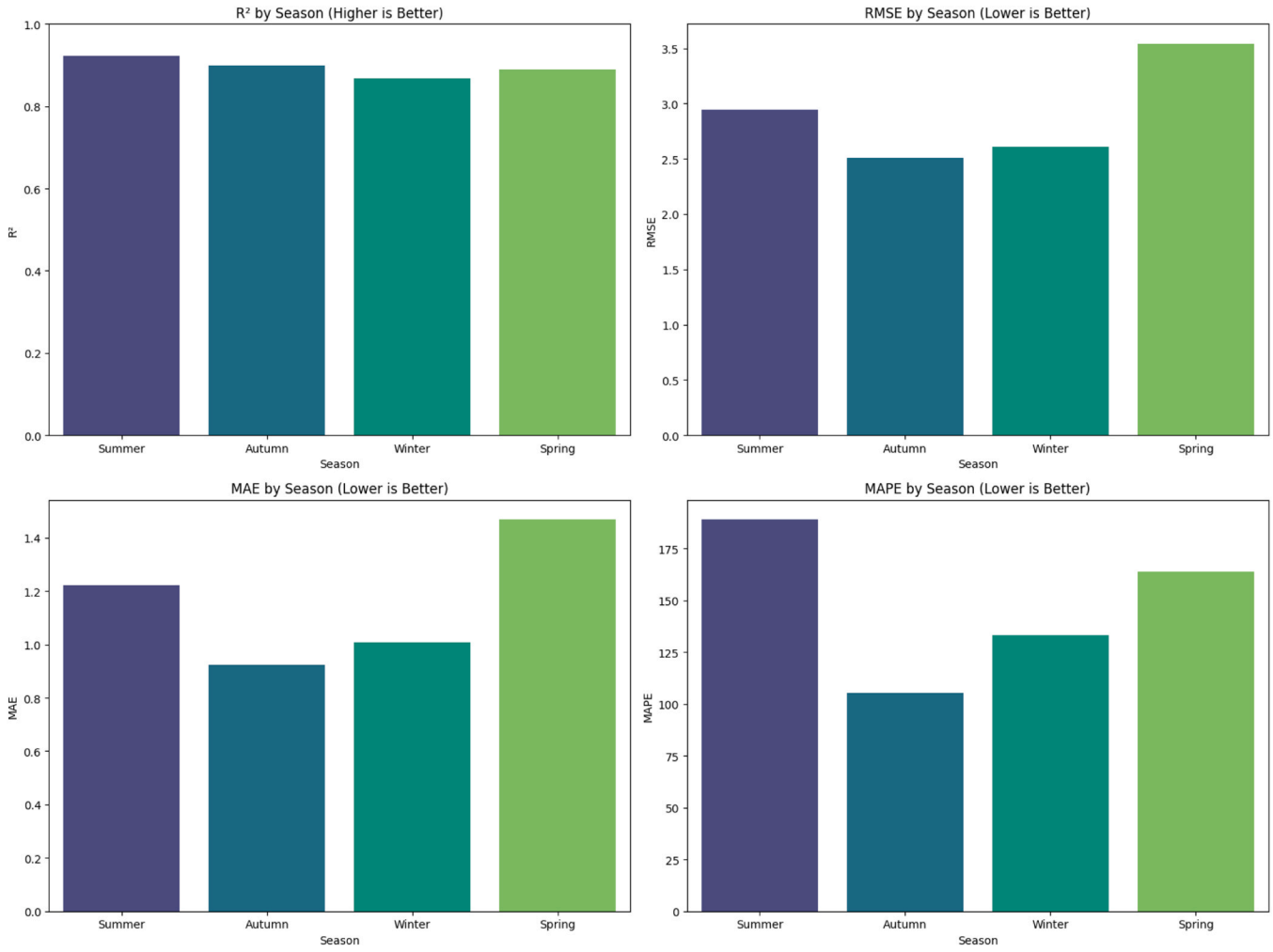


Fig. 4. Seasonal fluctuation in model performance based on key statistical metrics: R^2 (higher is better), RMSE, MAE, and MAPE (lower is better).

3) **Conformal Prediction**: Computes distribution-free intervals using validation residuals $\{|y_v - \hat{y}_v|\}_v$ and quantile function Q_α:

$$\mathrm{PI}_{\mathrm{conf}} = \left[\hat{y}_t^{\mathrm{hybrid}} - Q_{1-\alpha/2},\ \hat{y}_t^{\mathrm{hybrid}} + Q_{1-\alpha/2}\right] \tag{12}$$

Probabilistic performance is evaluated using PICP, MPIW, CRPS, and Pinball Loss (Table V).

