algorithms

MDPI

Article

Sub-Hourly Multi-Horizon Quantile Forecasting of Photovoltaic Power Using Meteorological Data and a HybridCNN–STTransformer

Guldana Taganova [1,2,*], Alma Zakirova [1], Assel Abdildayeva [3,*], Bakhyt Nurbekov [4], Zhanar Akhayeva [1] and Talgat Azykanov [5]

[1] Faculty of Information Technology, L. N. Gumilyov Eurasian National University, Astana 010000, Kazakhstan
[2] Higher School of Information Technology and Engineering, Astana International University, Astana 010000, Kazakhstan
[3] Department of Software Engineering, International Engineering-Technological University, Almaty 050000, Kazakhstan
[4] Mechanics and Mathematics Faculty, L. N. Gumilyov Eurasian National University, Astana 010000, Kazakhstan
[5] "Lumox-AI" Individual Entrepreneur, Astana 010000, Kazakhstan
* Correspondence: guldana.kileuzhanova@gmail.com (G.T.); a.abdildaeva@metu.kz (A.A.)

Abstract

The rapid deployment of photovoltaic generation increases uncertainty in power-system operation and strengthens the need for ultra-short-term forecasts with reliable uncertainty estimates. Point-forecasting approaches alone are often insufficient for dispatch and reserve decisions because they do not quantify risk. This study investigates probabilistic forecasting of short-horizon solar generation using quantile regression on a public dataset of solar output and meteorological variables. This study proposes a hybrid attention–convolution model that combines an attention-based encoder to capture long-range temporal dependencies with a causal temporal convolution module that extracts fast local fluctuations using only past information, preventing information leakage. The two representations are fused and decoded jointly across multiple future horizons to produce consistent quantile trajectories. Experiments against representative machine-learning and deep-learning baselines show improved probabilistic accuracy and competitive central forecasts, while illustrating an important sharpness–calibration trade-off relevant to risk-aware grid operation. Key novelties include a multi-horizon quantile formulation at 15 min resolution for one-hour-ahead PV increments, a HybridCNN–STTransformer that fuses causal temporal convolutions with Transformer attention, and a horizon-token decoder that models inter-horizon dependencies to produce consistent multi-step quantile trajectories; reliability/sharpness diagnostics and post hoc calibration are discussed for operational risk-aware use.

Keywords: photovoltaic forecasting; probabilistic forecasting; quantile regression; pinball loss; multi-horizon prediction; Transformer; CNN; hybrid deep learning; sub-hourly time series

Check for updates

Academic Editor: Vittorio Maniezzo

Received: 8 December 2025
Revised: 27 January 2026
Accepted: 30 January 2026
Published: 3 February 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) license.

1. Introduction

The rapid growth of installed photovoltaic (PV) capacity has increased the share of stochastic generation in modern power systems and reinforced the need for ultra-short-term forecasting methods that are both accurate and uncertainty-aware. At sub-

hourly horizons, PV dynamics are governed by fast-changing weather and cloud motion, producing multi-scale temporal patterns and abrupt regime shifts. This setting defines an algorithmically challenging high-resolution time-series problem where probabilistic multi-step predictions must be jointly learned to support risk-aware operational decisions such as reserve scheduling, congestion management, and intraday balancing.

Traditionally, short-term PV analytics has relied on point forecasting models optimized for MSE or MAE. However, they do not quantify predictive uncertainty and, therefore, offer limited support for risk-aware operational decisions. As a result, they offer limited support for assessing real operational risks. Probabilistic forecasting addresses this problem by describing the future not with a single number, but with a distribution, a forecast interval, or a set of quantiles. This makes it possible to more accurately evaluate the risk of under- and over-generation and to account for extreme scenarios in power system operations and electricity markets [1–3]. In related renewable contexts, probabilistic methodologies have been successfully applied to load and wind forecasting and are increasingly adopted for PV plants [4–9].

Among probabilistic frameworks, quantile regression remains a robust and interpretable approach for producing scenario-consistent forecasts across probability levels and for constructing prediction intervals with prescribed nominal coverage [10–12]. In this study, to ensure tail-awareness with explicit calibration diagnostic uncertainty estimates, we evaluate all models using a fixed quantile grid that characterizes both central tendencies and extreme under-/over-generation regimes relevant for operational risk. In addition, this study evaluates interval reliability using the empirical coverage and width of the central 90% prediction interval defined by the 5th and 95th quantiles, complementing quantile-wise pinball losses and median point errors. Strong baseline models for quantile forecasting include ensemble tree methods such as gradient boosting and XGBoost. They scale well, deliver high accuracy, and can be configured to predict multiple quantiles at once [13–15]. Recurrent neural networks, especially LSTM and GRU, are also widely used. They can capture complex nonlinear time dependencies, which makes them suitable for forecasting renewable energy generation and electricity load [16–18]. More recent deep learning pipelines integrate convolutional modules for local pattern extraction with attention mechanisms and Transformer-based encoders for long-range dependency modeling and cross-feature interactions [19–22]. In parallel, conformal prediction provides a complementary uncertainty framework that can deliver valid predictive intervals under minimal distributional assumptions, which is useful for operational settings with potential distribution shifts [22,23]. Recent reviews emphasize that very short-term solar forecasting remains sensitive to local atmospheric variability and highlight a growing role for hybrid deep learning, attention, remote sensing, and physics-informed features for reliable sub-hourly horizons [24–27]. Emerging paradigms, such as federated learning, indicate future opportunities for scalable and privacy-preserving probabilistic modeling across distributed PV assets [28].

Despite these advances, several algorithmic gaps remain in ultra-short-term multi-horizon probabilistic forecasting. First, many studies still prioritize point accuracy or provide limited probabilistic diagnostics, whereas practical deployment requires joint evaluation of probabilistic accuracy, calibration, and sharpness across quantiles and horizons [1–3,10–12]. Second, sub-hourly multi-step horizons (e.g., four consecutive 15 min steps ahead) remain less systematically studied than hourly or day-ahead settings, even though they pose distinct learning challenges driven by rapid regime changes and higher noise-to-signal ratios in PV dynamics [7–9,24–27]. Third, existing deep quantile pipelines often treat horizons as loosely coupled outputs, leaving inter-horizon dependency modeling insufficiently specified; this can reduce the temporal coherence of predicted quantile trajectories and limit their operational

interpretability [19–22]. Fourth, reproducible, apples-to-apples comparisons of ensemble, recurrent, and Transformer-family hybrid models within a single unified multi-horizon quantile framework on public PV datasets remain limited, complicating the assessment of relative model advantages and practical trade-offs [13–22]. To the best of our knowledge, explicit horizon-token decoding for cross-horizon coupling within a hybrid attention–convolution quantile framework remains underexplored for ultra-short-term PV forecasting.

From an algorithmic standpoint, our key contribution is a structured multi-horizon quantile architecture that couples global attention with local convolution and introduces explicit cross-horizon decoding. This work studies one-hour-ahead PV probabilistic forecasting at a 15 min resolution using the energy increment Energy delta [Wh] derived from the publicly available Renewable Power Generation and Weather Conditions dataset. The forecasting task is formulated as a multi-horizon, multi-quantile prediction for the next four 15 min steps. This study proposes HybridCNN–STTransformer, a hybrid attention–convolution architecture that integrates a Transformer encoder branch with a 1D causal convolution (Conv1D) branch and fuses them into a shared latent representation. A horizon-token decoder with self- and cross-attention is introduced to explicitly model inter-horizon dependencies and generate horizon-specific latent states, followed by a multi-quantile head trained with the pinball loss. Note that, in this manuscript, STTransformer refers to a spatio-temporal Transformer architecture, where "spatial" denotes interactions across the multivariate feature/channel dimension (not geographic space), and it is unrelated to the Stockwell (S-) transform.

The effectiveness of the proposed approach was evaluated using a unified and fully reproducible Python 3.11 (Python Software Foundation, Wilmington, DE, USA) pipeline implemented as Jupyter Notebooks. We benchmark against gradient boosting, XGBoost, LSTM, GRU, and a baseline spatio-temporal Transformer. The models are compared using probabilistic metrics (average pinball loss across quantiles, prediction-interval width, and empirical coverage) and point metrics for the median forecast. This evaluation design enables a systematic analysis of accuracy–calibration–sharpness trade-offs that is critical for deploying uncertainty-aware forecasts in short-term grid operation.

The main contributions of this paper can be summarized as follows:

(1) We introduce HybridCNN–STTransformer, a hybrid attention–convolution architecture for multi-quantile, multi-horizon sub-hourly PV forecasting.
(2) We propose a horizon-token decoding mechanism that explicitly captures inter-horizon dependencies, improving the temporal consistency of multi-step quantile trajectories.
(3) We provide a unified and reproducible benchmark comparing ensemble tree models, recurrent networks, and Transformer-based architectures under the same multi-horizon quantile learning protocol.
(4) We conduct a comprehensive probabilistic evaluation of accuracy, calibration, and sharpness alongside median point performance to assess operational suitability.

2. Materials and Methods

2.1. Dataset

We conduct experiments on the public 'Renewable Power Generation and Weather Conditions' dataset [Source: https://www.kaggle.com/datasets/pythonafroz/renewable-power-generation-and-weather-conditions, accessed on 8 November 2025], representing a single utility-scale PV plant with synchronized meteorological and astronomical variables at 15 min resolution. Specifically, we use the Renewable.csv file covering 1 January 2017–31 August 2022 (196,776 records, 17 variables). The target is the 15 min energy increment, Energy delta [Wh], which is operationally more interpretable for balancing and reserve planning than instantaneous power. The weather feature block includes GHI, temperature, pressure, humidity, wind speed, precipitation, and cloudiness, complemented by derived daylight-related indicators. The

dataset contains a large fraction of zero-generation nighttime intervals, motivating probabilistic modeling and careful evaluation of calibration and sharpness across sub-hourly horizons.

The dataset follows a nominal 15 min sampling scheme over 1 January 2017 to 31 August 2022. A check against the uniform 15 min grid reveals a small fraction of missing timestamps (1824 intervals, ≈0.918%), indicating generally high data completeness for the one-hour-ahead forecasting task.

As this study uses secondary, publicly released data, the authors did not perform primary measurements and did not operate measurement instruments. The public dataset documentation does not provide verifiable information on sensor manufacturers/models or their locations; therefore, these details cannot be reported in this work.

We structure the methodology to ensure an unambiguous link between the data characteristics, the multi-horizon quantile forecasting formulation, the proposed hybrid architecture, and a unified experimental protocol. The section first describes the dataset and target construction, then formalizes the forecasting task and feature design. Next, we summarize the baseline families and present the proposed HybridCNN–STTransformer in a single consolidated block. Finally, we detail the training procedure, evaluation metrics, and implementation settings used consistently across all models.

2.1.1. Target Variable and Input Features

The target variable y_t is defined as the energy produced during the 15 min interval $[t,\ t + 15\ \text{min}]$. For each time step t, we construct a vector of explanatory variables $x_t \in \mathbb{R}^d$ that includes.

Meteorological features

- GHI: global horizontal irradiance (W/m^2), characterizing the instantaneous solar resource;
- temp: near-surface air temperature (°C);
- pressure: atmospheric pressure (hPa);
- humidity: relative humidity (%);
- wind_speed: wind speed (m/s);
- rain_1h: precipitation (rain) accumulated over the last hour (mm);
- snow_1h: equivalent snowfall over the last hour (mm);
- clouds_all: cloudiness indicator, typically expressed as the percentage of sky cover.

Astronomical and derived features

- isSun: binary indicator of solar elevation above the horizon (day/night);
- sunlightTime: elapsed time from sunrise, i.e., duration of the current daylight period;
- dayLength: total length of the daylight period for the given calendar date;
- SunlightTime/dayLength: normalized fraction of the daylight period elapsed;
- weather_type: categorical code describing the prevailing weather conditions (clear, cloudy, rain, etc.).

Calendar features

- hour: hour of day (0–23);
- month: month of year (1–12).

This combination of physical (irradiance, temperature, humidity), astronomical (daylight-related), and calendar features provides a rich representation of the state of the system and supports both classical gradient-boosting models and more complex deep learning architectures. The marginal distribution of Energy delta [Wh] is highly skewed, with a large mass at zero due to nighttime intervals, and heavy tails during daytime generation. These properties motivate the use of quantile-based probabilistic methods rather than purely Gaussian assumptions.

2.1.2. Data Quality and Missing Values

A small fraction of missing timestamps (≈0.918%) is observed relative to the nominal 15 min grid. We align the series to a uniform 15 min index; missing target values are not imputed and are excluded from training windows, while meteorological covariates, when needed, are forward-filled over short gaps to maintain feature continuity without introducing artificial generation.

2.2. Exploratory Data Analysis

The empirical analysis is conducted on the Renewable Power Generation and Weather Conditions dataset with 15 min resolution. The sample spans from 1 January 2017 00:00 to 31 August 2022 17:45, providing long-term coverage of intra-day and seasonal variability. A check against the nominal 15 min time grid indicates a small amount of missing timestamps (1824 intervals, ≈0.918% of the expected grid), which is unlikely to distort distributional patterns at the one-hour forecasting scale. The target variable is the energy increment, Energy delta [Wh], which exhibits strong intermittency and pronounced day–night structure. As expected, nighttime values (isSun = 0) are identically zero, while daytime values display a highly right-skewed distribution with a non-negligible mass near zero. Overall, zeros constitute about 51.25% of the full sample, whereas the daytime zero share is much lower (≈6.25%), confirming that stochasticity is concentrated in daylight hours and increases with meteorological variability. This heteroskedastic and asymmetric behavior motivates the adopted multi-horizon quantile regression setup. To capture both median dynamics and operationally relevant tails, we fix the quantile set $Q = \{0.05, 0.50, 0.95\}$, and evaluate predictions across all horizons jointly.

2.2.1. Distributional Properties of the Target

We fix the quantile set $Q = \{0.05, 0.50, 0.95\}$ to capture the median behavior and both tails relevant for under- and over-generation risk at sub-hourly resolution. The target variable is the 15 min energy increment.

Energy delta [Wh]. The dataset spans from 1 January 2017 to 31 August 2022 with a nominal 15 min sampling rate. Nighttime samples (isSun = 0) are deterministically zero, while daytime samples exhibit a strongly right-skewed and intermittent distribution, motivating multi-quantile evaluation across horizons.

As shown in Figure 1, the full-sample distribution is dominated by the day–night mixture, producing a large probability mass at zero. Restricting to daylight reveals a strongly right-skewed and heteroskedastic daytime distribution with a substantially smaller zero share, highlighting the operational relevance of tail-aware probabilistic forecasting.

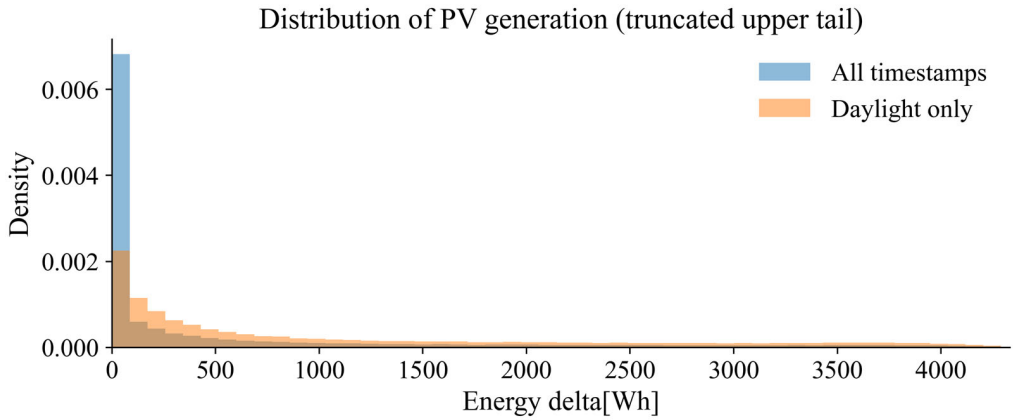


Figure 1. Normalized histograms of Energy delta [Wh] for the full sample and for daylight intervals only (isSun = 1).

