electronics

MDPI

Article

Short-Term Solar Power Forecasting Based on Pruning-Optimized Temporal Convolutional Networks for Edge Deployment

Ziran Guo *, Wenqiang Xie, Jian Liu, Xiaolong Xiao, Xiaoxing Lu and Fangfang Zhu

State Grid Jiangsu Electric Power Co., Ltd. Research Institute, Nanjing 211103, China
* Correspondence: jsdkygzr@163.com

Abstract

Accurate multi-horizon solar power forecasting is vital for modern power distribution networks, yet standard deep learning models face limitations due to weather stochasticity and prohibitive computational requirements on resource-constrained edge devices. To address these issues, this paper proposes a Physics-Informed Similar-Day Data Augmentation (PI-SDDA) framework coupled with a Pruning-Optimized Temporal Convolutional Network (POTCN). The PI-SDDA strategy integrates multi-variable meteorological dynamics with candidate cluster size optimization to capture diurnal thermal periodicity and transient weather variations. To enable lightweight deployment, POTCN incorporates Batch Normalization scaling factors for channel-level sparsification within dilated causal residual blocks, significantly reducing parameter redundancy while preserving temporal receptive fields. Experimental results demonstrate superior forecasting performance across fifteen-minute, one-hour, and two-hour prediction horizons under standard statistical metrics compared with existing benchmarks. Edge hardware deployment tests on an embedded edge computing platform further verify low inference latency and minimal power consumption. The proposed framework provides an accurate, resource-efficient solution for real-time edge intelligence in power systems.

Keywords: solar power forecasting; temporal convolutional network; edge computing; structural pruning

1. Introduction

Driven by global decarbonization initiatives and the accelerated transition toward clean energy, the penetration of photovoltaic (PV) power in modern electrical grids has expanded significantly worldwide [1]. However, the inherently stochastic, intermittent, and weather-dependent nature of solar generation poses severe operational challenges to power grid stability, dynamic dispatch, and energy storage coordination [2,3]. High-penetration renewable integration increases the risk of supply–demand imbalances, local voltage violations, and power quality degradation, rendering high-precision PV forecasting indispensable for reliable and economical grid operations [4,5].

Among various forecast horizons, short-term PV power forecasting plays a pivotal role in intra-hour dispatching, dynamic frequency regulation, and real-time edge control. Despite its criticality, achieving high accuracy remains exceptionally challenging due to micro-scale atmospheric turbulence and rapid cloud transients, which induce severe ramp events and non-stationary power fluctuations [6,7]. Addressing these sharp dynamics requires advanced forecasting frameworks capable of capturing transient solar behavior while maintaining computational efficiency for real-time field deployment.

Check for updates

Academic Editors: Cristina Ventura and Santi Agatino Rizzo

Received: 20 September 2026
Revised: 12 October 2026
Accepted: 13 October 2026
Published: 16 October 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.

To quantify operational risks and forecast uncertainty, probabilistic methods—encompassing both prediction intervals and full probability density estimations—have been extensively investigated [8]. Prediction interval approaches typically compute error bounds through time-series variance or nearest-neighbor lookup strategies [9], whereas full density forecasts rely on parametric distributions (such as Gaussian Process Regression [10]) or non-parametric quantile estimation using machine learning variants [11]. A critical drawback of these traditional probabilistic methods lies in their dependence on predefined statistical assumptions or static historical error metrics. Because real-world PV generation under volatile atmospheric conditions exhibits highly non-stationary and non-Gaussian behavior, parametric models struggle to represent actual error distributions during cloud transients. Furthermore, non-parametric and kernel-based extensions incur high computational overhead or rely heavily on dense high-resolution datasets, limiting their capability to provide sharp deterministic predictions for real-time edge control.

In recent years, data-driven machine learning algorithms have been widely applied to ultra-short-term photovoltaic solar power forecasting. Unlike traditional physical or statistical models, which often struggle to capture the highly volatile and non-stationary dynamics of PV generation under complex atmospheric conditions, deep neural networks can effectively learn the intricate, non-linear dependencies between multi-source meteorological parameters—such as solar irradiance, ambient temperature, and wind dynamics—and actual PV power output. Long Short-Term Memory (LSTM) networks excel at modeling dynamic temporal dependencies across sequential data [12,13], while Convolutional Neural Networks (CNNs) are adept at extracting spatial features and cross-channel correlations [14,15]. To leverage both strengths, hybrid spatiotemporal architectures—such as Convolutional LSTM (ConvLSTM)—have been developed to simultaneously capture spatial correlations and temporal trends in PV generation [16]. However, despite their representative capacity, these standard deep learning models and their hybrid variants operate as purely data-driven black boxes without explicit physical domain guidance. Consequently, under rapid cloud transients and volatile atmospheric dynamics, they remain vulnerable to prediction lag, overfitting, and degraded generalization when encountering unseen meteorological conditions. Unlike deep neural networks that perform continuous representation learning, tree-based models like LightGBM partition features through discrete tabular decision splits [17,18]. However, its piecewise decision boundaries fail to capture smooth temporal dynamics, and without physical data augmentation, it degrades significantly during rapid cloud transients [19]. Integrated Graph Convolutional Networks (GCNs) with Transformers model non-Euclidean spatial topologies and long-range temporal dependencies across renewable clusters. Despite their high capacity, such deep attention-graph architectures incur heavy computational complexity, and without explicit physical clear-sky constraints or sample-level augmentation, they remain vulnerable to prediction lag during severe micro-scale cloud transients [20,21]. A multi-model stacking ensemble has been employed to enhance forecasting accuracy. However, such heavy ensemble paradigms incur excessive memory and computational overhead, posing severe bottlenecks for real-time edge deployment on resource-constrained controllers. Therefore, a physics-informed yet lightweight forecasting method is urgently needed for robust, real-time edge deployment.