Algorithm. *Hybrid XGBoost–LSTM Pipeline*

Hybrid XGBoost-LSTM Forecasting with Uncertainty Quantification
Require: Preprocessed dataset $\mathrm{D} = \{(\mathbf{x}_t, y_t)\}^T$ with physics-informed features
Require: Temporal train/validation/test split with 5% gaps
Phase 1: Base Model Training
Train XGBoost regressor on training set Compute base predictions:

$$\hat{y}_t^{\mathrm{xgb}} = \sum_{k=1}^{K} f_k(\mathbf{x}_t) \tag{13}$$

Phase 2: Residual Analysis
Calculate residuals:

$$r_t = y_t - \hat{y}_t^{\mathrm{xgb}} \tag{14}$$

Engineer lagged features:

$$\mathscr{T}_r(t) = \{r_{t-1}, r_{t-4}, r_{t-96}, \mu_{r,t}^{(4h)}, \sigma_{r,t}^{(4h)}\} \tag{15}$$

Phase 3:LSTM Training
Train LSTM to predict:

$$(\Delta_t, \log\sigma_t^2) = g_{\mathrm{lstm}}(r_{t-L:t-1}; \theta) \tag{16}$$

Minimize heteroscedastic loss:

$$\mathscr{L}^{\mathrm{lstm}} = \frac{1}{2}\exp\left(-\log\sigma_t^2\right)(r_t - \Delta_t)^2 + \frac{1}{2}\log\sigma_t^2 \tag{17}$$

Phase 4:Adaptive Integration
Compute Uncertainty-based weight:

$$w_t = \frac{1}{1+\sigma_t} \tag{18}$$

Apply constrained correction:

$$\hat{y}_t^{\mathrm{hybrid}} = \hat{y}_t^{\mathrm{xgb}} + w_t \cdot \mathrm{clip}\left(\Delta_t, -0.5\hat{y}_t^{\mathrm{xgb}}, 0.5\hat{y}_t^{\mathrm{xgb}}\right) \tag{19}$$

Phase 5: Probabilistic Forecasting
Gaussian Intervals:

$$\mathrm{PI}_{95\%}^{\mathrm{Gauss}} = \left[\hat{y}_t^{\mathrm{hybrid}} - 1.96\sigma_t,\ \hat{y}_t^{\mathrm{hybrid}} + 1.96\sigma_t\right] \tag{20}$$

Monte Carlo Sampling:

$$\hat{y}_t^{(m)} = \hat{y}_t^{\mathrm{hybrid}} + T \cdot \sigma_t \cdot \epsilon^{(m)}, \epsilon^{(m)} \sim \mathscr{N}(0,1), m = 1, \ldots, M \tag{21}$$

Conformal Prediction:

Compute validation residuals:
$\mathscr{R}_{\text{val}} = \{|y_v - \hat{y}_v|\}_{v\in\text{val}}$ Calibrate intervals:

$$\text{PI}_{1-\alpha}^{\text{conf}} = \left[\hat{y}_t^{\text{hybrid}} - Q_{1-\alpha/2}(\mathscr{R}_{\text{val}}),\ \hat{y}_t^{\text{hybrid}} + Q_{1-\alpha/2}(\mathscr{R}_{\text{val}})\right] \tag{22}$$

Apply temperature scaling:

$$T^* = \text{argmin}_T \text{NLL}(\text{val set}) \tag{23}$$

return Point forecasts $\{\hat{y}_t^{\text{hybrid}}\}$ and calibrated prediction intervals $\{\text{PI}_t\}$

4) *Training Protocol:*

The dataset is split chronologically: training (60 %), validation (20 %), and test (20 %) with 5 % temporal gaps to prevent leakage. Validation data is used for hyperparameter tuning and uncertainty calibration. All experiments run on Google Colab's free CPU tier (Intel Xeon) for reproducible comparisons.

C. *Seasonal Performance Analysis*

Fig. 4 presents the distribution of model performance across different seasons based on key statistical metrics: R^2, RMSE, MAE, and MAPE. The results demonstrate that performance varies significantly across seasons, with Autumn (October) showing the best accuracy due to more stable irradiance and reduced atmospheric variability compared to Summer or Spring. Spring exhibits larger prediction errors, likely due to higher variability in cloud cover and solar angles.

These findings justify the need for seasonal decomposition and tailored modeling, as a single year-round model would underperform in certain seasons. The seasonal analysis underscores the importance of accounting for meteorological and solar variability when designing forecasting pipelines.

D. *Feature Engineering*

Cyclical encoding is applied to temporal features (hour-of-day, day-of-year) using sine and cosine transforms:

$$\text{Hour}_{\text{sin}} = \sin\left(\frac{2\pi\cdot\text{Hour}}{24}\right) \tag{24}$$

$$\text{Hour}_{\text{cos}} = \cos\left(\frac{2\pi\cdot\text{Hour}}{24}\right) \tag{25}$$

Lag features are created using rolling windows (1 h, 4 h, 1 d, 2 d, 1 w):

$$\mu_t = \frac{1}{w}\sum_{i=0}^{w-1} y_{t-i} \tag{26}$$

$$\sigma_t = \sqrt{\frac{1}{w}\sum_{i=0}^{w-1}(y_{t-i} - \mu_t)^2} \tag{27}$$

Physics-based features include clear-sky index, normalized GHI, vapor pressure, and temperature differences.