2.2.2. Seasonal and Diurnal Patterns

To characterize the structured nonstationarity typical for PV generation, we analyze aggregated profiles across months and hours. The monthly aggregation (mean with 5–95% percentile ranges) reveals a pronounced seasonal cycle, with substantially lower winter values and higher spring–summer production. The broad intra-month percentile bands underline strong weather-driven variability even under similar astronomical conditions.

The seasonal cycle of PV generation increments is summarized in Figure 2.

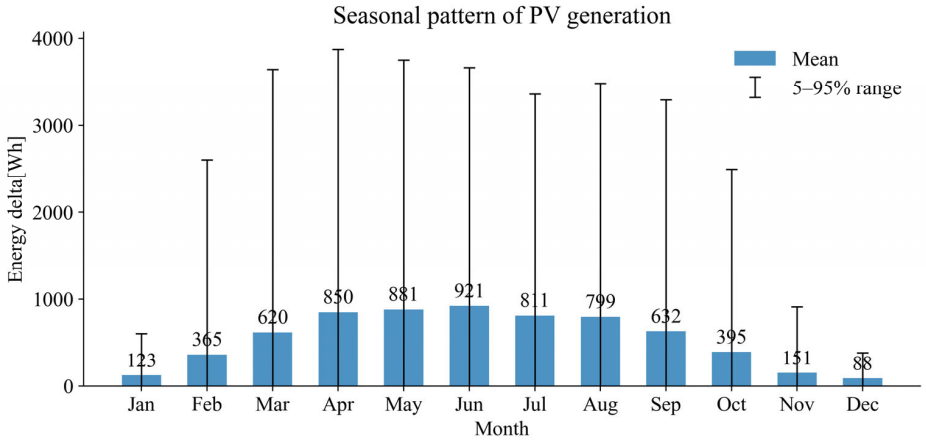


Figure 2. Seasonal profile of PV plant energy output increments, Energy delta [Wh] (mean values and 5–95% range).

Similarly, the hourly aggregation highlights a clear bell-shaped diurnal curve with near-zero generation during night hours and peak values around late morning to midday. The consistent gap between mean and median in daytime indicates right-skewness due to occasional high-generation spikes, while the wide 5–95% band around peak hours indicates maximal uncertainty during periods of high irradiance potential.

The diurnal profile of PV generation increments is shown in Figure 3.

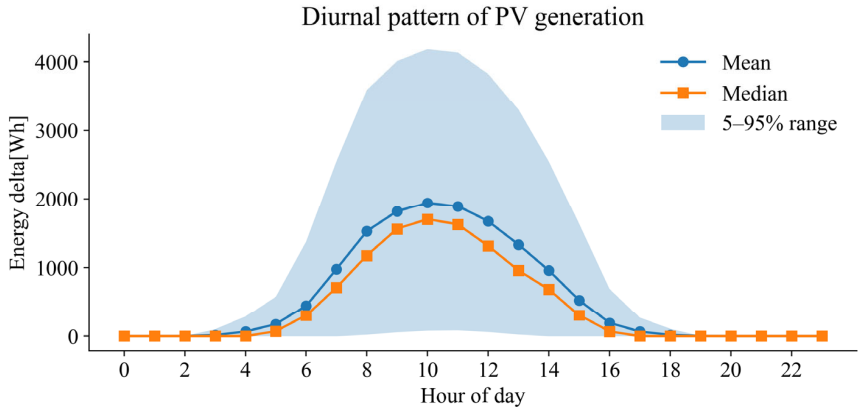


Figure 3. Diurnal profile of PV plant energy output increments, Energy delta [Wh] (mean, median, and 5th–95th percentile range).

2.2.3. Dependence Structure and Informative Predictors

We further examine linear associations between the target and candidate predictors. The largest correlations are expected for lagged and smoothed versions of Energy delta, reflecting short-term persistence, and for GHI and its temporal aggregates, consistent with the physical role of irradiance as the primary driver of PV output.

The most strongly correlated predictors with the target (Energy delta [Wh]) are summarized in Table 1.

Table 1. Features with the highest absolute correlation with the target variable Energy delta [Wh].

№	Feature	Correlation Coefficient
1	Energy delta [Wh]_lag1	0.956
2	Energy delta [Wh]_roll_mean_4	0.935
3	Energy delta [Wh]_lag2	0.925
4	GHI	0.915
5	GHI_lag1	0.907
6	Energy delta [Wh]_lag3	0.899
7	GHI_lag2	0.890
8	Energy delta [Wh]_roll_mean_8	0.888
9	GHI_roll_mean_4	0.888
10	GHI_lag3	0.871
11	Energy delta [Wh]_lag4	0.870
12	GHI_roll_mean_8	0.849
13	GHI_lag4	0.848
14	Energy delta [Wh]_roll_std_8	0.663
15	Energy delta [Wh]_roll_std_4	0.614
16	hour_cos	−0.566
17	isSun_roll_mean_8	0.551
18	hour_cos_lag1	−0.549
19	humidity	−0.545
20	isSun_roll_mean_4	0.542

Figure 4 presents the correlation structure between the target and the most informative features.

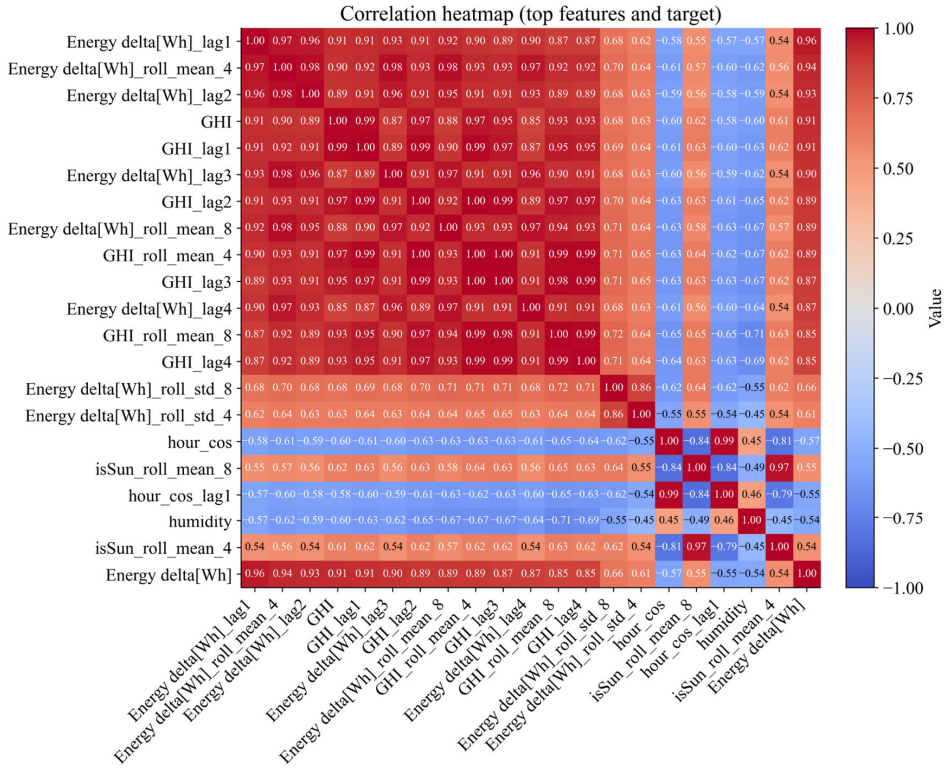


Figure 4. Correlation matrix of the target variable and the most informative features.

To refine the analysis beyond linear effects, we evaluate mutual information between the target and engineered predictors. The results are expected to confirm the dominant importance of the first lag of Energy delta, followed by GHI and short-window rolling statistics. This supports prioritizing lagged/rolling target and irradiance features in both tabular and sequence-based models.

The mutual information scores used to rank predictors are reported in Figure 5.

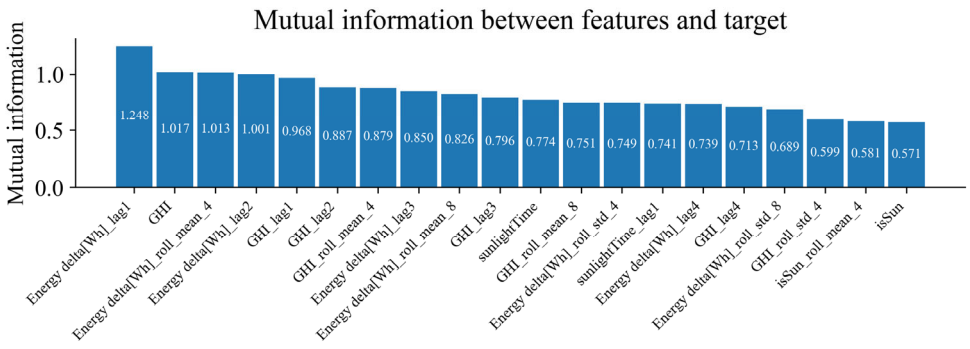


Figure 5. Mutual information between the features and the target variable Energy delta [Wh].

The most informative predictors with respect to the target (Energy delta [Wh]) are summarized in Table 2.

Table 2. Features with the highest mutual information with the target variable Energy delta [Wh].

№	Feature	Mutual Information
1	Energy delta [Wh]_lag1	1.248
2	GHI	1.017
3	Energy delta [Wh]_roll_mean_4	1.013
4	Energy delta [Wh]_lag2	1.001
5	GHI_lag1	0.968
6	GHI_lag2	0.887
7	GHI_roll_mean_4	0.879
8	Energy delta [Wh]_lag3	0.850
9	Energy delta [Wh]_roll_mean_8	0.826
10	GHI_lag3	0.796
11	sunlightTime	0.774
12	GHI_roll_mean_8	0.751
13	Energy delta [Wh]_roll_std_4	0.749
14	sunlightTime_lag1	0.741
15	Energy delta [Wh]_lag4	0.739
16	GHI_lag4	0.713
17	Energy delta [Wh]_roll_std_8	0.689
18	GHI_roll_std_4	0.599
19	isSun_roll_mean_4	0.581
20	isSun	0.571

2.2.4. Conditional Relationships with Key Weather Drivers

Figure 6 illustrates the empirical relationships between daylight PV generation increments and key meteorological drivers. Panel (a) shows a strong positive association

with GHI, consistent with the physical basis of PV generation. Panel (b) exhibits wide dispersion across cloud-cover levels, indicating a nonlinear and mediated cloud effect. Panel (c) suggests that temperature acts as a secondary efficiency modifier with high variance. Panel (d) shows similarly weak direct monotonicity for humidity, motivating multivariate modeling to capture interactions among predictors. For readability, Energy delta is plotted in kWh (Wh/1000) in Figure 6.

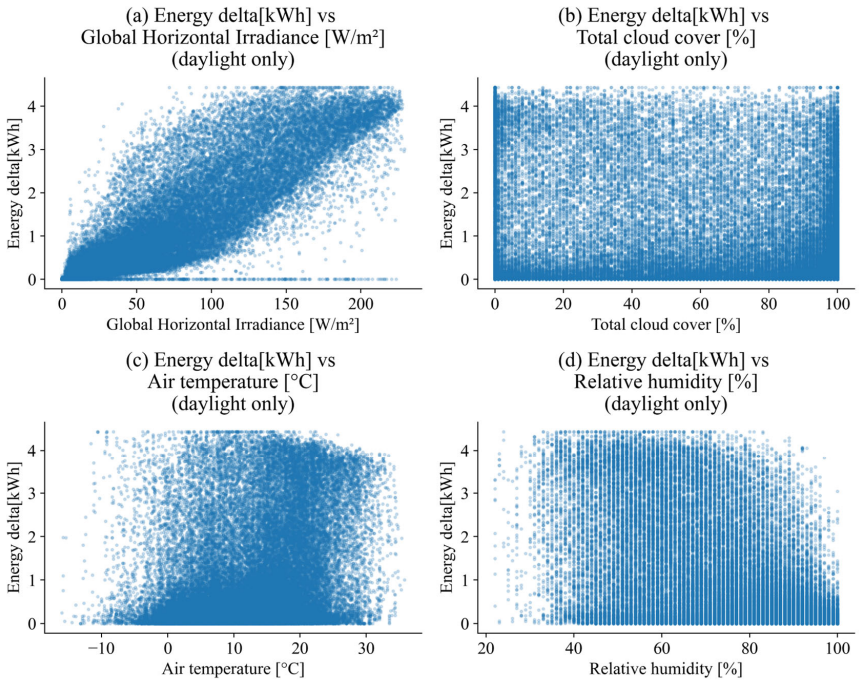


Figure 6. Scatter plots of daylight PV generation increments (Energy delta, kWh) versus key meteorological drivers: (**a**) global horizontal irradiance (GHI, W/m^2), (**b**) total cloud cover (%), (**c**) air temperature (°C), (**d**) relative humidity (%).

As shown in Figure 7, the conditional distribution of generation increments at fixed GHI values exhibits a clearly pronounced monotonically increasing pattern for all three quantiles (the 5th, 50th, and 95th percentiles).

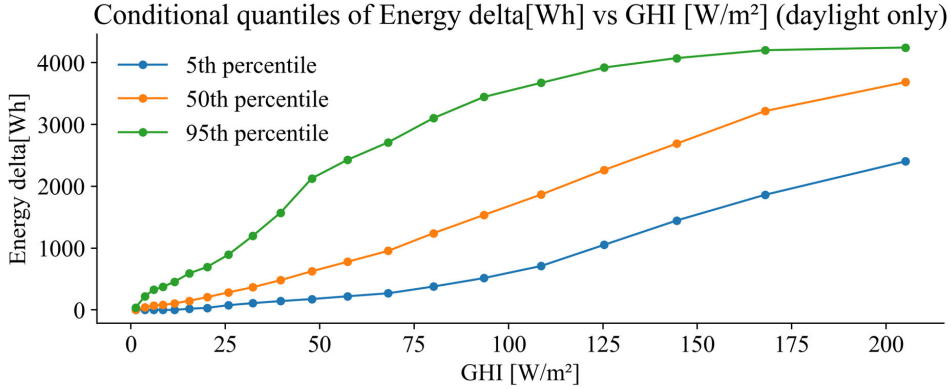


Figure 7. Conditional quantiles of Energy delta [Wh] generation increments as a function of global horizontal irradiance (GHI, W/m^2) (daytime intervals only).

At low irradiance levels (GHI < 50 W/m^2), all quantiles are close to zero, which corresponds to weak or absent insolation. As GHI increases, the median quantile (50th percentile) rises almost linearly, reflecting a steady increase in the typical generation level. The upper quantile (95th percentile) grows more steeply, especially in the mid-irradiance range, indicating the emergence of increasingly high potential output peaks under favorable conditions.

The lower quantile (5th percentile) remains substantially below the median but also increases with GHI, showing that even under unfavorable scenarios, generation at high irradiance is noticeably above zero. The widening gap between the 5th and 95th percentiles as GHI rises provides evidence of heteroskedasticity: forecast uncertainty is greatest precisely under high-irradiance conditions. This relationship confirms the dominant role of GHI in shaping the conditional distribution of generation and, at the same time, underscores the suitability of a quantile-based approach that can capture not only the mean response but also the asymmetry and width of predictive intervals.

As can be seen in Figure 8, the relationship between the conditional distribution of generation and the integrated cloudiness indicator is substantially weaker than that with irradiance, yet it follows the expected direction.