To address the aforementioned limitations, a novel Physics-Informed Similar-Day Data Augmentation framework coupled with a Pruning-Optimized Temporal Convolutional Network is proposed for accurate and resource-efficient multi-horizon solar forecasting on edge computing platforms. The main contributions of this work are as follows:

- A Physics-Informed Similar-Day Data Augmentation strategy is proposed to integrate physical meteorology with optimized similar-day candidate selection, effectively

capturing diurnal thermal periodicity and transient dynamics under stochastic weather conditions while enhancing data richness and model training robustness.

- A Pruning-Optimized Temporal Convolutional Network is designed by incorporating Batch Normalization scaling factors for channel-level sparsification within dilated causal 1D residual blocks, drastically reducing parameter redundancy while delivering superior forecasting accuracy across 15 min ahead, 1 h ahead, and 2 h ahead prediction horizons.
- An edge hardware deployment experimental testbed is established to evaluate host–edge communication and real-time inference on the NVIDIA Jetson Xavier NX edge platform, verifying low computational latency and minimal power consumption for real-time edge intelligence in modern power systems.

2. Datasets and Preprocessing

2.1. Dataset Description

The empirical data utilized in this study are derived from the Solar Radiation Monitoring Laboratory (SRML) maintained by the University of Oregon, recorded at the Portland monitoring station [22]. Located in the Pacific Northwest region, the Portland site experiences a Mediterranean marine climate, characterized by strong seasonal dynamics, frequent cloud transitions, and high microclimatic variability. Observational trend analysis over multi-year records highlights two prominent physical characteristics inherent to this dataset, namely seasonal bimodality and sub-hourly transient volatility. Summer months from June to August are dominated by clear-sky conditions with high daily clearness indices and Global Horizontal Irradiance (GHI) peaking above 900 W/m^2, whereas winter months from November to February exhibit prolonged cloud coverage, lower ambient temperatures, and high Diffuse Horizontal Irradiance (DHI) ratios where DHI/GHI > 0.8. Furthermore, rapid cloud movements driven by maritime air masses cause frequent sharp power ramp events, where GHI fluctuates by more than 400 W/m^2 within 15 min intervals. The primary target variables extracted from the raw SRML database include GHI, Direct Normal Irradiance (DNI), DHI, ambient temperature (T_{amb}), wind speed (v_w) and wind direction (φ_w).

To concretely illustrate these microclimatic dynamics, the monthly observational profiles from August 2023 at the Portland station provide a representative demonstration of the spatiotemporal features across all four target variables. As shown in Figure 1, under clear-sky conditions, GHI displays smooth parabolic diurnal envelopes peaking around 1000 W/m^2. However, under broken-cloud conditions, as highlighted in the detailed view of Day 9, passing cloud masses induce intense intra-day power fluctuations, where cloud-edge scattering produces brief GHI overshoot spikes exceeding 1000 W/m^2 alongside sudden drops below 400 W/m^2. This dynamic is closely coupled with strong anti-correlation between DNI and DHI. On unclouded days, DNI maintains a stable, high plateau around 950 W/m^2 while DHI remains suppressed below 200 W/m^2. Conversely, during cloud passage events like those on Day 9, direct sunlight experiences near-total optical attenuation, causing DNI to drop abruptly to 0 W/m^2 while atmospheric scattering drives DHI up to nearly 0 W/m^2.

In contrast to these rapid radiometric transients, ambient temperature exhibits smooth diurnal thermal periodicity ranging between 12 °C and 42 °C across the month, with Day 9 demonstrating a gradual thermal trajectory peaking near 28 °C, as shown in Figure 2.