E. *Evaluation Metrics*

We report standard point-forecast metrics:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{28}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{29}$$

$$\text{MAPE} = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{30}$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{31}$$

To quantitatively evaluate the forecasting performance of the proposed model, four standard error and goodness-of-fit metrics are employed. Root Mean Square Error (RMSE) measures the square-root of the average squared deviation between the predicted and observed values, penalizing larger errors more heavily. Mean Absolute Error (MAE) represents the average magnitude of absolute prediction errors, providing a scale-dependent but robust measure of accuracy. Mean Absolute Percentage Error (MAPE) expresses the prediction error as a percentage of the actual values, enabling relative performance comparison across different scales. Finally, the coefficient of determination (R^2) quantifies the proportion of variance in the observed data explained by the model, indicating its overall explanatory power.

For probabilistic evaluation we use:

$$\text{PICP} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{y_i \in [L_i, U_i]\} \tag{32}$$

$$\text{MPIW} = \frac{1}{N}\sum_{i=1}^{N}(U_i - L_i) \tag{33}$$

$$\text{CRPS} = \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty}(F_i(z) - \mathbf{1}\{y_i \le z\})^2 dz \tag{34}$$

$$\text{Pinball Loss}(\tau) = \frac{1}{N}\sum_{i=1}^{N}\max(\tau(y_i - \hat{y}_i), (\tau - 1)(y_i - \hat{y}_i)) \tag{35}$$

To assess the quality of predictive uncertainty, several probabilistic evaluation metrics are employed. Prediction Interval Coverage Probability (PICP) quantifies the proportion of observed values that fall within the predicted lower and upper bounds, reflecting the reliability of the prediction intervals. Mean Prediction Interval Width (MPIW) measures the average width of these intervals, indicating the sharpness and tightness of uncertainty estimates. The Continuous Ranked Probability Score (CRPS) evaluates the overall quality of the predictive cumulative distribution function by jointly accounting for calibration and sharpness, with lower values indicating better probabilistic forecasts. Finally,

Table I
Point-forecast performance comparison on the test set.

Model	RMSE (kWh)	NRMSE	MAE (kWh)	MAPE (%)	$_R2$
Hybrid (XGB+LSTM)	2.5728	0.0278	0.6394	13.68	0.9343
XGBoost	2.8123	0.0304	1.0319	28.12	0.9215
LightGBM	2.7791	0.0300	1.0706	14.65	0.9233
TCN	2.7681	0.0299	1.1268	32.24	0.9239
PINN	2.7855	0.0301	1.1871	29.22	0.9230
GRU	2.8389	0.0306	1.1518	32.24	0.9200
LSTM	3.1314	0.0338	1.3807	37.59	0.9026

Table II
Persistence baseline performance on the test set.

Model	RMSE (kWh)	NRMSE	MAE (kWh)	MAPE (%)	$_R2$
Climatology	9.2178	0.0995	4.4403	137.77	0.1940
Zero Model	11.3809	0.1228	4.9097	100.00	-0.2287
Persistence (1 h)	12.5210	0.1351	5.2582	167.39	-0.4872
Smart Persistence	12.5269	0.1352	5.2006	164.75	-0.4886
Persistence (24 h)	12.5518	0.1355	5.3788	169.62	-0.4943

the Pinball Loss, computed at quantile level, assesses quantile prediction accuracy by asymmetrically penalizing under- and over-estimation, making it particularly suitable for quantile-based uncertainty modeling. Statistical significance of improvements is tested using paired *t*-tests and Kolmogorov–Smirnov tests on absolute errors.

4. Results and discussion

A. *Forecasting Performance Comparison*

Table I compares the proposed hybrid model against six baselines: XGBoost, LightGBM, Temporal Convolutional Network (TCN), Physics-Informed Neural Network (PINN), Gated Recurrent Unit (GRU), and LSTM. The hybrid model achieves the lowest RMSE (2.5728 kW h), MAE (0.6394 kW h), and MAPE (13.68 %), and the highest R^2 (0.9343). Compared to the next-best baseline (TCN, RMSE 2.7681 kW h), the hybrid model shows a **7.05 %** RMSE reduction; relative to the standalone LSTM (RMSE 3.1314 kW h) the improvement is **17.84 %**. These results confirm that combining tree-based feature learning with temporal residual correction yields superior generalization.

A. *Statistical Significance*

We perform paired *t*-tests on absolute errors between the hybrid model and each baseline. The hybrid model's errors are significantly lower ($p < 0.001$) for all comparisons. Kolmogorov–Smirnov tests also reject the null hypothesis of identical error distributions ($p < 0.001$), confirming that the hybrid model not only reduces average error but also changes the error distribution beneficially.