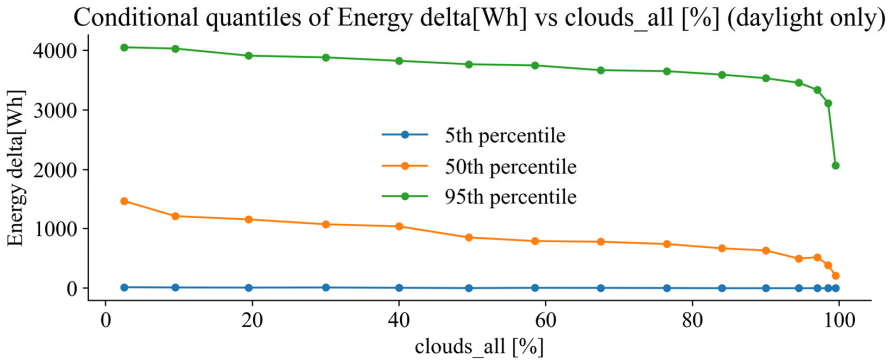


Figure 8. Conditional quantiles of Energy delta [Wh] generation increments as a function of cloud cover, clouds_all (daytime intervals only).

As cloudiness increases from low clouds_all values (≈0–20%) to heavily overcast conditions (≈80–100%), the median quantile (50th percentile) shows a consistent downward trend: the typical level of generation increments decreases by almost half. The upper quantile (95th percentile) also gradually declines with rising cloudiness, but remains at relatively high values (around 3.5–4.0 kWh), reflecting the possibility of episodically high generation even under substantial cloud cover, for instance, under broken or intermittent cloud conditions. The lower quantile (5th percentile) stays close to zero across the entire clouds_all range, emphasizing that in unfavorable scenarios, generation may nearly vanish regardless of the cloudiness level. Taken together, these results suggest that cloudiness exerts a smoothing, but not uniquely deterministic, influence on the generation distribution and are considered in the model jointly with more "hard" irradiance and astronomical predictors when constructing quantile forecasts.

2.2.5. Autocorrelation and Daily Periodicity

As shown in Figure 9, the autocorrelation function (ACF) of the Energy delta [Wh] series, computed from 15 min data and expressed in hourly lags, exhibits a pronounced daily periodicity. At small lags, very high positive autocorrelation values are observed, reflecting strong short-term persistence of the process: generation values in adjacent time intervals are tightly linked. As the lag increases, the coefficient decreases monotonically, crosses zero at around 12 h, and reaches a negative minimum. This pattern can be interpreted as an antiphase relationship between the daytime and nighttime segments of the daily cycle.

Next, around the 24 h lag, the autocorrelation becomes substantially positive again, forming a pronounced maximum that coincides with the length of the astronomical day (the dashed line in the plot). A similar wave-like pattern then repeats across subsequent daily cycles. This ACF shape simultaneously confirms strong short-term memory and stable daily seasonality in the generation series. These characteristics justify incorporating both short lags (to capture local

dynamics) and daily scale features and cyclic time transformations, as well as using sequential architectures (LSTM/GRU) that can explicitly model long-range dependencies.

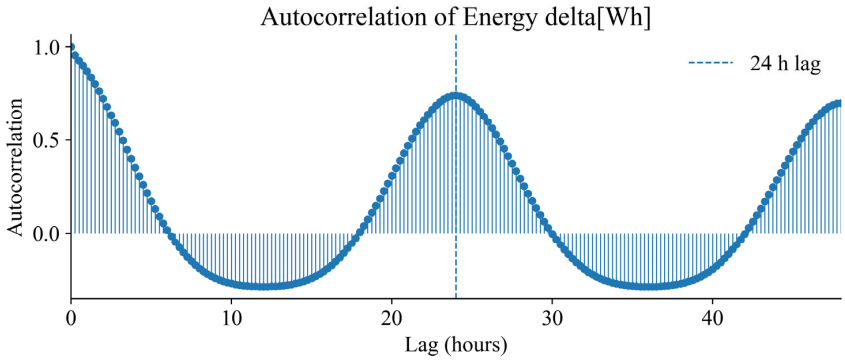


Figure 9. Autocorrelation function (ACF) of the Energy delta [Wh] time series. Blue dots indicate empirical ACF values at discrete hourly lags, while the solid line represents a smoothed trend to guide visual interpretation. The vertical dashed line marks the 24 h lag, highlighting daily periodicity.

Overall, the EDA shows that PV generation is intermittent, strongly shaped by daily and seasonal cycles, and becomes more variable under certain conditions. This supports our multi-horizon quantile regression setup and the proposed HybridCNN–STTransformer, which is designed to capture short-term persistence, nonlinear effects driven by irradiance, and dependencies across forecast horizons.

2.3. Feature Engineering and Problem Formulation

This study considers a 15 min photovoltaic (PV) generation time series, where the target corresponds to the Energy delta [Wh] produced over each 15 min interval and denotes the associated vector of meteorological, astronomical-derived, and calendar explanatory variables available in Renewable.csv. We address one-hour-ahead probabilistic forecasting using four 15 min horizons. For each reference time, the multi-horizon target is defined as corresponding to lead times of 15, 30, 45, and 60 min.

To provide historical context, we construct a sliding input window.

$$X_t = (x_{t-L+1}, \ldots, x_t) \in \mathbb{R}(L \times d)$$

This study set corresponds to a 4 h history at a 15 min resolution. The supervised learning dataset is thus formed by pairs over all valid time indices after windowing.

Guided by the EDA evidence of short-term persistence and diurnal structure, we retain the provided calendar indicators (hour, month) and the astronomical daylight descriptors (isSun, sunlightTime, dayLength, and their normalized ratio). These variables help the models distinguish nighttime zero-generation regimes from daytime variability.

This study adopts a quantile regression framework. Given a fixed set of quantile levels

$$Q = \{0.05,\ 0.5,\ 0.95\}$$

The model aims to approximate conditional quantile functions

$$\acute{y}_h^{(q)} = f_\theta(X_t, h,\ q),\ h \in \{1,\ 2,\ 3,\ 4\},\ q \in Q,$$

where θ denotes the model parameters. For each t, this yields a tensor of predicted quantiles.

$$\acute{Y}_t^{(Q)} = \left[\acute{y}_h^{(q_k)}\right]_{h=1,\ldots,4;\ k=1,\ldots,K} \in \mathbb{R}^{H \times K},$$

with $H = 4$ horizons and K quantile levels.

The parameters θ are estimated by minimizing the average pinball (quantile) loss over all training examples, horizons, and quantiles:

$$\mathcal{L}(\theta) = \frac{1}{|I_{train}|H\,K} \sum_{t \in I_{trin}} \sum_{h=1}^{H} \sum_{k=1}^{K} \rho_{q_k}\left(y_{t+h} - \hat{y}_h^{(q_k)}\right),$$

where $\rho_q(\varepsilon) = q\,\max(\varepsilon, 0) + (1-q)\max(-\varepsilon, 0)$ is the standard pinball loss for quantile level q. This objective encourages a coherent approximation of both the central part and the tails of the conditional distribution of PV output.

To ensure fair comparison across model families, the same multi-horizon formulation and quantile set are used for all baselines and the proposed HybridCNN–STTransformer. For sequential models (LSTM, GRU, STTransformer, and HybridCNN–STTransformer), input windows are provided as three-dimensional tensors, whereas for tree-based models (Gradient Boosting, XGBoost), each sample is represented as a flattened feature vector constructed from the most recent time steps and static features.

2.4. Data Preprocessing and Train–Validation–Test Split

The raw time series are inspected for missing values and time inconsistencies. To avoid any possibility of information leakage, missing meteorological values are imputed using a strictly causal strategy (forward-fill within each split). All scaling/normalization parameters are fitted on the training split only and then applied to validation/test splits.

To avoid information leakage from the future, we apply chronological splitting of the data into three non-overlapping subsets:

Training set: ≈70% of the oldest observations;

Validation set: ≈15% of subsequent observations;

Test set: ≈15% of the most recent observations.

Exact temporal cutoffs (chronological, no shuffling) are used in our released pipeline: the training split covers 1 January 2017 00:00 to 19 December 2020 02:15, the validation split covers 19 December 2020 02:30 to 25 October 2021 09:15, and the test split covers 25 October 2021 09:30 to 31 August 2022 16:45. All model parameters are fit on the training split only; the validation split is used both for early stopping and (when reported) as the calibration split for post hoc conformalized quantile correction (CQR), which is then applied once to the frozen test forecasts.

In the final configuration, this results in approximately 131 input features after adding cyclical time encodings (sin/cos of hour and day-of-year), daylight indicators (isSun, sunlight time), and lagged/rolling statistics. All lagged and rolling features are computed using only past values available at prediction time; rolling windows are trailing and aligned to the end of each 15 min interval to prevent temporal misalignment. All normalization parameters are fitted on the training split only, ensuring that no future information leaks into the training process.

Engineered lag/rolling features and time alignment. To make the feature construction explicit and avoid any temporal leakage, we compute lagged predictors using the previous 15 min steps only (e.g., Energy delta [Wh]_lag1–lag4 and GHI_lag1–lag4). In addition, this study includes trailing rolling means over recent windows (e.g., roll_mean_4 and roll_mean_8) computed strictly from past values within each split. Cyclical encodings (sin/cos of hour and day-of-year) and daylight indicators (isSun, sunlightTime) are appended as time/context features. All lagged and rolling predictors are aligned to the right endpoint of each 15 min interval (i.e., features at time t use information available up to t), and all preprocessing/scaling parameters are fitted on the training split only and applied to validation/test.

For sequential models (LSTM, GRU, STTransformer, and HybridCNN–STTransformer), input windows are provided as three-dimensional tensors of shape (N, L, d), where N is the number of samples. For tabular models (gradient boosting and XGBoost), each sample is represented as a flattened feature vector that concatenates the most recent time steps and static features. The same train–validation–test split is used for all models to ensure fair comparison.

2.5. Forecasting Models

Within the proposed multi-horizon quantile regression setup, we evaluate four model families: (i) tree-based ensemble methods, (ii) recurrent neural networks, (iii) a baseline spatio-temporal Transformer, and (iv) the proposed HybridCNN–STTransformer. All models predict the same set of quantiles for the next four 15 min horizons using a common historical input window and identical train–validation–test splits to ensure fair comparison.

2.5.1. Tree-Based Baseline Models

As strong baselines for tabular data, we employ gradient boosting and XGBoost in their quantile regression variants. For each horizon h and each quantile level $q \epsilon Q$, a separate model is trained with the pinball loss as objective, or equivalently, through appropriate asymmetric weighting in the loss function. Input features include lagged PV outputs, meteorological variables, and calendar indicators. Hyperparameters (number of trees, tree depth, learning rate, and subsampling rates) are selected based on performance on the validation set.

2.5.2. Recurrent Neural Networks

This study considers two recurrent architectures: LSTM and GRU. Both models take the historical window as input and process it through one or two stacked recurrent layers, producing a hidden representation that summarizes the recent history of PV output and meteorological conditions. This representation is then passed to fully connected layers that output predicted quantiles for all horizons:

Recurrent models are trained end-to-end using the multi-horizon pinball loss described in Section 2.2. Dropout and L_2 weight regularization are applied to mitigate overfitting, and early stopping based on validation loss is used to select the final model.

2.5.3. STTransformer Baseline

The spatio-temporal Transformer (STTransformer) baseline treats the sequence $(x_{t-L+1}, \ldots, x_t)$ as a sequence of input tokens and employs multi-head self-attention to capture global temporal dependencies and interactions among features. After positional encoding and a stack of Transformer encoder layers, the model produces a sequence of hidden states, which are then aggregated (e.g., via global average pooling or selection of the last token) into a compact vector $z_{tr} \in \mathbb{R}^{d_{tr}}$. A feed-forward head maps this representation to the tensor of predicted quantiles $\acute{Y}_t^{(Q)} \in \mathbb{R}^{H \times K}$.

Terminology note: In this single-site PV setting, the "spatial" dimension in STTransformer refers to the multivariate feature/channel axis (cross-variable interactions) rather than geographic space or multiple PV sites.

2.5.4. Proposed HybridCNN–STTransformer

The proposed HybridCNN–STTransformer architecture combines the strengths of convolutional and Transformer-based modeling. It consists of two parallel branches that process the same input window $X_t \in \mathbb{R}^{L \times d}$.

A Transformer branch identical in spirit to the STTransformer described above, responsible for capturing long-range temporal dependencies and global interactions among meteorological and astronomical features. Its output is a context vector $Z_{tr} \in \mathbb{R}^{d_{tr}}$.

A convolutional branch that applies one-dimensional causal convolutions along the temporal axis, with several Conv1D layers interleaved with nonlinear activations and pooling. This branch extracts localized patterns related to rapid changes in irradiance and PV output, yielding a feature vector $Z_{cnn} \in \mathbb{R}^{d_{cnn}}$.

The two vectors are concatenated and passed through a series of fully connected layers with normalization and dropout, resulting in a joint latent representation $z \in \mathbb{R}^{d_z}$:

$$\boldsymbol{z} = \varnothing([\boldsymbol{z}_{tr}; \boldsymbol{z}_{cnn}])$$

where $[\cdot;\cdot]$ denotes concatenation and $\varnothing(\cdot)$ is a nonlinear mapping implemented by dense layers. The output head finally maps z to the tensor of quantile forecasts for all horizons:

By jointly parameterizing all horizons and quantiles, the hybrid model can exploit correlations between neighboring lead times and enforce a more coherent shape of the predictive distribution over the one-hour horizon.

The proposed HybridCNN–STTransformer employs a dual-branch encoder that processes the same input window in parallel. The Transformer branch captures long-range temporal dependencies via linear projection, positional embeddings, and stacked multi-head self-attention blocks, while the causal Conv1D branch extracts local high-frequency patterns without future leakage. The two representations are concatenated and projected back to the d_model space with layer normalization to form a shared hybrid encoder output. Forecasting is performed by a horizon-token decoder: learnable horizon embeddings first interact through self-attention across horizons and then attend to the hybrid encoder states through cross-attention. A lightweight quantile head maps each horizon-specific latent vector to the fixed quantile set for all four 15 min lead times.

The overall architecture of the proposed HybridCNN–STTransformer is illustrated in Figure 10.

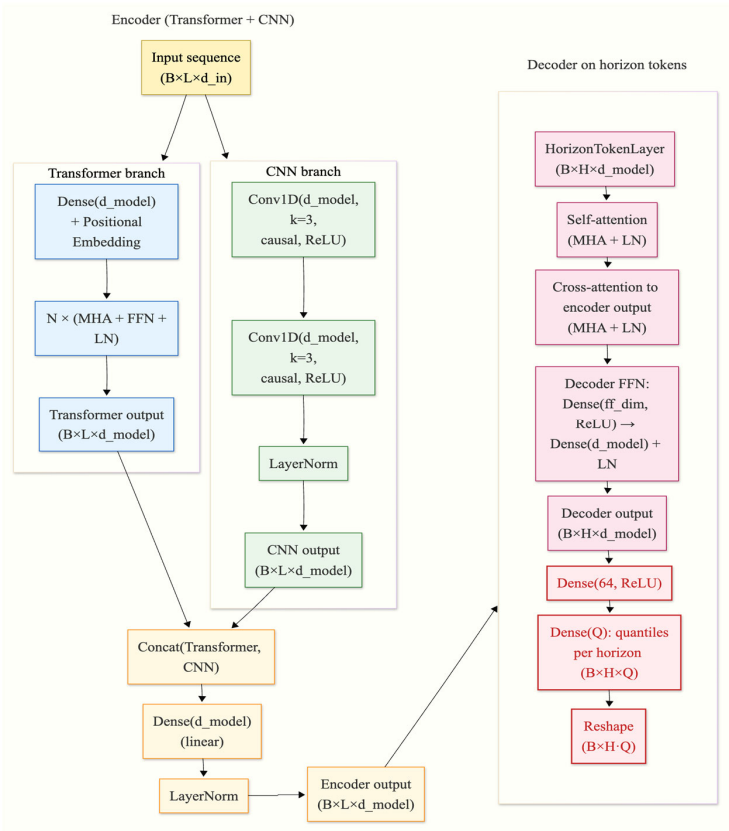


Figure 10. Architecture of the proposed HybridCNN–STTransformer model. Blue blocks denote the Transformer-based branch, green blocks indicate the CNN-based branch, and red blocks correspond to the decoder and output layers. Arrows represent the data flow between model components.