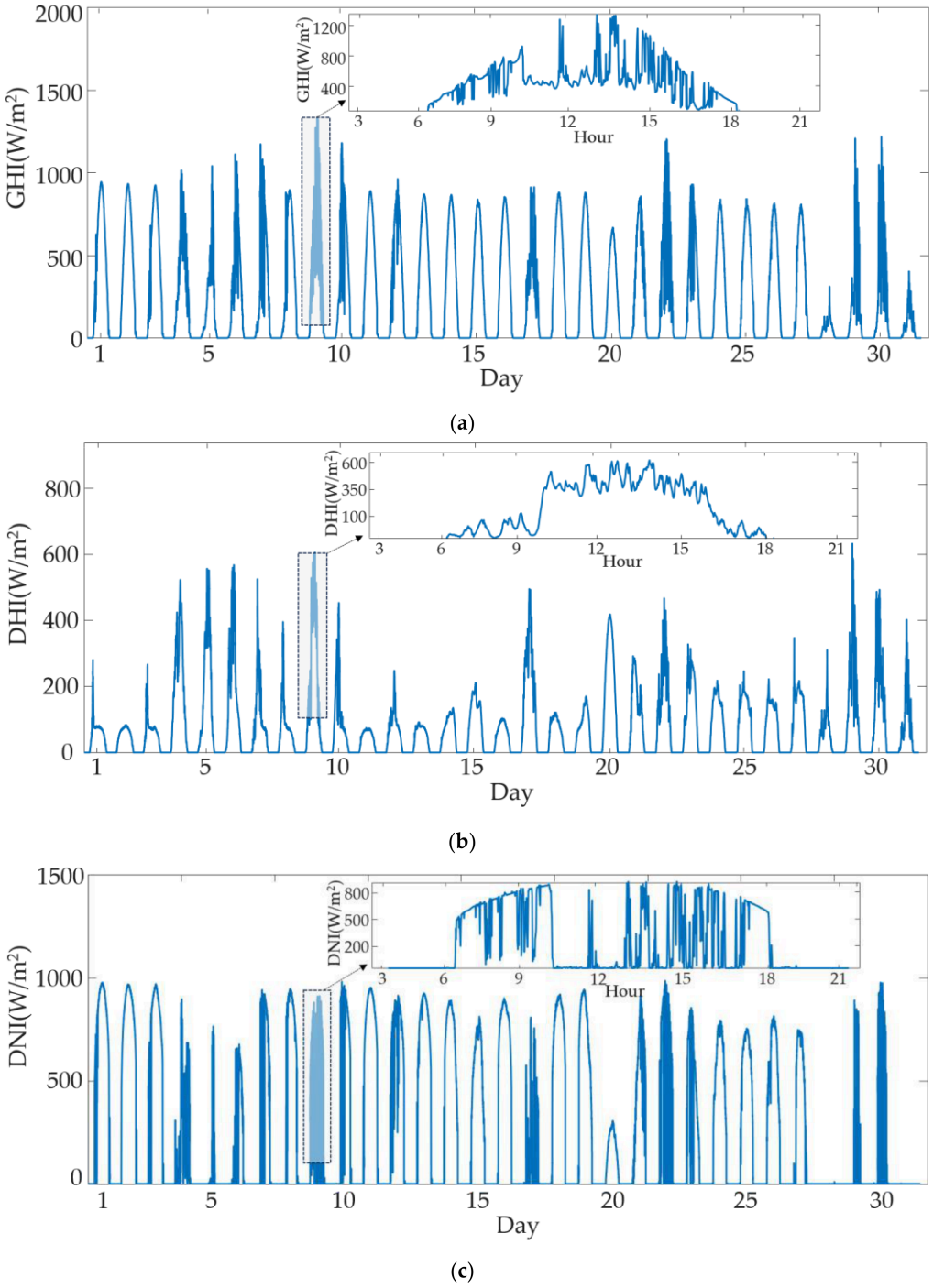


Figure 1. Monthly profiles and diurnal transient dynamics of solar irradiance parameters at the Portland monitoring station (August 2023): (**a**) GHI; (**b**) DHI; (**c**) DNI.

Meanwhile, as shown in Figure 3, v_w fluctuates continuously between 0 m/s and 6 m/s, exhibiting high-frequency stochastic turbulence superimposed on a diurnal boundary-layer convective trend, as evidenced by Day 9′s inset showing rapid velocity oscillations between 0.5 m/s and 5 m/s. Furthermore, φ_w spans 0° to 360° across the month, capturing meso-scale air mass circulation and microclimatic wind vectors. As depicted in the updated Day 9 inset, φ_w exhibits distinct phase-wise directional shifts, transitioning from smoother variations around 200–300° in the early morning (3:00–6:00) to a lower band around 100–200° (6:00–8:30), before jumping to 200–300° with rapid, dense oscillations during peak daytime hours.

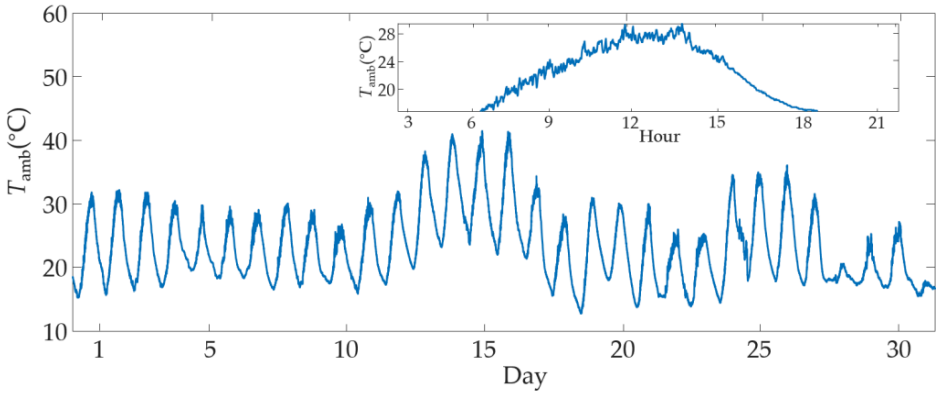


Figure 2. Monthly profile and diurnal thermal periodicity of ambient temperature (T_{amb}) at the Portland station, with inset highlighting intra-day thermal dynamics on Day 9.