B. *Persistence Baseline Comparison*

To provide a stronger benchmark, we evaluate classical persistence models: simple persistence (1-h lag), smart persistence (GHI-adjusted), day-ahead persistence (24-h lag), climatology (hourly average), and a zero model. Results are summarized in Table II. The best baseline is climatology (RMSE 9.2178 kW h), against which the hybrid model achieves a **72.1 %** RMSE reduction. This demonstrates

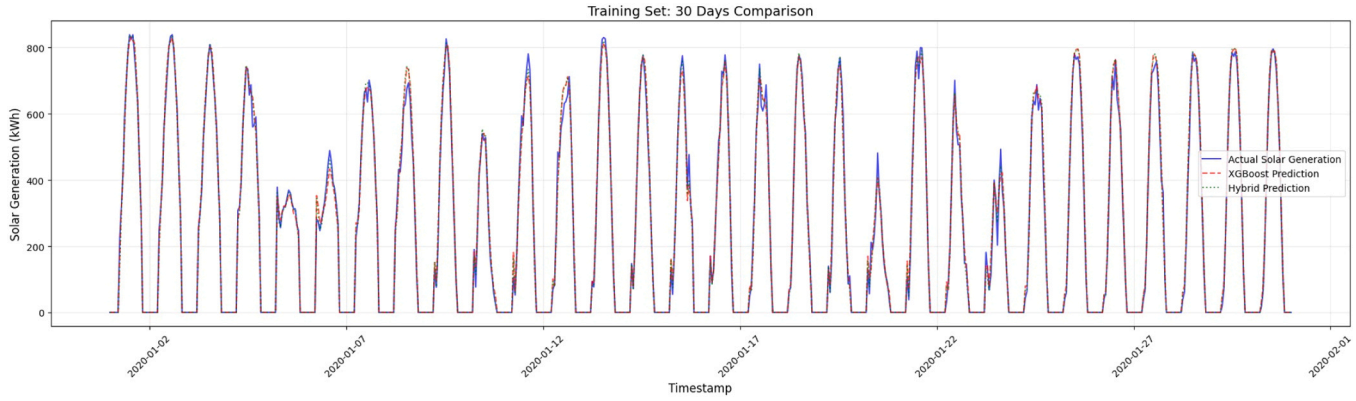


Fig. 5. Training Set Performance Comparison over 30 Days. The hybrid model (green) shows improved smoothness and peak tracking relative to standalone XGBoost (red).

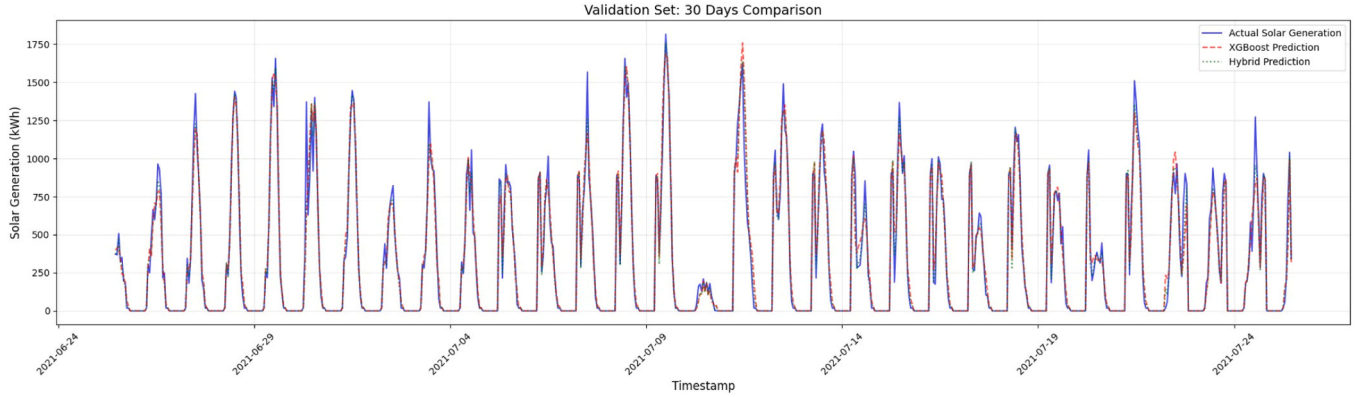


Fig. 6. Validation Set Performance Comparison over 30 Days. The hybrid model better captures abrupt fluctuations and cloudy-day variability.