2.6. Training Procedure

All models are trained under a unified experimental protocol to ensure fair comparison within the proposed multi-horizon quantile forecasting setting. For neural architectures (LSTM,

GRU, STTransformer, and the proposed HybridCNN–STTransformer), we optimize the multi-quantile pinball loss aggregated across the four 15 min horizons and all quantile levels. This objective jointly calibrates the central part and the tails of the predictive distribution.

This study uses the Adam optimizer with a fixed learning rate, batch size of 256, and a maximum of 50 epochs. To prevent overfitting, early stopping is applied if the validation pinball loss does not improve for five consecutive epochs. The same training settings are used for all neural baselines and the proposed hybrid model to isolate architectural effects.

For tree-based baselines (Gradient Boosting and XGBoost), quantile models are trained using the corresponding quantile objectives (pinball loss), following a direct multi-horizon strategy where a separate regressor is fitted for each quantile–horizon pair. Hyperparameters are selected based on validation performance and summarized in Table 3.

Table 3. Key hyperparameters of quantile forecasting models.

Model	Hyperparameter	Value
GBM (Gradient Boosting)	Number of trees n_estimators	400
	Learning rate learning_rate	0.05
	Maximum depth max_depth	3
	Minimum samples per leaf min_samples_leaf	50
	Subsample ratio subsample	0.7
	Loss function	Quantile (pinball) loss
XGBoost	Number of boosting rounds num_boost_round	400
	Learning rate eta	0.05
	Maximum depth max_depth	3
	Minimum child weight min_child_weight	1.0
	Subsample	0.7
	Column subsample ratio colsample_bytree	0.8
	Loss function	Custom pinball loss
LSTM-baseline	Encoder structure	LSTM (64, return_seq) → LSTM (32)
	Fully connected layer	Dense (64, ReLU)
	Output layer	Dense(H × |Q|)
	Optimizer/lr	10^{-3} Adam (fixed learning rate)
	Batch size/epochs	256/50, early stopping (5)
GRU-baseline	Encoder structure	GRU (64, return_seq) → GRU (32)
	Other parameters	Same as LSTM
STTransformer-baseline	Model dimension d_{model}	64
	FFN dimension ff_dim	128
	Number of attention heads num_heads	8
	Number of encoder layers	2
	Number of decoder layers	1
	Attention/FF dropout	0.1/0.1
	Other training settings	Adam, batch 256, 50 epochs, early stopping
HybridCNN–STTransformer	Transformer branch	Same as STTransformer-baseline
	CNN branch	Conv1D (64, k = 3, causal, ReLU) × 2
	Fusion layer	Concat → Dense (64) → LayerNorm
	Decoder	Same as STTransformer-baseline with horizon tokens
	Loss function	Multi-quantile pinball loss
	Training settings	Adam, batch 256, 50 epochs, early stopping

For LSTM-baseline, LSTM(64, return_seq) → LSTM(32) denotes a stacked LSTM architecture, where the first LSTM layer has 64 hidden units and returns the full sequence to the next LSTM layer with 32 hidden units.

2.7. Evaluation Metrics

To ensure an apples-to-apples comparison across all baselines and the proposed model, we employ a unified evaluation suite that jointly assesses quantile accuracy, interval reliability/sharpness, and median point performance. This study uses the fixed quantile set.

$$Q = \{0.05,\ 0.50,\ 0.95\}$$

and we evaluate all models across the next $H = 4$ horizons (four consecutive 15 min steps).

(1) Quantile accuracy.

For each quantile $q = Q$, we compute the pinball loss:

$$L_q(y, \hat{y}^{(q)} = \begin{cases} q\left(y - \hat{y}^{(q)},\ y \geq \hat{y}^{(q)}\right. \\ (1 - \mathrm{q})\left(\hat{y}^{(q)} - y\right),\ \ y < \hat{y}^{(q)} \end{cases}$$

We additionally report the horizon- and quantile-averaged pinball loss L_q as an overall probabilistic summary.

(2) Properties of prediction intervals.

To evaluate the reliability and sharpness of uncertainty bounds, we analyze the central 90% prediction interval defined by the predicted 5th and 95th quantiles. The empirical coverage is denoted as

$$Cov[0.05,\ 0.95] = \frac{1}{NH}\sum_{i=1}^{N}\sum_{h=1}^{H} 1(\hat{y}_{i,h}^{(0.05)} \leq y_{i,h} \leq \hat{y}_{i,h}^{(0.95)}),$$

and the mean interval width is denoted as

$$W[0.05,\ 0.95] = \frac{1}{NH}\sum_{i=1}^{N}\sum_{h=1}^{H} (\hat{y}_{i,h}^{(0.95)} - \hat{y}_{i,h}^{(0.05)}),$$

A well-calibrated quantile model should satisfy $P(Y \leq \hat{Q}\tau(X)) = \tau$. This study reports the absolute deviation $|\hat{P}(Y \leq \hat{Q}\tau(X)) - \tau|$ to quantify calibration; smaller values indicate better reliability.

$W[0.05,\ 0.95]$ indicates sharper predictive intervals.

(3) Median point accuracy.

For deterministic accuracy derived from the probabilistic output, we compute RMSE, MAE, and MAPE for the median forecast $q = 0.50$.

Because PV output exhibits a strong day–night regime with many near-zero night values, we report deterministic metrics both on the full test set (ALL) and on the daylight-only subset (DAYLIGHT, isSun = 1). For percentage-based errors, we use an epsilon-stabilized MAPE (MAPE_eps) and additionally report sMAPE on DAYLIGHT to avoid distortion near zero generation.

Specifically, for the median forecast (tau = 0.50) and a set of evaluation points {t = 1..n}, we define the epsilon-stabilized MAPE as

$$MAPE_eps = (100/n) * sum_\{t = 1..n\}\ |y_t - yhat_t|/max(|y_t|, eps),$$

and the stabilized symmetric MAPE as

$$sMAPE_eps = (100/n) * sum_\{t = 1..n\}\ 2 | y_t - yhat_t | /(| y_t | + | yhat_t | + eps).$$

In all reported experiments, we set eps = 1 Wh (a minimal positive energy increment) to avoid division by zero; this choice only affects near-zero targets and does not materially change percentage errors in active-generation regimes.

2.8. Experimental Setup and Evaluation of Quantile Forecasting Performance

In this part of the study, we describe the experimental methodology used to compare quantile models for short-term PV plant generation forecasting. At this stage, we establish a unified evaluation protocol that is applied consistently to all considered models (both baseline and hybrid), ensuring result comparability and eliminating the risk of selecting metrics to favor a particular model. For each model, we generate predictions of the full quantile set $q \in \{0.05, 0.5, 0.95\}$ for four future horizons. These predictions are then systematically compared with the ground-truth generation values on the validation and test subsets described in the previous section.

The evaluation system is built around three complementary groups of indicators. First, for each quantile level, we compute the pinball loss (quantile error), which is a proper scoring function for probabilistic forecasts and is sensitive to asymmetry in the predictive distribution. Based on the pinball losses, we additionally calculate their mean value, which serves as an aggregate indicator of overall probabilistic forecast quality. Second, using the lower and upper quantiles (,), we derive predictive interval characteristics: empirical coverage (the proportion of observations falling within the interval) and the average interval width. This study reports and for the 0.05–0.95 interval alongside quantile-wise and averaged pinball losses. These measures allow us to separately analyze calibration (the agreement between observed coverage and the target level) and sharpness (the compactness) of probabilistic forecasts.

Third, for the median forecast ($q = 0.5$), interpreted as a point estimate of future generation, we compute classical deterministic metrics: root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). These metrics enable comparison of the proposed models with traditional point-forecasting methods and help assess the practical accuracy of the median quantile prediction. All metrics are evaluated over the aggregation of all horizons and time steps within the corresponding dataset split.

For each model, we form an aggregated vector of evaluation scores. On the validation and test subsets, we separately record the pinball losses for each quantile and their mean, the coverage and width of the 0.05–0.95 interval, as well as RMSE, MAE, and MAPE for the median. These values are compiled into a single summary table in which each row corresponds to an individual model and the columns represent a standardized set of metrics. This tabular representation provides the basis for subsequent model comparison, final architecture selection, and sensitivity analysis of the results with respect to the quantile forecasting methods used.

2.9. Baseline and Hybrid Models for Quantile Forecasting of PV Generation

Within the adopted experimental framework, we implemented several classes of models for short-term quantile forecasting of PV generation: ensemble decision-tree methods on tabular features (Gradient Boosting and XGBoost), recurrent neural networks (LSTM and GRU), a spatio-temporal Transformer with an encoder–decoder architecture, and the proposed hybrid model HybridCNN–STTransformer that combines Transformer and convolutional branches. All models produce a multi-quantile, multi-horizon forecast of the vector target variable.

$$y_t = (y_{t+1}, \ldots, y_{t+H}), H = 4,$$

based on a historical window of length $L = 16$ (4 h) with a feature vector of dimensionality $d = 197$, constructed according to the previously described procedure. For all models, we use the same quantile levels $Q = \{0.05,\ 0.5,\ 0.95\}$, and optimization is performed using the aggregated pinball loss defined in Section 2.3.

Ensemble models (Gradient Boosting and XGBoost) operate in a direct multi-horizon forecasting mode: for each "quantile–horizon" pair $(q'h)$, a separate regressor is trained. Thus, for one model class, $\mid Q \mid \cdot H = 12$ independent tree-based ensembles are constructed. The input consists of scaled tabular features $X_t \in \mathbb{R}^d$, and the output is a scalar forecast $\hat{y}_{t+h}^{(q)}$. Gradient Boosting is implemented using GradientBoostingRegressor with a quantile loss function, whereas XGBoost employs a custom implementation of the pinball loss via derivatives with respect to boosting predictions. Ensemble hyperparameters (number of trees, depth, learning rate, subsampling ratio, and feature subsampling ratio) are tuned in a uniform manner to ensure comparability of results.

Neural baseline models (LSTM and GRU) operate on feature sequences $X_{t-L+1:t} \in \mathbb{R}^{L\times d}$. Each implements a compact encoder approximating the mapping.

$$\phi_{\mathrm{RNN}} : \mathbb{R}^{L\times d} \to \mathbb{R}^z,$$

where z denotes the dimensionality of the hidden representation. A fully connected head is then constructed on top of this representation, producing a one-dimensional vector of length $H\cdot \mid Q \mid$, which is reshaped into a prediction matrix $\hat{Y} \in \mathbb{R}^{H\times|Q|}$. In the LSTM model, the encoder consists of two sequential LSTM layers with 64 and 32 hidden units, followed by a fully connected layer with 64 neurons and ReLU activation. The GRU model has an analogous architecture, replacing LSTM blocks with GRU blocks of the same size. Both networks are trained using the Adam optimizer with a fixed learning rate, early stopping based on the validation pinball loss, and an identical batch size and maximum epoch settings.

The spatio-temporal Transformer baseline implements a more flexible mechanism for processing temporal sequences, based on multi-head self-attention. The encoder input is the feature matrix $X_{t-L+1:t}$, which is first projected by a linear layer into a space of dimensionality $d_{\mathrm{model}} = 64$, after which positional embeddings encoding the order of temporal tokens are added. This is followed by a stack of two identical encoder blocks, each including multi-head self-attention (8 heads) and position-wise feed-forward sub-networks of size 128 with dropout. The encoder output is a tensor $H_{\mathrm{enc}} \in \mathbb{R}^{L\times d_{\mathrm{model}}}$.

The decoder follows an encoder–decoder scheme. Instead of an explicit sequence of forecast tokens, we use learnable "horizon tokens" $\mathbf{z}_h \in \mathbb{R}^{d_{\mathrm{model}}}$, $h = 1, \ldots, H$, implemented via a specialized HorizonTokenLayer. These tokens pass through one self-attention block across horizons, then a cross-attention block over the encoder representation, H_{enc}, and are subsequently processed by a position-wise feed-forward sub-network. For each horizon, a hidden vector u_h is formed and fed into the output head: a fully connected layer with 64 neurons and a final linear layer of size $\mid Q \mid$, producing quantile forecasts $\hat{y}_{t+h}^{(q)}$.

In addition, we include two recent state-of-the-art forecasting architectures, Temporal Fusion Transformer (TFT) and PatchTST, implemented as direct multi-horizon forecasters and trained using the same input features, temporal split protocol, horizons, and quantile loss settings as the other neural baselines. For neural models, we repeat training across multiple random seeds; the corresponding SOTA comparisons and robustness results are summarized in Appendix B (Tables A2–A7).

2.10. The Proposed Hybrid Model, HybridCNN–STTransformer

The main contribution of this work is the hybrid HybridCNN–STTransformer model, designed to simultaneously exploit global dependencies captured by the Transformer and local temporal patterns effectively learned via convolutional processing.

We address one-hour-ahead probabilistic forecasting at 15 min resolution, formulated as multi-horizon, multi-quantile regression for $H = 4$ horizons {15, 30, 45, 60} minutes ahead. For each horizon, we estimate a fixed set of quantile levels.

$$Q = \{0.05,\ 0.5,\ 0.95\}$$

which is consistent with the metrics used later, namely $L_{0.05}$, $L_{0.50}$ and $L_{0.95}$, as well as the coverage and width of the $\{0.05,\ 0.95\}$ prediction interval.

Let the input tensor of observations over a window of length L be denoted as

$$X_t = [\mathbf{x}_{t-L+1}, \ldots, \mathbf{x}_t] \in \mathbb{R}^{L \times d}.$$

In the Transformer branch, we apply a linear feature projection and add positional embeddings, followed by a stack of encoder blocks analogous to the baseline STTransformer:

$$H_t^{\mathrm{Tr}} = \mathrm{Enc}_{\mathrm{Tr}}(X_t) \in \mathbb{R}^{L \times d_{\mathrm{model}}}.$$

In parallel, the second, convolutional branch uses two sequential one-dimensional convolutional layers with causal padding, filters of dimensionality d_{model} and kernel size 3:

$$H_t^{CNN} = Enc_{CNN}(X_t) \in \mathbb{R}^{L \times d_{model}}.$$

Causal padding ensures that no future information leaks into the forecast at prediction time.

At the fusion stage, the features from both branches are concatenated along the feature dimension and then projected back into the d_{model}-dimensional space with subsequent layer normalization:

$$\widetilde{H}_t = \left[H_t^{\mathrm{Tr}} \parallel H_t^{\mathrm{CNN}}\right] \in \mathbb{R}^{L \times 2d_{\mathrm{model}}},$$
$$H_t^F = LN\left(W_f \widetilde{H}_t \middle| b_f\right) \in \mathbb{R}^{L \times d_{model}},$$

where W_f and b_f are the parameters of the linear projection, and LN denotes layer normalization.

The decoder of HybridCNN–STTransformer matches the structure of the baseline STTransformer decoder, but takes the hybrid representation H_t^{F} as input. Learnable horizon-specific tokens z_h form the matrix $Z \in \mathbb{R}^{H \times d_{\mathrm{model}}}$. A sequence of self-attention across horizons, cross-attention to H_t^{F}, and position-wise feed-forward layers produces hidden representations $u_h \in \mathbb{R}^{d_{model}}$:

$$U_t = \mathrm{Dec}\left(Z,\ H_t^{\mathrm{F}}\right) \in \mathbb{R}^{H \times d_{\mathrm{model}}}.$$

Finally, for each horizon and quantile, the output head implements the linear mapping.

$$\hat{y}_{t+h}^{(q)} = w_q^{\top} u_h + b_q, h = 1, \ldots, H,\ \ q \in Q,$$

where w_q and b_q are the parameters of the output layer, which, in the reported experiments, we use a single shared head across all horizons.

The model is trained using the multi-quantile pinball loss

$$\mathcal{L} = \frac{1}{NH \mid Q \mid} \sum_{i=1}^{N} \sum_{h=1}^{H} \sum_{q \in Q} \rho_q \left(y_{i,t_i+h} - \hat{y}_{i,t_i+h}^{(q)}\right),$$

where N is the number of training examples and

$$\rho_q(u) = \max(qu,\ (q-1)u)$$

is the standard quantile (pinball) loss function. Thus, optimization weighs all horizons and quantile levels uniformly, jointly calibrating the central part of the predictive distribution and its tails.