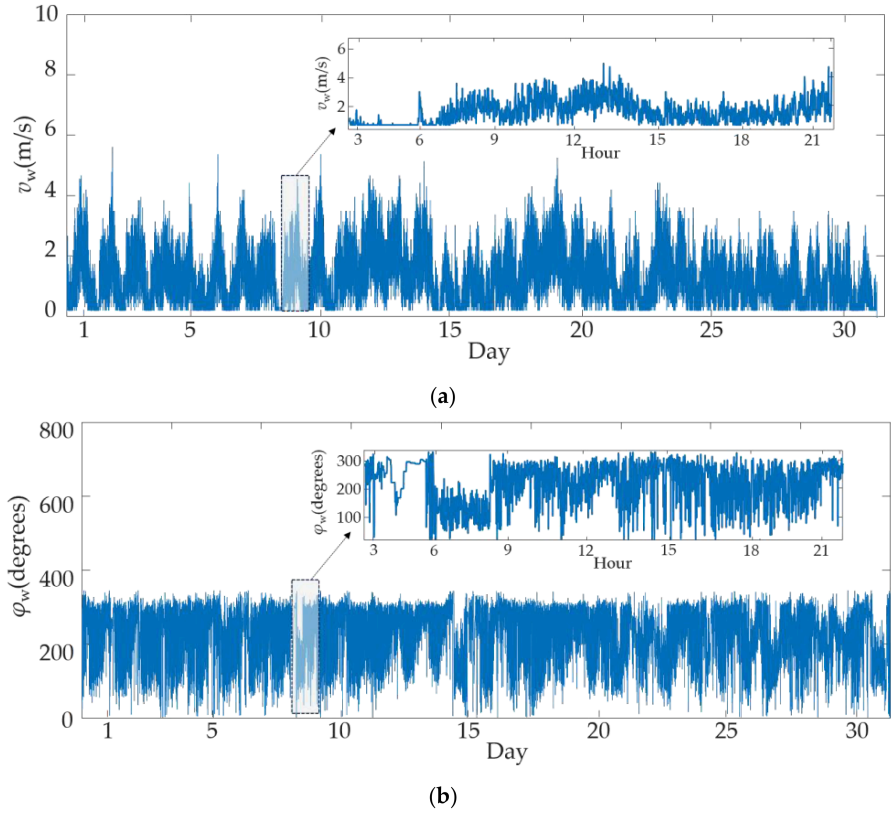


Figure 3. Monthly profiles and high-frequency stochastic dynamics of wind parameters at the Portland station: (**a**) wind speed (v_w); (**b**) wind direction (φ_w).

While GHI, DNI, and DHI determine optical energy input and T_{amb} sets the ambient thermal baseline, v_w and φ_w jointly govern forced convective heat transfer coefficients across PV module surfaces relative to array orientation, while implicitly providing spatial cues regarding cloud advection vectors. Incorporating both wind speed and directional dynamics alongside radiometric and thermal features enables the predictive framework to comprehensively capture optical transients, spatial cloud movements, and dynamic thermal dissipation mechanisms.

To ensure high data fidelity and eliminate measurement artifacts across these extracted variables prior to predictive modeling, a systematic quality control protocol is implemented. Given the strong physical interdependencies and distinct dynamic behaviors of input data, corrupted records in any single variable could propagate errors into downstream feature representations. Raw SRML records pair every physical measurement column with a

dedicated quality flag column $f(t)$. To eliminate corrupted observations caused by sensor shading, snow cover, or telemetry interruptions, raw entries are filtered according to

$$x_{clean}(t) = \begin{cases} x_{raw}(t), & \text{if } f(t) = 11 \\ \text{NaN}, & \text{otherwise} \end{cases} \quad (1)$$

where $f(t)$ = 11 designates fully validated, high-quality measurements. Short missing sequences ($\Delta t \leq 15$ min) are reconstructed using piecewise linear interpolation:

$$x(t) = x(t_1) + \frac{x(t_2) - x(t_1)}{t_2 - t_1}(t - t_1),\ t \in (t_1, t_2) \quad (2)$$

Days containing continuous corrupted segments exceeding 15 min are entirely removed from the historical dataset pool. To suppress high-frequency physical noise while maintaining micro-ramp dynamics, the cleaned 1 min raw time series are resampled into 15 min moving averages, serving as the standardized input granularity for downstream modules.