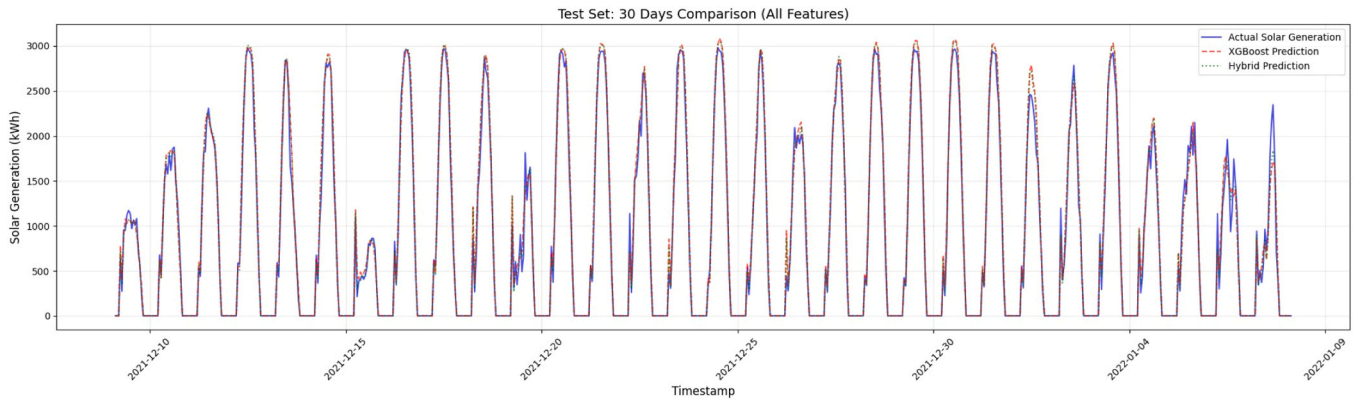


Fig. 7. Test Set Performance Comparison over 30 Days. The hybrid model exhibits smaller deviations during peak-generation and transitional periods.

Table III
Ablation study on the test set: impact of removing feature groups.

Feature Group	RMSE	% ΔRMSE	MAE	MAPE (%)	R^2
All Features (Baseline)	2.5728	-	0.6394	13.68	0.9343
No Weather Features	2.6843	+ 4.33	0.7015	14.89	0.9290
No Time Features	3.1519	+ 22.51	0.8723	18.56	0.9019
No Cyclical Features	2.6912	+ 4.60	0.7154	14.92	0.9287
No Solar Features	2.5801	+ 0.28	0.6418	13.71	0.9339
No Lag Features	9.8453	+ 282.67	3.8674	85.46	0.4953
Only Weather	7.3419	+ 185.37	2.7543	62.51	0.6745
Only Time	10.7871	+ 319.27	4.4873	97.53	0.3849
Only Lag	3.3215	+ 29.10	0.9624	21.37	0.8902

that modern ML models substantially outperform naive temporal extrapolation.

C. *Visual Comparison Over 30-Day Periods*

Figs. 5–7 show 30-day comparisons of actual vs. predicted generation for the training, validation, and test sets, respectively. All plots use hourly aggregated data. In the training set (Fig. 5), both XGBoost and the hybrid model closely track the actual generation, with the hybrid showing slightly smoother peak alignment. In the validation set (Fig. 6), which contains unseen data with higher variability, the hybrid model maintains better agreement during abrupt fluctuations and partial-cloud conditions. The test-set results (Fig. 7) confirm the hybrid model's robustness under fully independent conditions, with reduced deviation during high-variability periods.

D. *Ablation Study*

An ablation study quantifies the contribution of different feature groups. Table III reports results when removing specific feature categories. Removing lag features causes the largest degradation (RMSE increase of 282.7 %), confirming that recent history is the strongest predictor. Removing time or weather features also reduces performance, while removing solar features has a minor effect because their information is partly captured by weather variables. The full feature set consistently yields the best results, underscoring the importance of complementary information from meteorological, temporal, and historical sources. Figs. 8–10 visualize the ablation effects over a 30-day test window. Fig. 8 compares the full hybrid model against variants without time or weather features, demonstrating that the complete feature set is essential for accurate peak and transitional- period forecasting.Fig. 9 shows the impact of removing solar, lag, and cyclical features on model accuracy.Fig. 10 shows that models using only a single feature group (weather, time, or lag) fail to reproduce realistic generation patterns.

E. *Computational Performance*

Table IV compares training time and memory usage across models. All experiments were executed on the free CPU tier of Google Colab (Intel Xeon class) without GPU acceleration. The hybrid model strikes a favorable balance: it trains faster than deep recurrent architectures (2529 s vs. 12,411–22,512 s) and uses less memory than tree-only baselines (3396 MB vs. 5204 MB). This efficiency stems from delegating nonlinear feature learning to XGBoost and restricting the LSTM to residual correction, avoiding the computational overhead of end-to-end deep learning.