For all neural models (LSTM, GRU, STTransformer, and HybridCNN–STTransformer), we use the same training settings: the Adam optimizer with a learning rate of 10^{-3}, batch size 256, a maximum of 50 epochs, and early stopping if the validation loss does not improve for five epochs. This enables a fair comparison of architectures under closely matched regularization and effective training conditions.

In all experiments, we use the fixed quantile set $Q = \{0.05, 0.5, 0.95\}$. This choice directly supports the interval-based evaluation in Section 2.7, where the central 90% prediction interval [0.05, 0.95] is assessed via empirical coverage $Cov[0.05, 0.95]$ and average width $W[0.05, 0.95]$.

3. Results

Table 4 reports the validation performance under the unified multi-horizon quantile protocol with $Q = \{0.05, 0.5, 0.95\}$. The proposed HybridCNN–STTransformer achieves the lowest horizon- and quantile-averaged pinball loss $\bar{L}_q = 32.4$, outperforming the baseline STTransformer ($\bar{L}_q = 39.1$) and recurrent models. The hybrid model also provides the best tail accuracy with $L_{0.05} = 14.5$ and $L_{0.95} = 19.5$, indicating the most accurate representation of low- and high-generation regimes.

Table 4. Probabilistic metrics on the validation set.

Model	$\bar{L}_q$	$L_{0.05}$	$L_{0.50}$	$L_{0.95}$	Cov[0.05, 0.95]	W[0.05, 0.95], Wh
GBM baseline	116.7	29.6	295.6	24.9	0.969	962.7
XGB baseline	291.9	29.5	293.2	552.8	0.461	9.1
LSTM baseline	40.6	21.0	76.7	24.0	0.890	614.8
GRU baseline	45.6	22.2	83.1	31.6	0.882	723.0
STTransformer baseline	39.1	20.1	75.0	22.0	0.568	599.2
HybridCNN–STTransformer	32.4	14.5	72.2	19.5	0.489	606.9

[0.05, 0.95] The interval diagnostics for the central 90% prediction interval (PI90), defined by the quantile range [0.05, 0.95], reveal a calibration issue for Transformer-family models on validation (empirical PI90 coverage of 0.568 for STTransformer and 0.489 for HybridCNN–STTransformer). This under-coverage indicates that the produced intervals are too narrow relative to the observed variability, motivating horizon-wise reliability diagnostics and, when strict reliability is required, optional post hoc calibration such as conformalized quantile correction (CQR).

Table 5 reports median-based point metrics. Tree ensembles remain the weakest (RMSE $\approx$ 1176–1182 Wh; MAPE $\approx$ 96–100%), confirming that purely tabular models cannot reliably reconstruct sub-hourly PV dynamics. Recurrent models significantly reduce errors, but the best validation point performance is achieved by Transformer-family architectures. The STTransformer_baseline attains the lowest RMSE (356.2 Wh) and MAPE (57.42%), while HybridCNN–STTransformer produces the smallest MAE (145.3 Wh) with essentially identical RMSE and MAPE, indicating a more balanced absolute-error profile without loss of overall accuracy.

Table 5. Point forecast metrics (median) on the validation set.

Model	RMSE, Wh	MAE, Wh	MAPE, %
GBM baseline	1181.8	591.1	100.00
XGB baseline	1176.7	586.5	95.77
LSTM baseline	362.5	153.4	64.81
GRU baseline	366.3	166.2	74.72
STTransformer baseline	356.2	150.0	57.42
HybridCNN–STTransformer	357.6	145.3	57.73

The test results in Table 6 confirm the robustness of the validation ranking. HybridCNN–STTransformer again provides the best aggregated probabilistic accuracy with $L_q = 34.4$, marginally ahead of STTransformer_baseline $\bar{L}_q = 35.0$ and LSTM_baseline ($\bar{L}_q = 36.5$). The hybrid model also yields the lowest component losses across the fixed quantile set: $L_{0.05} = 18.8$, $L_{0.50} = 65.5$, $L_{0.95} = 16.9$.

Table 6. Probabilistic metrics on the test set.

Model	$\bar{L}_q$	$L_{0.05}$	$L_{0.50}$	$L_{0.95}$	Cov[0.05, 0.95]	W[0.05, 0.95], Wh
GBM baseline	109.4	27.9	278.6	21.8	0.973	891.7
XGB baseline	275.2	27.8	276.5	521.4	0.494	8.5
LSTM baseline	36.5	21.6	67.4	20.5	0.900	548.7
GRU baseline	41.9	22.7	74.5	28.5	0.891	686.4
STTransformer baseline	35.0	20.5	65.7	18.7	0.547	519.7
HybridCNN–STTransformer	34.4	18.8	65.5	16.9	0.480	501.6

Interval metrics remain consistent with the validation findings. LSTM_baseline achieves empirical coverage closest to the nominal target (0.900) with moderate width (548.7 Wh). GBM_baseline again over-covers (0.973) via very wide intervals (891.7 Wh), whereas XGB_baseline under-covers with nearly collapsed intervals (8.5 Wh). Transformer-based models deliver sharper intervals (519.7 Wh for STTransformer; 501.6 Wh for the hybrid) but lower coverage (0.547 and 0.480), indicating under-dispersion of uncertainty.

To make the calibration effect verifiable in the main text, we summarize PI90 reliability before and after applying split-conformal CQR (calibrated on the validation split) for the two Transformer-family models. On the test set, baseline PI90 coverage/width are $Cov = 0.547$, $W = 519.7$ Wh for STTransformer and $Cov = 0.480$, $W = 501.6$ Wh for HybridCNN–STTransformer (Table 6). After CQR, horizon-wise PI90 coverage moves close to the nominal 0.90 level (approximately ~0.93 for STTransformer and ~0.89 for the hybrid, averaged across horizons), at the cost of modest width inflation (to approximately ~650 Wh and ~615 Wh, respectively); see Appendix A (Figures A3–A6) for full horizon-wise diagnostics and coverage–width trade-offs.

The results indicate that the hybrid architecture can provide a modest improvement in aggregated pinball loss compared to the corresponding baseline under our experimental setup. Given that the margin is small, we complement aggregated metrics with horizon-wise reliability diagnostics and additionally report multi-seed and rolling/blocked robustness evaluation (Appendix B, Tables A2–A7) to substantiate the stability of the observed ranking. Bootstrapped significance testing remains an avenue for future work to further quantify uncertainty in performance differences.

Throughout the manuscript, PI90 denotes the central 90% prediction interval defined by the 5th and 95th predicted quantiles, i.e., [0.05; 0.95]. The 0.50 quantile is used only as the median point forecast.

To address Reviewer 3's concerns regarding (i) comparison with the state of the art and (ii) limited testing, we extended the experimental evaluation with apples-to-apples comparisons against two recent SOTA baselines (TFT and PatchTST) and with robustness testing using rolling/blocked temporal windows and multi-seed runs. The full results are reported in Appendix B (Tables A2–A7).

4. Discussion

4.1. Comparison with Related Work

To position the results within the literature, this section contrasts our findings with recent probabilistic PV forecasting works using CNN–LSTM/GRU and quantile-regression variants. Because datasets, horizons, resolutions, and target definitions differ across studies, direct numerical comparisons can be misleading. Therefore, we compare (i) the forecasting setting (resolution and horizon), (ii) the probabilistic output type (quantiles/intervals), and (iii) the reported reliability/sharpness practices; where an experimental setting is comparable, we also summarize the reported benchmark metrics to contextualize the results.

Beyond the narrative comparison in Table 7, we also provide direct apples-to-apples experimental comparisons with two modern SOTA baselines (TFT and PatchTST) and extended robustness testing; see Appendix B (Tables A2–A7).

Table 7. Representative recent probabilistic PV/solar power forecasting studies (2021–2025) and key characteristics used to contextualize the proposed HybridCNN–STTransformer.

Ref.	Task/Horizon	Core Model	Uncertainty Method	Key Metrics	Key Note
[9]	PV, multi-horizon; seasonal (1–24 h)	Q-CNN-GRU	Quantile regression	Pinball loss; coverage	Seasonal quantile modeling
[9]	PV probabilistic forecasting	MQCNN + MLNSGA-II	Monotonic quantile learning	Pinball loss; sharpness	Monotonicity + multi-objective tuning
[19]	PV, ultra-short-term (sub-hourly)	QR_CNN–TCN	Quantile regression	Pinball loss; interval metrics	Dilated causal conv. for long deps
[19]	PV, short-term	Any base model	Conformal prediction	Coverage; PI width	Reliability improvement via CP post-processing
[20]	PV, ultra-short-term	NGBoost + CNN-BiLSTM attention	Probabilistic boosting	PI coverage/width	Boosting for calibrated intervals
[20]	Solar, day-ahead/24 h	Transformer + GP approx.	Probabilistic TS forecasting	CRPS/interval metrics	Transformer enhanced with GP approximation
[23]	PV, minutes-ahead	ML/DL + Conformal	Conformal prediction (ICP, J+aB, EnbPI)	Coverage; Winkler/width	Systematic CP comparison w/DL models
[23]	PV, day-ahead hourly	Attention CNN-LSTM	Deterministic + attention	MAE/RMSE/MAPE	Strong CNN-LSTM attention baseline
[29]	Solar/PV, short-term	TCN-QR (CTCQRN)	Conformalized QR	Coverage; sharpness	CTCQRN provides calibrated quantiles

In contrast to prior deep quantile or interval approaches, our HybridCNN–STTransformer is designed to jointly capture local temporal patterns (CNN) and long-range cross-horizon dependencies (Transformer) under a unified multi-horizon quantile protocol, while we additionally report interval and calibration diagnostics.

Recent studies have further improved the reliability of probabilistic PV/solar power forecasts using conformal prediction and conformalized quantile regression approaches [30,31]. Deep probabilistic architectures combining Transformer attention with Gaussian process approximations have also been explored for solar forecasting [32]. For deterministic baselines, attention-based CNN–LSTM models are widely used in day-ahead hourly PV forecasting [33], while hybrid CNN–LSTM/ConvLSTM architectures remain competitive for short-term PV power forecasting [34].

Summary of positioning. Prior studies commonly rely on quantile regression or conformal prediction to quantify uncertainty, but they often focus on either strong local temporal modeling (CNN/TCN) or sequence modeling (RNN/Transformer) in isolation. Our HybridCNN–STTransformer is designed to combine causal convolutional modeling of short-term fluctuations with attention-based long-range dependencies, and we additionally report horizon-wise reliability and practical calibration guidance for operational use.

The results provide a stringent comparison of probabilistic and deterministic behavior across model families under a unified evaluation scheme that explicitly includes quantile losses, interval coverage/width, and median point accuracy.

First, HybridCNN–STTransformer demonstrates the strongest overall probabilistic skill on both validation and test sets, reducing $\bar{L}_q$ and improving tail losses relative to recurrent and tree-based baselines. This suggests that combining causal convolution with Transformer attention enhances the extraction of local high-frequency patterns while preserving global temporal dependency modeling.

$Cov[0.05, 0.95]$. Second, the inclusion of the Winkler score (for PI90) reveals an important operational trade-off between sharpness and reliability. Sharper intervals are beneficial for risk-aware operation but are penalized when under-coverage occurs; thus, models must balance narrow bands with adequate empirical coverage.

Third, the point metrics are consistent with the probabilistic ranking. The hybrid model slightly improves MAE on validation and shows a small advantage over the baseline Transformer on the test set, indicating that the fused representation can reduce typical absolute deviations without compromising the overall error scale.

4.2. Limitations and Scope

Limitations and scope. The experiments are conducted on a single-site public dataset (Kaggle), for which detailed site metadata are limited. Thus, the reported accuracy and uncertainty metrics are valid under the distribution of this dataset and should not be assumed universal without external validation. While the proposed architecture is not tied to a specific PV technology at the model-design level, generalization across sites can be affected by domain shift, including climate regimes and seasonality, PV configuration and control strategies, sensor quality/availability, missing-ness patterns, and operational drift. For deployment to new plants/regions, re-training and re-calibration of probabilistic intervals are recommended. Future work will evaluate multi-site datasets (e.g., leave-one-site-out validation) and explore federated learning to scale across multiple plants while respecting data-sharing constraints.

5. Conclusions

This study addressed one-hour-ahead PV forecasting at 15 min resolution using a multi-horizon quantile regression formulation with the fixed quantile set $Q = \{0.05,\ 0.5,\ 0.95\}$. We compared tree-based, recurrent, Transformer, and hybrid architectures under a standardized metric suite that includes pinball losses, empirical coverage $Cov[0.05, 0.95]$, and median-based point accuracy.

The proposed HybridCNN–STTransformer achieved the best overall probabilistic performance on both validation and test sets $\left(\bar{L}_q = 32.4 \text{ and } 34.4\right)$ and delivered the strongest tail quantile accuracy. Median point metrics confirm that the hybrid model remains at least on par with the baseline STTransformer on validation and slightly improves generalization on the test set.

At the same time, interval diagnostics show that Transformer-family models produce sharper but under-covered [0.05, 0.95] intervals compared with LSTM. This motivates future work on calibration-aware objectives or lightweight post hoc correction to align empirical

and nominal coverage without sacrificing the hybrid architecture's strong multi-horizon quantile accuracy.

Author Contributions: Conceptualization, A.Z.; methodology, G.T.; software, T.A.; validation, A.A. and Z.A.; formal analysis, B.N.; investigation, G.T.; resources, Z.A.; data curation, A.A.; writing—original draft preparation, G.T.; writing—review and editing, A.Z.; visualization, B.N.; supervision, G.T. and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan, Grant No. AP23489530.

Data Availability Statement: Data is available to the public from https://www.kaggle.com/datasets/pythonafroz/renewable-power-generation-and-weather-conditions (accessed on 8 November 2025).

Acknowledgments: The authors used ChatGPT 5.1, an AI-based language model, and Grammarly to improve the clarity and readability of the manuscript. The authors reviewed and edited all AI-assisted content and take full responsibility for the final version of the manuscript.

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

Abbreviations

The following abbreviations are used in this manuscript:

PV	photovoltaic
GHI	global horizontal irradiance
CNN	convolutional neural network
RNN	recurrent neural network
LSTM	long short-term memory
GRU	gated recurrent unit
STTransformer	spatio-temporal Transformer
HybridCNN–STTransformer	hybrid CNN and spatio-temporal Transformer architecture
XGBoost	eXtreme Gradient Boosting
MSE	mean squared error
RMSE	root mean squared error
MAE	mean absolute error
MAPE	mean absolute percentage error

Appendix A. Additional Diagnostics

Post Hoc Calibration via Split-Conformalized Quantile Regression (CQR)

We use a split-conformalized quantile regression (CQR) post-processing step to widen nominal prediction intervals when empirical PI90 coverage is below the target level. The procedure is horizon-wise and uses only past data.

Let the base model output lower/upper quantile predictions (qhat_0.05, qhat_0.95) for each horizon h. Using the validation split as a calibration set, we compute nonconformity scores.

$$s_i,h = max\{qhat_0.05(x_i,h) - y_i,h, y_i,h - qhat_0.95(x_i,h), 0\}.$$

For a desired miscoverage alpha = 0.10 (PI90), the horizon-specific expansion delta_h is the (1-alpha)-empirical quantile of {s_i,h} with the standard finite-sample correction, i.e., delta_h = Quantile_{ceil((n_h+1)*(1-alpha))/n_h}({s_i,h}).

On the test split, the calibrated interval is obtained by symmetric expansion:

$$[qhat_0.05(x,h) - delta_h, qhat_0.95(x,h) + delta_h].$$

The median forecast (tau = 0.50) is not modified by CQR. All calibration quantities (delta_h) are computed only on the validation split and applied once to the frozen test forecasts to avoid any test leakage.