2.2. Similar-Day-Guided Data Augmentation

Directly training deep learning models on raw, multi-year temporal sequences subjects the optimization process to massive data redundancy and conflicting gradient signals from contrasting weather regimes. During stochastic gradient descent, training simultaneously on clear-sky and heavily overcast days generates orthogonal or opposing gradient vectors, which impairs convergence speed, causes model overfitting, and increases memory consumption during edge deployment. Integrating a similar-day (SD) retrieval mechanism addresses these limitations by extracting historical days that share underlying atmospheric, radiometric, and thermal states with the forecast target day, thereby focusing model capacity on localized weather patterns and minimizing gradient update interference [23,24].

However, conventional SD retrieval algorithms based on standard Euclidean distance over raw irradiance profiles suffer from severe physical and mathematical vulnerabilities. First, point-to-point Euclidean evaluation is highly sensitive to minor temporal phase shifts induced by moving cloud fronts, where a 15 min delay in cloud passage yields artificially large distance penalties despite identical wave morphologies. Second, wind direction measurements $\varphi_w(t)$ exhibit a non-Euclidean 0° to 360° circular boundary discontinuity, where slight angular oscillations across the north (e.g., 1° and 359°) produce erroneous scalar distance spikes. Third, raw GHI(t) curves conflate deterministic astronomical solar geometry with stochastic atmospheric attenuations, masking true meteorological similarities across different months. To overcome these limitations, this study proposes a Physics-Informed Similar-Day Data Augmentation (PI-SDDA) framework, which synergistically integrates a Physics-Informed Similar-Day Retrieval (PI-SDR) scheme with an adaptive physical data augmentation strategy.

To isolate stochastic atmospheric attenuations from deterministic solar trajectories, the PI-SDR scheme first converts the raw Global Horizontal Irradiance GHI(t) profile into a dimensionless clear-sky index profile $k_c(t)$:

$$k_c(t) = \frac{GHI(t)}{GHI_{cs}(t)} \quad (3)$$

where $GHI_{cs}(t)$ denotes theoretical clear-sky horizontal irradiance derived from solar zenith geometry and the Ineichen–Perez clear-sky model. By normalizing raw GHI against its astronomical upper bound, $k_c(t)$ isolates atmospheric transmissivity, effectively eliminating seasonal solar elevation bias. Concurrently, to eliminate non-Euclidean boundary disconti-

nuities inherent in directional angular measurements $\varphi_{\mathrm{w}}(t)$ and to construct a continuous velocity field, wind speed $v_{\mathrm{w}}(t)$ and wind direction $\varphi_{\mathrm{w}}(t)$ are mapped into orthogonal Cartesian wind velocity vectors:

$$u_w(t) = W_s(t) \cdot \cos\phi_w(t), \quad v_w(t) = W_s(t) \cdot \sin\phi_w(t) \tag{4}$$

This vector transformation maps scalar wind variables into a continuous Euclidean vector space, explicitly representing zonal and meridional advection components. Each historical day d_i is thus represented as a multi-variable feature vector containing high-frequency diurnal trajectories and daily scalar indicators:

$$\boldsymbol{F}_{d_i} = \left[\{k_c(t)\}_{t=1}^{N}, \{\boldsymbol{w}_v(t)\}_{t=1}^{N}, \overline{T}_{amb,d}, K_{T,d}\right] \tag{5}$$

where $\boldsymbol{w}_v(t) = [u_w(t), v_w(t)]^T$, N = 48 for a 12 h daylight window at 15 min resolution, $\overline{T}_{amb,d}$ is the daily mean ambient temperature, and $K_{\mathrm{T,d}}$ is the daily clearness index defined as

$$K_{T,d} = \frac{\sum_{t=1}^{N} \mathrm{GHI}(t)}{H_{0,d}} \tag{6}$$

where $\sum_{t=1}^{N} \mathrm{GHI}(t)$ represents the daily accumulated Global Horizontal Irradiance, and $\mathrm{H}_{0,d}$ denotes the daily extraterrestrial solar radiation on a horizontal surface, calculated based on astronomical solar geometry for the given latitude and day of the year. Subsequently, the constrained dynamic time warping (CDTW) distance between target day d_t and candidate historical day d_i is calculated as

$$D_{CDTW}(d_t, d_i) = \min_{\pi} \sum_{(k,l)\in\pi} |k_c^{d_t}(k) - k_c^{d_i}(l)|, \quad \text{s.t. } |k-l| \le w \tag{7}$$

where π represents the optimal non-linear warping path. The Sakoe–Chiba constraint window w bounds maximum temporal realignment to $\pm$w time steps, permitting elastic temporal matching caused by shifting cloud advection while preventing non-physical excessive profile distortion. To prevent look-ahead data leakage during online inference, the target-day meteorological features used as queries in the PI-SDR module—including GHI, wind statistics, temperature, and the clearness index—are not actual ground-truth observations. Instead, these variables are derived from day-ahead Numerical Weather Prediction (NWP) forecasts. The wind vector field discrepancy $D_{\mathrm{wind}}(d_t, d_{\mathrm{i}})$ between target day d_t and candidate historical day d_i captures both advection vectors and turbulent kinetic energy:

$$D_{wind}(d_t, d_i) = \left\| \begin{bmatrix} \overline{u}_w^{d_t} - \overline{u}_w^{d_i} \\ \overline{v}_w^{d_t} - \overline{v}_w^{d_i} \end{bmatrix} \right\|_2 + \left| \sigma_{u_w}^{d_t} - \sigma_{u_w}^{d_i} \right| + \left| \sigma_{v_w}^{d_t} - \sigma_{v_w}^{d_i} \right| \tag{8}$$

where $\overline{u}_w, \overline{v}_w$ and $\sigma_{u_w}, \sigma_{v_w}$ represent the daily mean velocities and standard deviations of the orthogonal wind components. Combining mean advection vectors with variance terms accounts for both macro-scale cloud transportation vectors and micro-scale turbulent fluctuations. The scalar meteorological distances for ambient temperature D_{temp} and daily clearness index D_{KT} are calculated using Z-score standardized L_1-norms:

$$D_{temp}(d_t, d_i) = \frac{|\overline{T}_{amb}^{d_t} - \overline{T}_{amb}^{d_i}|}{\sigma_T}, \quad D_{KT}(d_t, d_i) = \frac{|K_{T,d}^{d_t} - K_{T,d}^{d_i}|}{\sigma_{KT}} \tag{9}$$

where σ_T and σ_{KT} are the standard deviations across the entire historical dataset, converting physical scalar differences into dimensionless normalized quantities. The composite distance metric $D_{composite}(d_t,d_i)$ is formulated via information entropy-derived dynamic weighting:

$$D_{composite}(d_t, d_i) = w_1 \cdot \widetilde{D}_{CDTW} + w_2 \cdot \widetilde{D}_{wind} + w_3 \cdot \widetilde{D}_{temp} + w_4 \cdot \widetilde{D}_{KT} \tag{10}$$

where $\widetilde{D}$ represents min-max normalized distance components, and $w_m(m \in \{1,2,3,4\})$ are dynamically assigned feature weights based on candidate set information entropy. Utilizing information entropy objectively measures the dispersion and information density of each distance channel, automatically increasing the weight of features exhibiting higher discriminatory power.

Rather than enforcing a fixed *K*-nearest neighbor constraint that risks introducing non-similar noisy samples during anomalous weather regimes or omitting valid candidates on typical days, an Adaptive Density Radius (ADR) threshold δ_{th} is deployed to construct the optimal similar-day candidate cluster $\mathcal{S}_{SD}$:

$$\mathcal{S}_{SD} = \{d_i | D_{composite}(d_t, d_i \leq \delta_{th}\} \tag{11}$$

By dynamically bounding the feature space neighborhood according to spatial density, the ADR mechanism strictly identifies historically relevant meteorology while eliminating outlier noise during extreme or rare weather events. To prevent a combinatorial explosion during the subsequent convex interpolation phase, a maximum capacity bound, N_{sim}, is applied to the retrieved cluster. If the number of days satisfying δ_{th} exceeds N_{sim}, only the top N_{sim} candidates with the lowest composite distances are retained to form the final similar-day set $\mathcal{S}_{SD}$. Conversely, if fewer days satisfy the threshold, the entire density-filtered subset is retained. This two-stage procedure guarantees both meteorological purity and computational tractability.

Although retrieving $\mathcal{S}_{SD}$ eliminates redundant historical data, training a deep neural network solely on a constrained subset risks overfitting due to limited sample diversity. To expand the feature manifold without violating diurnal physical boundaries, the data augmentation stage of PI-SDDA is performed. Convex morphological interpolation is applied to any arbitrary pair of retrieved similar days d_a,$d_b \in \mathcal{S}_{SD}$ across all continuous feature channels $\boldsymbol{X} = [GHI, DNI, DHI, T_{amb}, u_w, v_w]^T$:

$$\boldsymbol{X}_{syn}(t) = \alpha \boldsymbol{X}_{d_a}(t) + (1-\alpha)\boldsymbol{X}_{d_b}(t) \tag{12}$$

where setting $\alpha \sim U(0.3,0.7)$ enforces balanced convex combination while preserving diurnal physical boundary constraints. Because d_{a} and d_{b} naturally satisfy solar boundary conditions (e.g., zero night-time irradiance and smooth clear-sky envelopes), their convex linear combination inherently preserves these single-variable physical bounds without introducing non-physical profile jumps or negative values. Furthermore, the physical consistency among the generated GHI, DNI, and DHI samples is implicitly inherited from the natural correlations of the actual historical parent days.