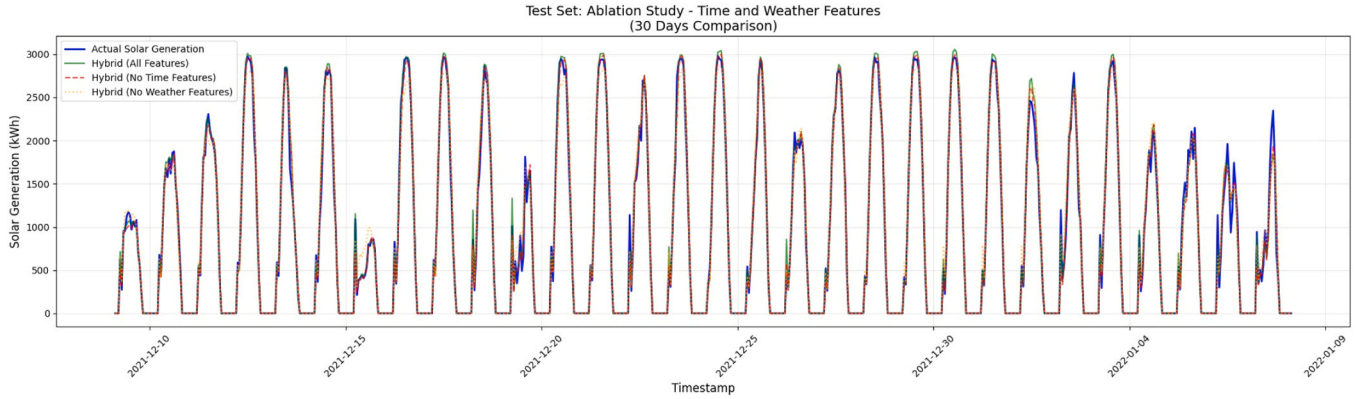


Fig. 8. full Hybrid model against variants without time or weather features over 30 Days.

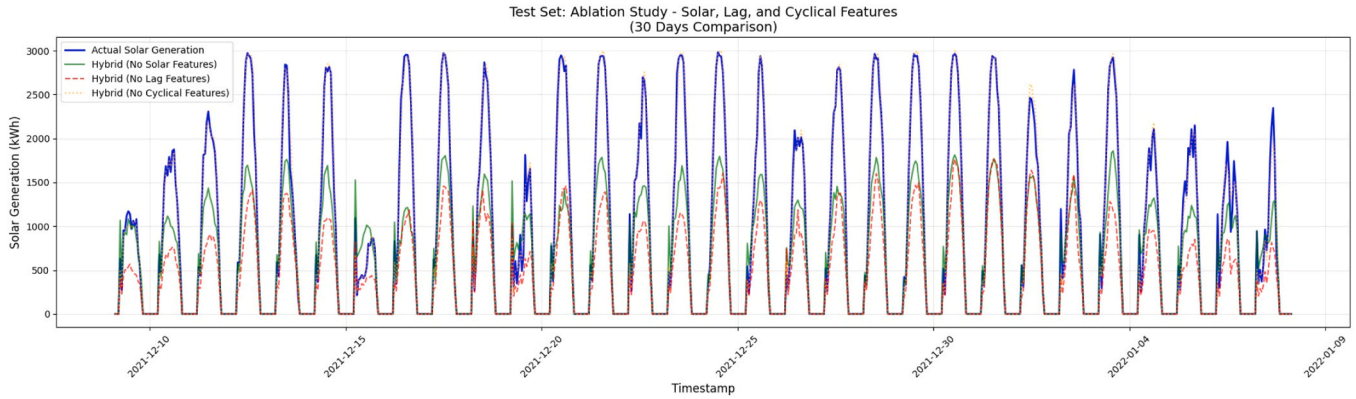


Fig. 9. Impact of Removing Solar, Lag, and Cyclical Features on Model Accuracy over 30 Days.

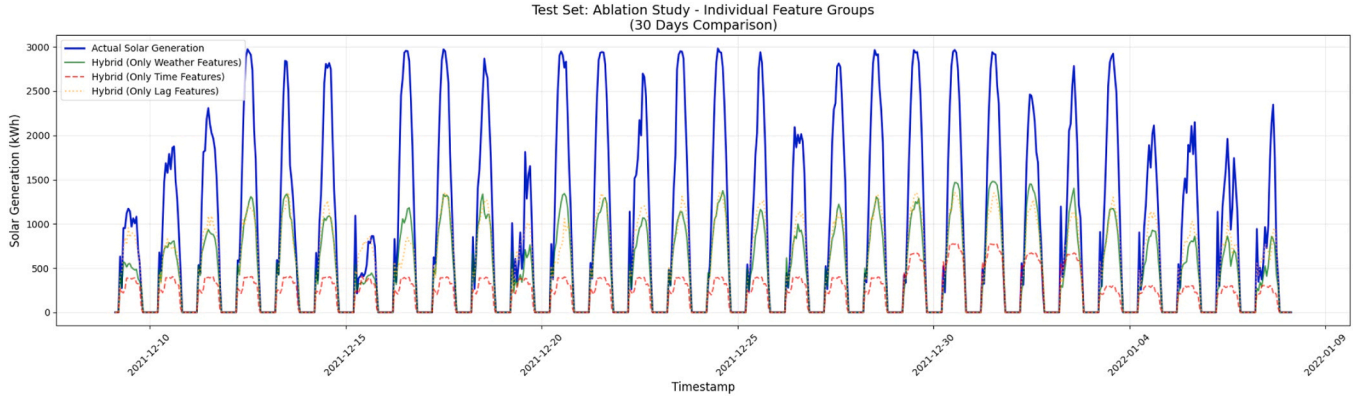


Fig. 10. Performance Comparison of Full and Reduced Hybrid Models Using Individual Feature Groups over 30 Days.