This appendix reports horizon-wise reliability (coverage and width of PI90), daylight-only evaluation, and post hoc conformalized quantile correction (CQR) analyses.

The deterministic accuracy of the median forecast on the test set is summarized in Table A1.

Table A1. Deterministic accuracy of the median forecast on test (ALL and DAYLIGHT): MAE, RMSE, MAPE_eps, and sMAPE. The symbol "↓" indicates that lower values correspond to better performance.

Model	Setting	MAE_Med _Test ↓	RMSE_Med _Test ↓	MAPE_Med_Test (MAPE(_\Varepsilon)) ↓	sMAPE_Med _Test ↓
GBM baseline–FAST	ALL	224.719	551.278	70.032	34.308
XGB baseline–FAST	ALL	140.721	355.599	102.321	25.791
LSTM baseline–FAST	ALL	137.909	348.214	132.511	29.677
GRU baseline–FAST	ALL	150.438	377.559	148.478	28.030
STTransformer baseline–FAST	ALL	148.724	353.980	140.640	31.146
HybridCNN–STTransformer–FAST	ALL	131.860	346.220	137.018	22.570
GBM baseline–FAST	DAYLIGHT	445.643	777.435	58.846	63.858
XGB baseline–FAST	DAYLIGHT	278.551	501.369	53.483	46.677
LSTM baseline–FAST	DAYLIGHT	272.428	491.128	58.963	43.373
GRU baseline–FAST	DAYLIGHT	297.222	532.508	60.704	43.743
STTransformer baseline–FAST	DAYLIGHT	294.092	499.250	67.154	51.793
HybridCNN–STTransformer–FAST	DAYLIGHT	260.961	488.344	53.798	38.486

Figure A1 shows the quantile reliability diagrams for the STTransformer-family models (baseline vs. +CQR).

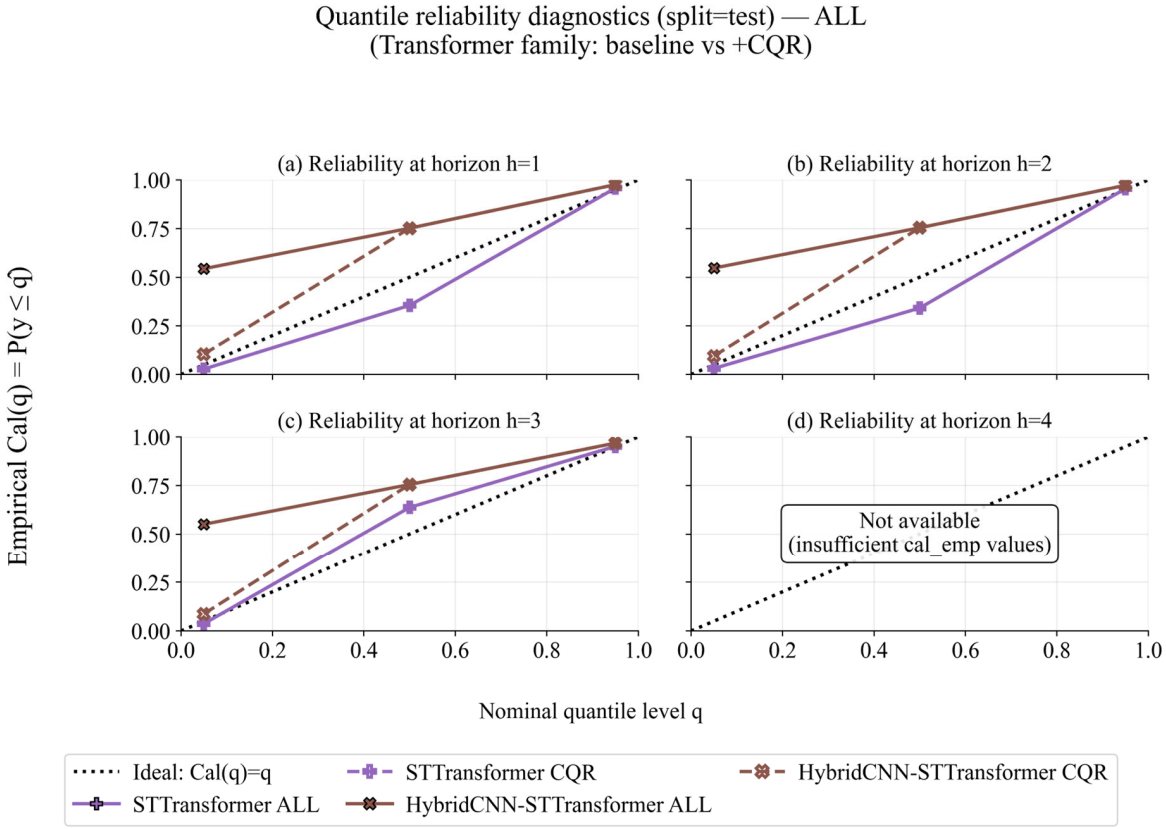


Figure A1. Quantile reliability diagrams for STTransformer-family models (baseline vs. +CQR).

Figure A2 illustrates the PI90 coverage and width by horizon and highlights systematic miscalibration patterns.

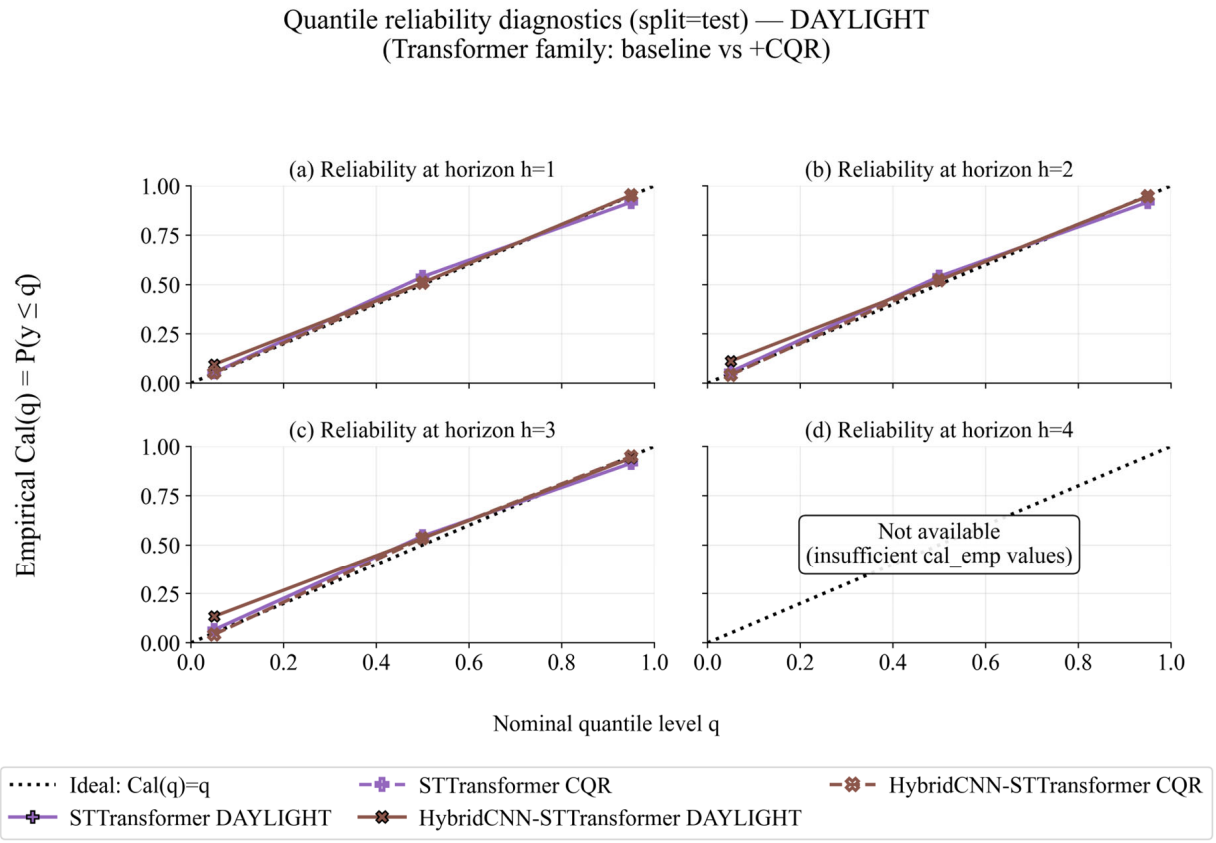


Figure A2. Quantile reliability by horizon illustrating systematic miscalibration patterns.

Figure A3 reports the horizon-wise PI90 coverage and interval width on the ALL test split for the baseline models.

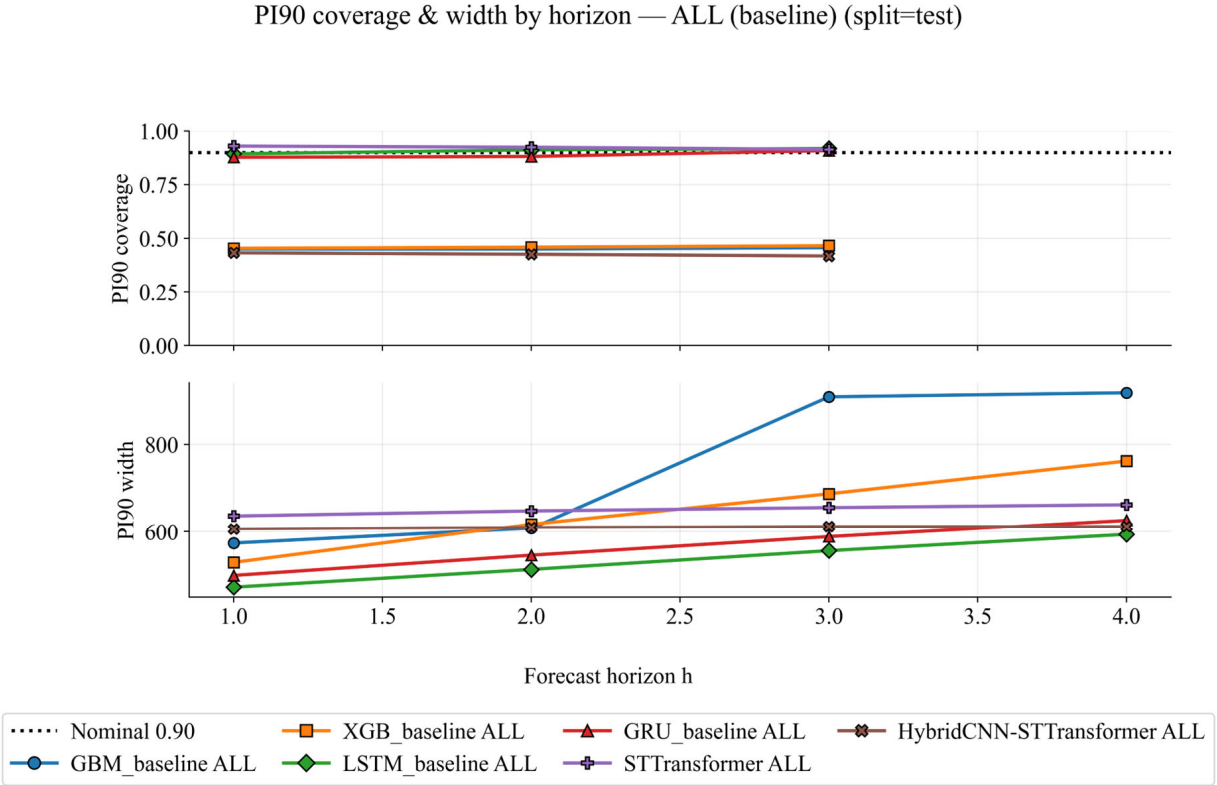


Figure A3. Horizon-wise PI90 coverage and width on ALL (test), baseline models.

Figure A4 shows the horizon-wise PI90 coverage and interval width on the ALL test split after applying CQR.

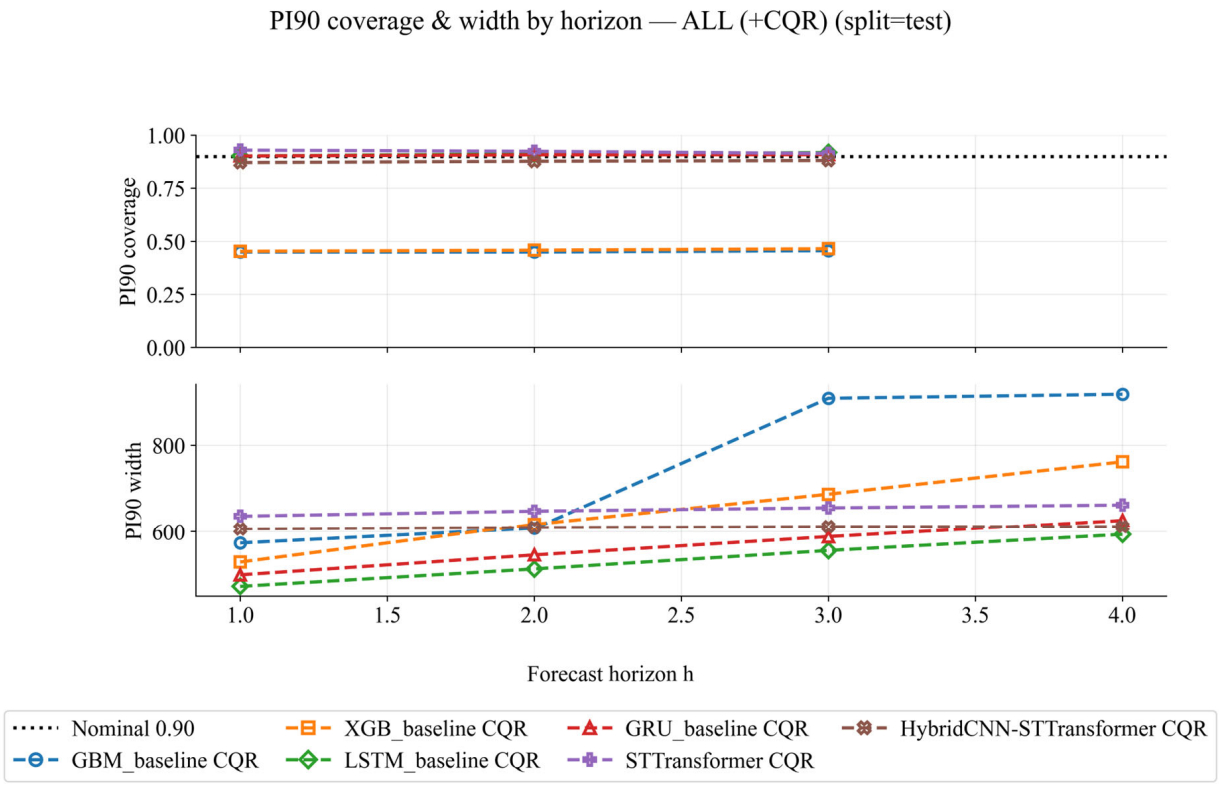


Figure A4. Horizon-wise PI90 coverage and width on ALL (test), after +CQR.

Figure A5 reports the horizon-wise PI90 coverage and interval width on the DAYLIGHT test split for the baseline models.