Subsequently, to mimic sub-hourly cloud dynamics and atmospheric turbulence, localized perturbation scaled by the instantaneous diffuse fraction $\eta(t)$ is injected into synthesized radiometric profiles:

$$\boldsymbol{X}_{aug}(t) = \boldsymbol{X}_{syn}(t) + \eta(t) \cdot \epsilon(t) \tag{13}$$

$$\eta(t) = \frac{DHI_{syn}(t)}{GHI_{syn}(t) + \epsilon_0} \tag{14}$$

where $\epsilon(t) \sim \mathcal{N}(0, \sigma_n^2 \boldsymbol{I})$ represents a zero-mean Gaussian noise vector, and $\epsilon_0 = 10^{-5}$ prevents numerical division by zero. Scaling perturbation amplitude by the diffuse fraction $\eta(t)$ serves as a dynamic atmospheric scattering gauge. During clear-sky periods ($\eta(t) \approx 0.1$), low diffuse radiation suppresses noise injection to retain smooth, deterministic profile envelopes. Conversely, during cloudy or broken-cloud intervals ($\eta(t) \geq 0.6$), heightened diffuse scattering scales up stochastic perturbations, enriching sample diversity specifically during high-volatility regimes to bolster model robustness against real-time solar ramp events.

To construct complete, well-defined supervised learning pairs ($\boldsymbol{X}_{aug}(t)$,$y_{aug}(t)$), the target PV power generation sequence $y_{\text{aug}}(t)$ is synthesized simultaneously by applying the identical interpolation weight α to the historical ground-truth power generation profiles of similar days d_{a} and d_{b}:

$$\boldsymbol{y}_{aug}(t) = \alpha \boldsymbol{y}_{d_a}(t) + (1 - \alpha)\boldsymbol{y}_{d_b}(t) \tag{15}$$

where $y_{d_a}(t)$ and $y_{d_b}(t)$ represent the recorded historical PV power time series of d_{a} and d_{b}, respectively. By enforcing the exact same interpolation coefficient α across both the input feature space and the output target space, the resulting augmented dataset maintains structural co-variability, providing fully specified supervised targets for model training.

To prevent potential physical violations—such as negative values, unrealistic irradiance spikes, or inconsistencies with theoretical atmospheric limits—that might otherwise be induced by subsequent stochastic perturbations, a strict post-processing constraint enforcement layer is integrated into the PI-SDDA framework. Specifically, all synthesized power generation profiles $y_{\text{aug}}(t)$ and meteorological inputs (e.g., GHI, DNI, DHI) are subjected to non-negativity clipping ($X \geq 0$) and upper-bound verification based on theoretical clear-sky models and installed capacity limits (y_{cap}). Consequently, while convex interpolation guarantees convex hull boundedness, the subsequent perturbation phase is strictly guarded by these physical boundary checks, ensuring that all augmented samples maintain realistic, bounded, and physically consistent operational profiles.

To sum up, the overall workflow of the proposed PI-SDDA framework—encompassing physics-informed feature preprocessing, dynamic entropy weighting, adaptive density retrieval, and similar-day-guided data augmentation—is illustrated in Figure 4.

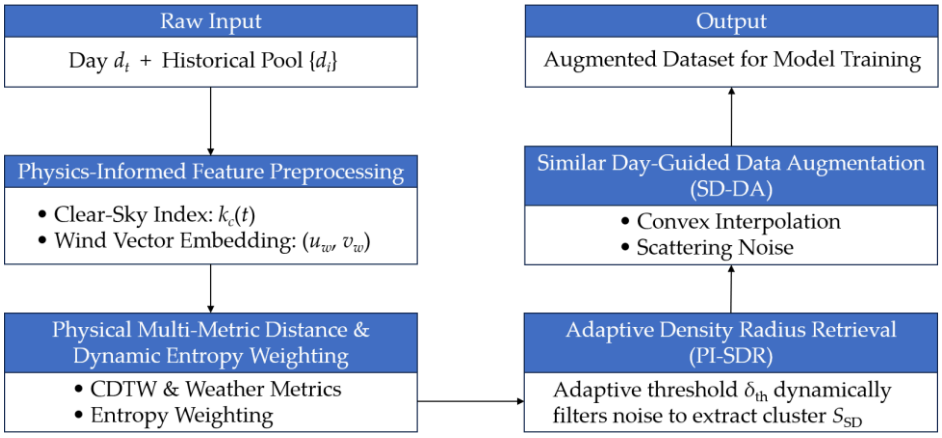


Figure 4. Overall operational workflow of the proposed PI-SDDA framework.

3. Solar Power Forecasting Based on POTCN

3.1. Temporal Convolutional Network

A Temporal Convolutional Network (TCN) is a specialized 1D fully convolutional architecture engineered for sequence-to-sequence mapping, as shown in Figure 5 [25,26]. Given an input sequence containing $\boldsymbol{T}_{\text{d}}$ historical time steps with C_{in} feature dimensions

$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{T_d}]^T \in \mathbb{R}^{T_d \times C_{in}}$, the TCN maps $\boldsymbol{X}$ to a predicted power sequence $\hat{\boldsymbol{Y}} = [\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2, \ldots, \hat{\boldsymbol{y}}_{T_d}]^T \in \mathbb{R}^{T_d \times C_{out}}$ through a sequence of non-linear transformation functions:

$$\hat{\boldsymbol{Y}} = f_{\mathrm{TCN}}(\boldsymbol{X}; \Theta) \tag{16}$$

where Θ represents the set of all trainable network parameters. In short-term solar power forecasting, historical inputs X consist of non-stationary meteorological parameters (such as Global Horizontal Irradiance, surface temperature, and relative humidity) alongside historical solar power output. Unlike traditional Recurrent Neural Networks (RNNs) or LSTMs that process temporal steps iteratively as $\boldsymbol{h}_t = g(\boldsymbol{x}_t, \boldsymbol{h}_{t-1})$, a TCN performs non-recurrent parallel transformations across the temporal axis, eliminating backpropagation-through-time (BPTT) bottlenecks and enabling highly efficient extraction of multi-variate solar features.