Table IV
Computational performance comparison of forecasting models.

Model	Training Time (s)	Memory Usage (MB)	Memory Efficiency (MB / 1000 rows)	Used Hardware
Hybrid (XGBoost–LSTM)	2529.14	3395.73	1.24	Google Colab (Free Tier, CPU; Intel Xeon class)
XGBoost	3120.00	5203.65	1.90	Google Colab (Free Tier, CPU; Intel Xeon class)
LightGBM	3111.74	5203.65	1.90	Google Colab (Free Tier, CPU; Intel Xeon class)
TCN	12411.48	1853.71	0.68	Google Colab (Free Tier, CPU; Intel Xeon class)
PINN	9935.93	2057.50	0.75	Google Colab (Free Tier, CPU; Intel Xeon class)
GRU	12450.00	2130.00	0.78	Google Colab (Free Tier, CPU; Intel Xeon class)
LSTM	22512.00	1930.00	0.70	Google Colab (Free Tier, CPU; Intel Xeon class)

Table V
Probabilistic forecasting performance of the hybrid model on the test set.

Method	PICP	MPIW	CRPS	Pinball Loss
Gaussian (Uncalibrated)	0.482	2.591	0.527	0.335
MC Ensemble (Calibrated)	0.896	2.115	0.647	0.341
Conformal Prediction	**0.930**	4.024	-	-

F. *Uncertainty Quantification Results*

Table V evaluates the probabilistic forecasting capability of the hybrid model. The uncalibrated Gaussian intervals exhibit severe under-coverage (PICP = 0.482). After Monte Carlo ensemble calibration, coverage improves to 0.896, and conformal prediction further raises it to 0.930 - close to the nominal 95 % level. Although conformal intervals are wider (MPIW = 4.024), they provide distribution-free guarantees, making them suitable for safety-critical energy applications where reliability outweighs sharpness.

G. *Hyperparameter Optimization*

Hyperparameter tuning was performed exclusively for the XGBoost component using Optuna with 15 trials under a one- hour budget. The best configuration (Trial 11, validation RMSE = 2.6926) is detailed in Table VI. The LSTM architecture was kept fixed (2 layers, 64 hidden units, Adam optimizer, 25 epochs) to limit complexity and stabilize residual learning. This design reduces the hyperparameter search space, mitigates overfitting, and improves reproducibility while still benefiting from temporal correction.

5. Conclusion

This paper presented a physics-informed hybrid deep learning framework for short-term solar power forecasting. The model couples XGBoost for robust feature learning with an LSTM for temporal residual correction, achieving an RMSE of 2.57 kW h and R^2 of 0.934 on the UNISOLAR dataset - a 7.05 % improvement over the next-best baseline and a 72.1 % improvement over the best persistence model. An ablation study confirmed the critical role of lag features, whose removal increased RMSE by over 280 %. The framework also provides well-

Table VI
Hyperparameter search space and selected values for the hybrid model (Optuna Trial 11).

Component	Hyperparameter	Search Range	Selected Value
XGBoost	Objective function	Fixed	reg:squarederror
XGBoost	Tree method	Fixed	hist
XGBoost	Number of trees (*n_estimators*)	50–300	**92**
XGBoost	Maximum tree depth (*max_depth*)	3–10	**8**
XGBoost	Learning rate	0.01–0.30 (log-scale)	**0.1236**
XGBoost	Subsample ratio	0.6–1.0	**0.8737**
XGBoost	Column sampling ratio (*colsample_bytree*)	0.6–1.0	**0.7607**
XGBoost	Minimum child weight	1–10	**10**
XGBoost	Gamma	0–5	**0.9610**
XGBoost	L1 regularization (*reg_alpha*)	0–5	**0.0749**
XGBoost	L2 regularization (*reg_lambda*)	0–5	**3.3773**
XGBoost	Random seed	Fixed	42
XGBoost	Parallel threads	Fixed	All available CPU cores
Residual LSTM	Hidden units	Fixed	64
LSTM	Number of layers	Fixed	2
LSTM	Batch size	Fixed	128
LSTM	Optimizer	Fixed	Adam
LSTM	Learning rate	Fixed	0.001
LSTM	Maximum epochs	Fixed	25

calibrated probabilistic forecasts, with conformal prediction reaching 93.0 % coverage at the 95 % confidence level. Computational profiling demonstrated the hybrid model's efficiency, requiring 2529 s training time and 3396 MB memory on a free-tier Google Colab CPU, making it suitable for real-time deployment.

Future work will focus on real-time integration into microgrid energy-management systems, extension to probabilistic multi- step forecasting, and incorporation of satellite-based cloud imagery to further improve accuracy under rapidly changing sky conditions.