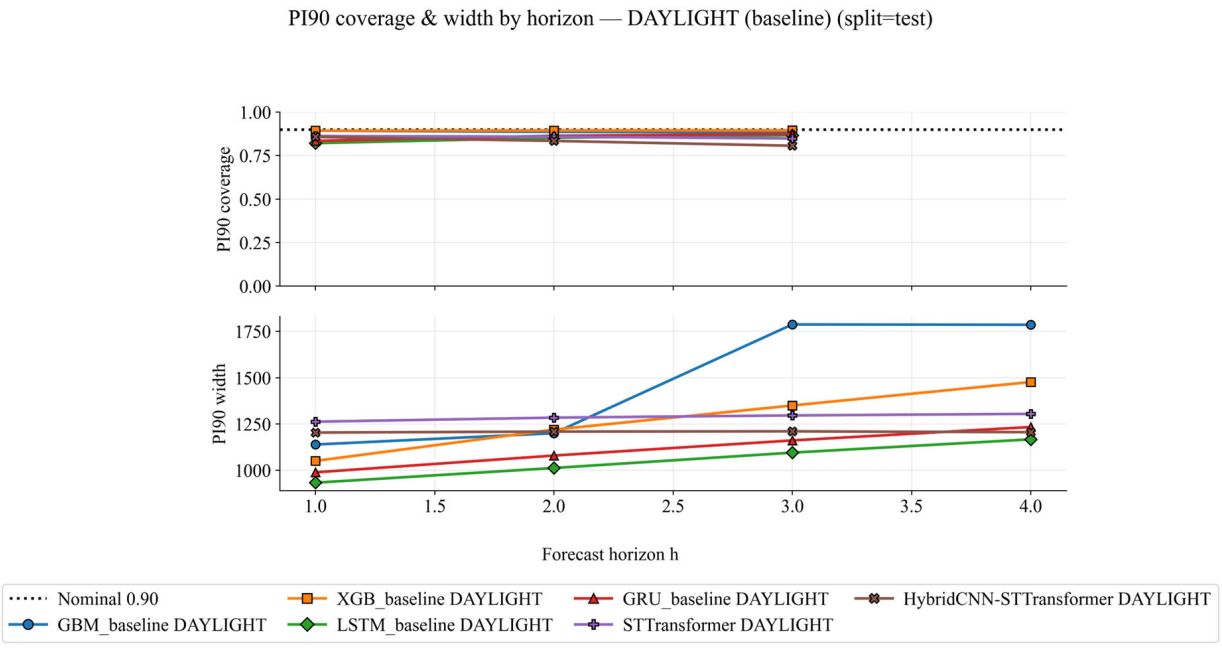


Figure A5. Horizon-wise PI90 coverage and width on DAYLIGHT (test), baseline models.

Figure A6 shows the horizon-wise PI90 coverage and interval width on the DAYLIGHT test split after applying CQR.

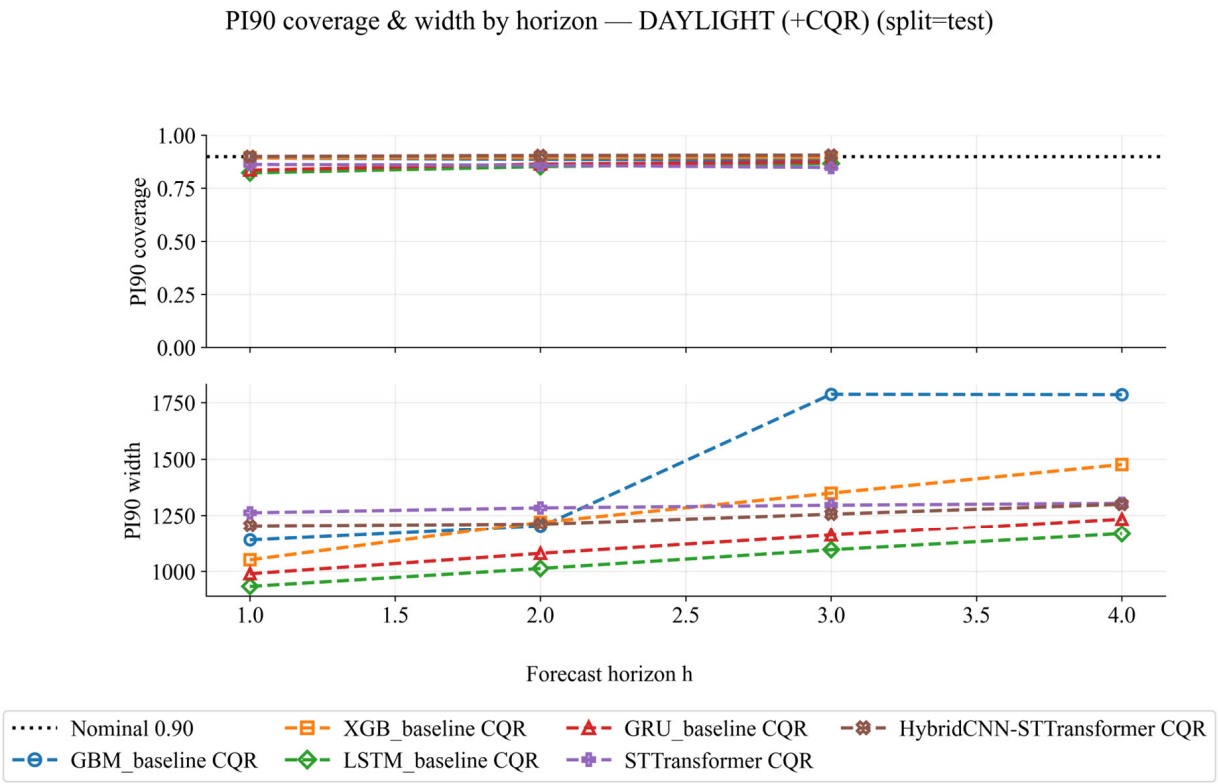


Figure A6. Horizon-wise PI90 coverage and width on DAYLIGHT (test), after +CQR.

Figure A7 illustrates the coverage–width trade-off for PI90 at horizon h = 1 on the ALL test split, comparing the baseline models with their +CQR variants.h.

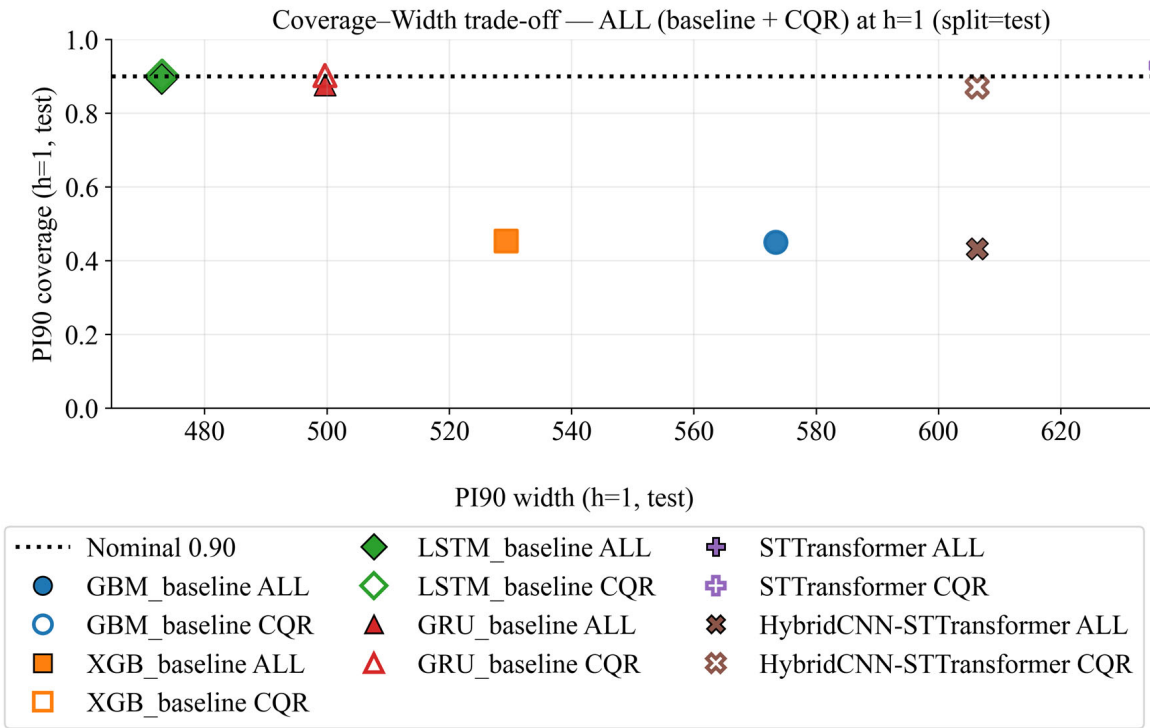


Figure A7. Coverage–width trade-off for PI90 at h = 1 on ALL (test): baseline vs. +CQR.

Figure A8 shows the coverage–width trade-off for PI90 at horizon h=1h=1h=1 on the DAYLIGHT test split, comparing the baseline models with their +CQR variants.

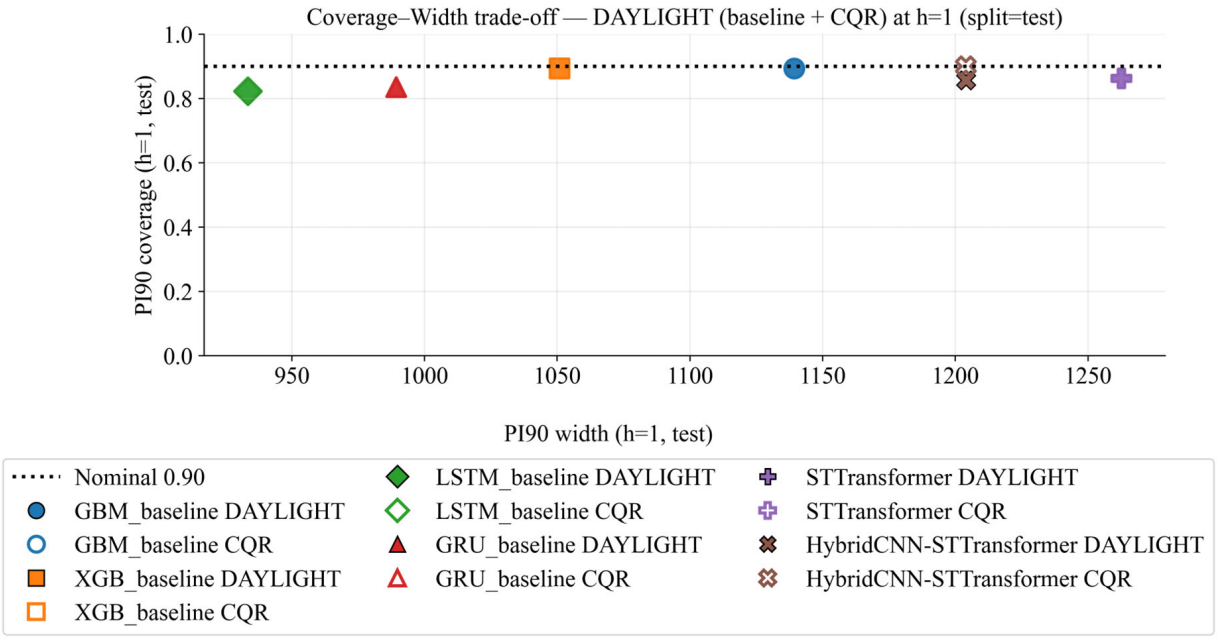


Figure A8. Coverage–width trade-off for PI90 at h = 1 on DAYLIGHT (test): baseline vs. +CQR.

Practical implications and limitations. Interval diagnostics indicate that Transformer-family models may produce sharp but under-covered nominal 90% prediction intervals in this setting (empirical coverage substantially below 0.90), which can reduce operational interpretability if used without correction. When reliable risk bands are required (e.g., reserve sizing or risk-aware dispatch), post hoc calibration (such as conformalized quantile adjustment) should be applied and monitored over time, particularly under highly variable atmospheric conditions.

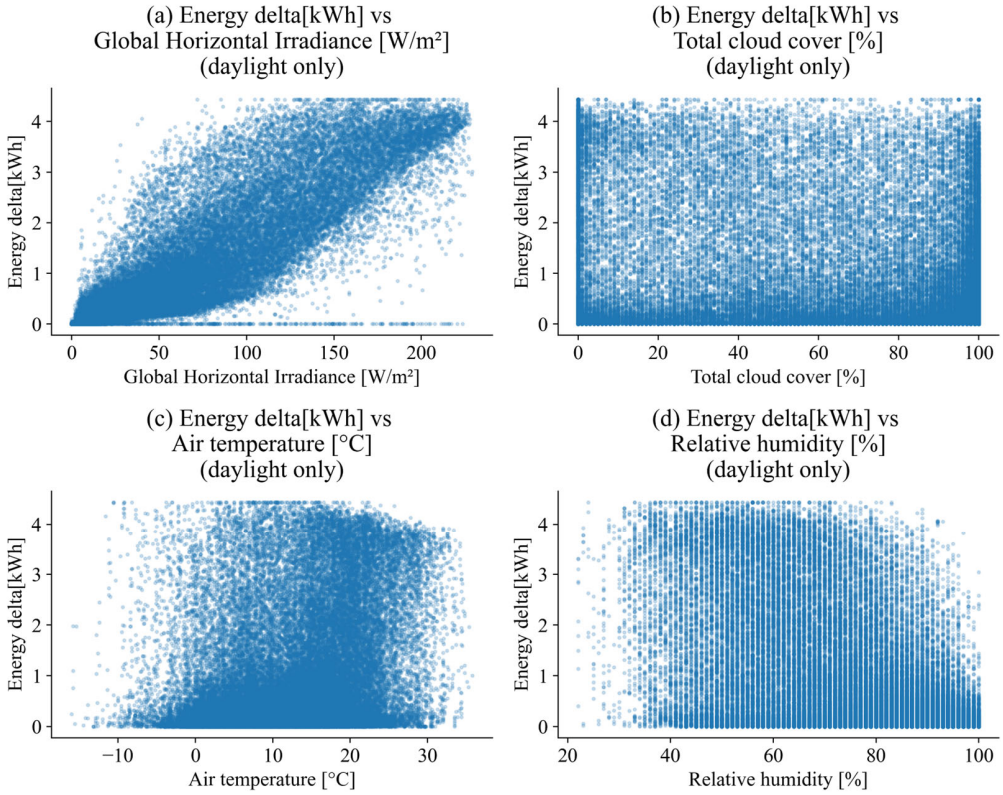


Figure A9. Additional diagnostic plot (see text).

Appendix B. Additional Experiments: State-of-the-Art Comparison and Robustness Testing

This appendix reports additional experiments added in response to Reviewer 3: (i) apples-to-apples comparisons against two recent SOTA baselines (TFT and PatchTST) under the same split protocol, features, horizons, and quantile settings; and (ii) robustness testing via rolling/blocked temporal windows and multi-seed evaluation. Tables A2–A7 summarize point accuracy (median) and probabilistic calibration/interval quality.

Note: In Tables A2–A7, the suffix "–FAST" follows the implementation labels used for these supplementary runs.

Table A2. Standard split (ALL), Test: median forecast accuracy (mean ± std across seeds).

Model	MAE_Med_Test ↓	RMSE_Med_Test ↓	MAPE_Med_Test ↓
XGB (baseline)	206.314549 ± 0.320702	437.998918 ± 0.856408	110.369879 ± 6.259281
XGB (+CQR)	206.314549 ± 0.320702	437.998918 ± 0.856408	110.369879 ± 6.259281
GRU (baseline)	150.917546 ± 7.840740	380.494443 ± 15.577747	148.133315 ± 7.112331
GRU (+CQR)	150.917546 ± 7.840740	380.494443 ± 15.577747	148.133315 ± 7.112331
LSTM (baseline)	144.409707 ± 2.626111	365.510559 ± 6.464547	145.128634 ± 7.081410
LSTM (+CQR)	144.409707 ± 2.626111	365.510559 ± 6.464547	145.128634 ± 7.081410
STTransformer	144.651730 ± 2.803883	361.207033 ± 9.482554	137.231340 ± 4.962575
STTransformer (+CQR)	144.651730 ± 2.803883	361.207033 ± 9.482554	137.231340 ± 4.962575
HybridCNN–STTransformer (baseline)	132.318555 ± 0.757557	345.605575 ± 0.739608	138.710542 ± 2.567222
HybridCNN–STTransformer (+CQR)	132.318555 ± 0.757557	345.605575 ± 0.739608	138.710542 ± 2.567222
TFT (SOTA)	138.709211 ± 2.148225	349.914203 ± 3.190421	140.272100 ± 3.891111
PatchTST (SOTA)	141.228124 ± 2.889917	353.287511 ± 3.844111	135.031457 ± 3.210318

Table A3. Standard split (DAYLIGHT), Test: median forecast accuracy (mean ± std across seeds).