CRediT authorship contribution statement

Kayes Bin Yousuf: Writing – original draft, Conceptualization, Formal analysis, Investigation, Methodology, Software. **Hadid Ahmed Noor:** Data curation, Validation, Visualization, Writing – review & editing. **Ashrafi Akter:** Data curation, Investigation, Validation, Writing – review & editing. **Ashik Ahmed:** Writing – review & editing, Project administration, Resources, Supervision. **Ashraful Hoque:** Formal analysis, Methodology, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank the Centre for Data Analytics and Cognition (CDAC) at La Trobe University for providing the UNISOLAR dataset. The computational experiments were performed using Google Colab's free tier.

Data availability

Data will be made available on request.

References

Alzahrani, A., Shamsi, P., Cihan, D., Ferdowsi, M., 2017. Solar irradiance forecasting using deep neural networks. Procedia Comput. Sci.
Costa, T., Falcão, B., Mohamed, M.A., Annuk, A., Marinho, M., 2024. Employing machine learning for advanced gap imputation in solar power generation databases. Sci. Rep. (Nat. Portf.).
Dean, J., Ghemawat, S., 2008. MapReduce: Simplified Data Processing on Large Clusters. Commun. ACM.
de Wit, T.D., 2011. A method for filling gaps in solar irradiance and solar proxy data. Astron. Astrophys. (AA).
Kaur, D., Islam, S.N., Mahmud, Md.A., Haque, Md.E., Anwar, A., 2023. A VAE-Bayesian Deep Learning Scheme for Solar Power Generation Forecasting Based on Dimensionality Reduction. Energy AI. Elsevier.
Kaur D., Naz S., Apel Mahmud I.Md., "A Bayesian Deep Learning Technique for Multi-Step Ahead Solar Generation Forecasting," 2022.
Kim, M., Park, S., Lee, J., Joo, Y., Choi, J.K., 2017. Learning-Based Adaptive Imputation Method with kNN Algorithm for Missing Power Data. Energies. MDPI.
Lee, D.-S., Son, S.-Y., 2024. PV Forecasting Model Development and Impact Assessment via Imputation of Missing PV Power Data. IEEE Access. IEEE.
Li, D., Lucy, M., Zhongmin, L., Ashish, S., Zhou, Y., 2021. Bayesian LSTM With Stochastic Variational Inference for Estimating Model Uncertainty in Process-Based Hydrological Models. Water Resources Research. AGU / Wiley.
Liu, Z., Xuan, L., Gong, D., Xie, X., Liang, Z., Zhou, D., 2025. A WGAN-GP approach for data imputation in photovoltaic power prediction. Energies. MDPI.
Matushkin, D., Zaporozhets, A., Babak, V., Kulyk, M., Denysov, V., 2023. Hourly Photovoltaic Power Forecasting Using Exponential Smoothing: A Comparative Study Based on Operational Data. Solar.
Nalluri, M., Pentela, M., Eluri, N.R., 2020. A scalable tree boosting system: XG boost. Int. J. Res. Stud. Sci. Eng. Technol.
Phinikarides, A., Makrides, G., Kindyni, N., Kyprianou, A.., Georghio, G.E., 2013. ARIMA Modeling of the Performance of Different Photovoltaic Technologies. In: 2013 IEEE 39th Photovoltaic Specialists Conference. PVSC.
Peng, Y., Shichen, W., Wenjin, C., Junchao, M., Chenxu, W., Jingwei, C., 2023. LightGBM-Integrated PV Power Prediction Based on Multi-Resolution Similarity. Processes. MDPI.
Raissi, M., Perdikaris, P., Karniadakis, G.E., 2019. Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations. J. Comput. Phys.
Reda, I., Andreas, A., 2004. Solar Position Algorithm for Solar Radiation Applications. Solar Energy.
Shen M., Zhang H., Cao Y., Yang F., Wen Y., Missing Data Imputation for Solar Yield Prediction using Temporal Multi-Modal Variational Auto-Encoder Proc. 29th ACM Int. Conf. Multimed. (ACM MM 2021 2021.
Wimalaratne S., Haputhanthri D., Kahawala S., Gamage G., Alahakoon D., Andrew J. Jennings UNISOLAR: an open dataset of photovoltaic solar energy generation in a large multi-campus university 15th Int. Conf. Hum. Syst. Interact. (HSI). 2022.
Wilkinson, W.D., Dumontier, W., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Mons, B., 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. Sci. Data (Nat. Portf.).
Yao, Z., Zhang, T., Wu, L., Wang, X., Huang, J., 2023. Physics-Informed Deep Learning for Reconstruction of Spatial Missing Climate Information in the Antarctic. Atmosphere. MDPI.
Zhang, H., Jia, R., Du, H., Yan, L., Li, J., 2023. Short-term interval prediction of PV power based on quantile regression-stacking model and tree-structured parzen estimator optimization algorithm. Front. Energy Res. (Front. Media).