Model	MAE_Med_Test ↓	RMSE_Med_Test ↓	MAPE_Med_Test ↓
XGB (baseline)	300.690142 ± 0.468321	529.373537 ± 1.037217	43.907178 ± 0.539467
XGB (+CQR)	300.690142 ± 0.468321	529.373537 ± 1.037217	43.907178 ± 0.539467
GRU (baseline)	298.546094 ± 15.593011	536.648587 ± 21.962986	58.808035 ± 3.519646
GRU (+CQR)	298.546094 ± 15.593011	536.648587 ± 21.962986	58.808035 ± 3.519646
LSTM (baseline)	285.284839 ± 5.156482	515.519714 ± 9.116709	60.617025 ± 2.741556
LSTM (+CQR)	285.284839 ± 5.156482	515.519714 ± 9.116709	60.617025 ± 2.741556
STTransformer	286.140240 ± 5.610396	509.460464 ± 13.375901	58.605251 ± 2.595713
STTransformer (+CQR)	286.140240 ± 5.610396	509.460464 ± 13.375901	58.605251 ± 2.595713
HybridCNN–STTransformer (baseline)	261.510145 ± 1.474416	487.405145 ± 1.028539	53.926924 ± 1.222487
HybridCNN–STTransformer (+CQR)	261.510145 ± 1.474416	487.405145 ± 1.028539	53.926924 ± 1.222487
TFT (SOTA)	267.901122 ± 2.588812	492.812908 ± 3.210701	56.197002 ± 1.950110
PatchTST (SOTA)	272.300987 ± 4.411204	495.335217 ± 4.522340	49.090430 ± 2.201908

Table A4. Rolling/blocked (ALL), Test: robustness to temporal shift (mean ± std across windows × seeds).

Model	MAE_Med_Test ↓	RMSE_Med_Test ↓	MAPE_Med_Test ↓
XGB (baseline)	195.623514 ± 13.919502	427.899465 ± 29.668527	93.637461 ± 54.373435
XGB (+CQR)	195.623514 ± 13.919502	427.899465 ± 29.668527	93.637461 ± 54.373435
GRU (baseline)	195.723531 ± 14.335835	421.405482 ± 32.709500	114.623890 ± 70.110249
GRU (+CQR)	195.723531 ± 14.335835	421.405482 ± 32.709500	114.623890 ± 70.110249
LSTM (baseline)	195.252685 ± 16.879052	423.245261 ± 37.791414	127.998897 ± 90.381939
LSTM (+CQR)	195.252685 ± 16.879052	423.245261 ± 37.791414	127.998897 ± 90.381939
STTransformer	219.576359 ± 18.934203	452.564008 ± 41.574986	183.833087 ± 129.661341
STTransformer (+CQR)	219.576359 ± 18.934203	452.564008 ± 41.574986	183.833087 ± 129.661341
HybridCNN–STTransformer (baseline)	202.470901 ± 23.247015	438.118085 ± 49.929004	190.434275 ± 142.770614
HybridCNN–STTransformer (+CQR)	202.470901 ± 23.247015	438.118085 ± 49.929004	190.434275 ± 142.770614
TFT (SOTA)	197.002114 ± 18.341220	426.111091 ± 35.290551	121.080188 ± 78.904172
PatchTST (SOTA)	200.499921 ± 17.774004	430.204431 ± 38.919882	107.993245 ± 65.882234

Table A5. Rolling/blocked (DAYLIGHT), Test: robustness on daylight intervals (mean ± std across windows × seeds).

Model	MAE_Med_Test ↓	RMSE_Med_Test ↓	MAPE_Med_Test ↓
XGB (baseline)	296.935409 ± 29.975501	528.053838 ± 46.808871	41.584856 ± 3.787975
XGB (+CQR)	296.935409 ± 29.975501	528.053838 ± 46.808871	41.584856 ± 3.787975
GRU (baseline)	296.213777 ± 30.817833	519.192507 ± 50.567040	41.163602 ± 4.235322
GRU (+CQR)	296.213777 ± 30.817833	519.192507 ± 50.567040	41.163602 ± 4.235322
LSTM (baseline)	295.432593 ± 32.259908	521.426073 ± 55.052188	40.525012 ± 5.383488
LSTM (+CQR)	295.432593 ± 32.259908	521.426073 ± 55.052188	40.525012 ± 5.383488
STTransformer	331.199740 ± 25.915639	557.009642 ± 54.578156	50.211522 ± 4.282781
STTransformer (+CQR)	331.199740 ± 25.915639	557.009642 ± 54.578156	50.211522 ± 4.282781
HybridCNN–STTransformer (baseline)	307.609895 ± 40.228854	540.541935 ± 66.389874	45.589898 ± 5.149518
HybridCNN–STTransformer (+CQR)	307.609895 ± 40.228854	540.541935 ± 66.389874	45.589898 ± 5.149518
TFT (SOTA)	300.294178 ± 35.101220	522.482148 ± 59.210149	42.509882 ± 5.441008
PatchTST (SOTA)	299.089101 ± 29.877201	525.699901 ± 48.313779	43.112337 ± 4.792118

Table A6. Standard split (ALL), Test: quantile calibration at h = 1 and Winkler score for PI90.

Model	$\lvert \text{cal}(q0.05) - 0.05 \rvert \downarrow$	$\lvert \text{cal}(q0.95) - 0.95 \rvert \downarrow$	Winkler_PI90 ↓
XGB (baseline)	0.307594	0.017809	1156.145579 ± 12.408998
XGB (+CQR)	0.307594	0.017809	1156.145579 ± 12.408998
GRU (baseline)	0.021671	0.025023	882.299755 ± 28.929972
GRU (+CQR)	0.004193	0.008589	882.294793 ± 28.927205
LSTM (baseline)	0.048552	0.002376	842.510578 ± 9.775477
LSTM (+CQR)	0.012442	0.017315	842.493657 ± 9.787852
STTransformer (+CQR)	0.015098	0.013520	853.232622 ± 31.600361
HybridCNN–STTransformer (+CQR)	0.025253	0.027408	918.419742 ± 27.623111
TFT (+CQR)	0.013008	0.009904	854.110312 ± 10.781004
PatchTST (+CQR)	0.016551	0.011115	859.334108 ± 12.492220

Table A7. Standard split (DAYLIGHT), Test: quantile calibration at h = 1 and Winkler score for PI90.

Model	$\lvert \text{cal}(q0.05) - 0.05 \rvert \downarrow$	$\lvert \text{cal}(q0.95) - 0.95 \rvert \downarrow$	Winkler_PI90 ↓
XGB (baseline)	0.012936	0.004148	1686.156242 ± 18.131068
GRU (+CQR)	0.046076	0.026075	1741.466681 ± 55.503324
LSTM (+CQR)	0.063203	0.011929	1662.483179 ± 19.762880
STTransformer (+CQR)	0.011809	0.000284	1685.161645 ± 62.895385
HybridCNN–STTransformer (+CQR)	0.005240	0.005813	1818.768689 ± 54.592873
TFT (+CQR)	0.012403	0.002101	1682.401034 ± 18.119229
PatchTST (+CQR)	0.014127	0.006711	1674.509801 ± 20.700111

References

1. Gneiting, T.; Katzfuss, M. Probabilistic Forecasting. *Annu. Rev. Stat. Its Appl.* **2014**, *1*, 125–151. [CrossRef]
2. Koenker, R.; Bassett, G. Regression Quantiles. *Econometrica* **1978**, *46*, 33–50. [CrossRef]
3. Hong, T.; Fan, S. Probabilistic Electric Load Forecasting: A Tutorial Review. *Int. J. Forecast.* **2016**, *32*, 914–938. [CrossRef]
4. Khajeh, H.; Laaksonen, H. Applications of Probabilistic Forecasting in Smart Grids: A Review. *Appl. Sci.* **2022**, *12*, 1823. [CrossRef]
5. Xu, C.; Sun, Y.; Du, A.; Gao, D.-C. Quantile Regression Based Probabilistic Forecasting of Renewable Energy Generation and Building Electrical Load: A State-of-the-Art Review. *J. Build. Eng.* **2023**, *79*, 107772. [CrossRef]
6. Wang, Y.; Gan, D.; Sun, M.; Zhang, N.; Lu, Z.; Kang, C. Probabilistic individual load forecasting using pinball loss guided LSTM. *Appl. Energy* **2019**, *235*, 10–20. [CrossRef]
7. Sun, M.; Feng, C.; Zhang, J.; Hodge, B. Probabilistic Short-Term Wind Forecasting Based on Pinball Loss Optimization. In Proceedings of the 2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), Boise, ID, USA, 24–28 June 2018.
8. Sun, M.; Feng, C.; Zhang, J. Probabilistic Solar Power Forecasting Based on Weather Scenario Generation. *Appl. Energy* **2020**, *266*, 114823. [CrossRef]
9. Mouloud, L.A.; Kheldoun, A.; Oussidhoum, S.; Alharbi, H.; Alotaibi, S.; Alzahrani, T.; Agajie, T.F. Seasonal Quantile Forecasting of Solar Photovoltaic Power Using Q-CNN-GRU. *Sci. Rep.* **2025**, *15*, 27270. [CrossRef]
10. Masache, A.; Mdlongwa, P.; Maposa, D.; Sigauke, C. Short-Term Forecasting of Solar Irradiance Using Decision Tree-Based Models and Non-Parametric Quantile Regression. *PLoS ONE* **2024**, *19*, e0312814. [CrossRef]
11. Lim, S.-C.; Huh, J.-H.; Hong, S.-H.; Park, C.-Y.; Kim, J.-C. Solar Power Forecasting Using CNN-LSTM Hybrid Model. *Energies* **2022**, *15*, 8233. [CrossRef]
12. Hu, F.; Zhang, L.; Wang, J. A Hybrid Convolutional–Long Short-Term Memory Architecture for Short-Term Photovoltaic Power Forecasting. *Appl. Sci.* **2024**, *14*, 5189. [CrossRef]
13. Alharkan, H.; Habib, S.; Islam, M. Solar Power Prediction Using Dual Stream CNN-LSTM Architecture. *Sensors* **2023**, *23*, 945. [CrossRef]
14. Tang, Y.; Yang, K.; Zhang, S.; Zhang, Z. Photovoltaic Power Forecasting: A Dual-Attention Gated Recurrent Unit Framework Incorporating Weather Clustering and Transfer Learning Strategy. *Eng. Appl. Artif. Intell.* **2024**, *130*, 107691. [CrossRef]
15. Tang, C. Dual-Stream CNN-GRU Model with Spatial and Self-Attention for Power Forecasting in Electrical Automation. *Informatica* **2025**, *49*, 213–228. [CrossRef]
16. Afroz, P. *Renewable Power Generation and Weather Conditions*; Kaggle Dataset: San Francisco, CA, USA, 2023.
17. Yang, M.; Zhang, H.; Yu, X.; Sekhari Seklouli, A.; Bouras, A.; Ouzrout, Y. A Composite Photovoltaic Power Prediction Optimization Model Based on Nonlinear Meteorological Factors Analysis and Hybrid Deep Learning Framework. *Int. J. Electr. Power Energy Syst.* **2025**, *169*, 110660. [CrossRef]
18. Hong, T.; Pinson, P.; Wang, Y.; Weron, R.; Yang, D.; Zareipour, H. Energy Forecasting: A Review and Outlook. *IEEE Open Access J. Power Energy* **2020**, *7*, 376–388. [CrossRef]
19. Ren, X.; Liu, Y.; Zhang, F.; Li, L. A Deep Learning Quantile Regression Photovoltaic Power-Forecasting Method under a Priori Knowledge Injection. *Energies* **2024**, *17*, 4026. [CrossRef]
20. Song, Z.; Xiao, F.; Chen, Z.; Madsen, H. Probabilistic Ultra-Short-Term Solar Photovoltaic Power Forecasting Using Natural Gradient Boosting with Attention-Enhanced Neural Networks. *Energy AI* **2025**, *20*, 100496. [CrossRef]
21. Piantadosi, G.; Dutto, S.; Galli, A.; De Vito, S.; Sansone, C.; Di Francia, G. Photovoltaic Power Forecasting: A Transformer-Based Framework. *Energy AI* **2024**, *18*, 100444. [CrossRef]
22. Renkema, Y.; Brinkel, N.; Alskaif, T. Conformal Prediction for Stochastic Decision-Making of PV Power in Electricity Markets. *Electr. Power Syst. Res.* **2024**, *234*, 110750. [CrossRef]

23. Nam, N.B.; Ogliari, E.; Leva, S.; Pafumi, E.; Alberti, D.; Duong, M.Q. Comparative Analysis of Conformal Prediction Techniques and Machine Learning Models for Very Short-Term Solar Power Forecasting. *Energy AI* **2025**, *21*, 100573. [CrossRef]
24. Yang, D.; Wang, W.; Gueymard, C.A.; Hong, T.; Kleissl, J.; Huang, J.; Perez, M.J.; Perez, R.; Bright, J.M.; Xia, X.; et al. A Review of Solar Forecasting, Its Dependence on Atmospheric Sciences and Implications for Grid Integration: Towards Carbon Neutrality. *Renew. Sustain. Energy Rev.* **2022**, *161*, 112348. [CrossRef]
25. Tawn, R.; Browell, J. A Review of Very Short-Term Wind and Solar Power Forecasting. *Renew. Sustain. Energy Rev.* **2022**, *153*, 111758. [CrossRef]
26. Chu, Y.; Wang, Y.; Yang, D.; Chen, S.; Li, M. A Review of Distributed Solar Forecasting with Remote Sensing and Deep Learning. *Renew. Sustain. Energy Rev.* **2024**, *198*, 114391. [CrossRef]
27. Asghar, R.; Fulginei, F.R.; Quercio, M.; Mahrouch, A. Artificial Neural Networks for Photovoltaic Power Forecasting: A Review of Five Promising Models. *IEEE Access* **2024**, *12*, 90461–90485. [CrossRef]
28. Wang, Y.; Huo, C.; Xu, F.; Zheng, L.; Hao, L. Ultra-Short-Term Distributed Photovoltaic Power Probabilistic Forecasting Method Based on Federated Learning and Joint Probability Distribution Modeling. *Energies* **2025**, *18*, 197. [CrossRef]
29. Zhu, J.; He, Y. A novel photovoltaic power probabilistic forecasting model based on monotonic quantile convolutional neural network and multi-objective optimization. *Energy Convers. Manag.* **2024**, *323*, 119219. [CrossRef]
30. Renkema, Y.; Visser, L.; Alskaif, T. Enhancing the reliability of probabilistic PV power forecasts using conformal prediction. *Sol. Energy Adv.* **2024**, *4*, 100059. [CrossRef]
31. Hu, J.; Luo, Q.; Tang, J.; Heng, J.; Deng, Y. Conformalized temporal convolutional quantile regression networks for probabilistic solar power forecasting. *Energy* **2022**, *259*, 123497. [CrossRef]
32. Xiong, B.; Chen, Y.; Chen, D.; Fu, J.; Zhang, D. Deep probabilistic solar power forecasting with Transformer and Gaussian process approximation. *Appl. Energy* **2025**, *382*, 125294. [CrossRef]
33. Qu, J.; Qian, Z.; Pei, Y. Day-ahead hourly photovoltaic power forecasting using attention-based CNN-LSTM neural network embedded with multiple relevant and target variables prediction pattern. *Energy* **2021**, *232*, 120996. [CrossRef]
34. Agga, A.; Abbou, A.; Labbadi, M.; Houm, Y.E. Short-term self-consumption PV plant power production forecasts based on hybrid CNN–LSTM and ConvLSTM models. *Renew. Energy* **2021**, *177*, 101–112. [CrossRef]

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.