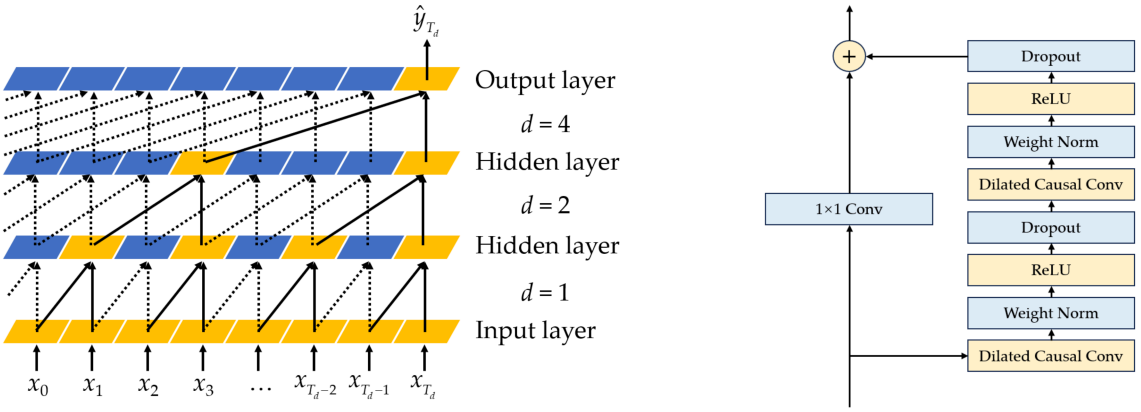


Figure 5. Architecture of the standard TCN: (**a**) hierarchical structure of dilated causal 1D convolutions across receptive fields; (**b**) micro-architecture of the TCN residual block.

To prevent information leakage from future time steps into past predictions—a fundamental requirement in real-time operational solar forecasting—the TCN enforces a strict causality constraint on its 1D convolution operations. The causal 1D convolution ensures that the output at time index t depends exclusively on elements from time t and earlier in the preceding layer. Formally, for a 1D input sequence $\boldsymbol{x} \in \mathbb{R}^T$ and a convolutional filter $f = [f_0, f_1, \ldots, f_{k-1}] \in \mathbb{R}^k$ of size k, the standard causal convolution at time step t is expressed as

$$y(t) = (f * \boldsymbol{x})(t) = \sum_{i=0}^{k-1} f_i \cdot \boldsymbol{x}_{t-i} \tag{17}$$

In actual grid-connected PV operational scenarios, power generation must be predicted strictly using real-time and historical sensor data up to time t. The causal formulation mathematically guarantees that the network respects time-arrow constraints, eliminating unphysical look-ahead bias during both training and real-time edge inference.

To capture long-range temporal dependencies without exponentially expanding filter parameters or network depth, the TCN incorporates dilated convolutions. For an input sequence $\boldsymbol{x} \in \mathbb{R}^T$ and a kernel f of size k, the dilated causal convolution operation with a dilation factor d at time step t is defined as

$$y(t) = (\boldsymbol{x} *_d f)(t) = \sum_{i=0}^{k-1} f_i \cdot \boldsymbol{x}_{t-d \cdot i} \tag{18}$$

where $d \cdot i$ specifies the spatial displacement back into historical steps. As illustrated in Figure 5a, causality is strictly preserved by ensuring that any output node at time t connects solely to nodes at time step t and earlier in preceding layers. By exponentially expanding the dilation factor at layer level $l \in \{1, 2, \ldots, L\}$ according to $d_l = 2^{l-1}$, the total receptive field RF of an L-layer TCN with a uniform kernel size k expands as

$$RF = 1 + \sum_{l=1}^{L} (k-1) \cdot d_l = 1 + (k-1) \cdot (2^L - 1) \tag{19}$$

This exponential growth of RF matches the multi-scale dynamic characteristics of PV power output. As depicted in the structural hierarchy of Figure 5a, lower layers with small dilations ($d_1 = 1$, $d_2 = 2$) capture high-frequency, localized power fluctuations caused by transient cloud cover, whereas large dilations (d_{L-1}, d_L) span broad temporal windows to capture diurnal solar elevation cycles and macro-scale weather patterns. To train deeper network structures without encountering vanishing or exploding gradients, the TCN employs residual blocks instead of simple stacked convolutional layers. The detailed micro-architecture of each residual block is shown in Figure 5b. The transformation within the l-th residual block is mathematically represented as

$$z^{(l)} = \sigma(\mathcal{H}(z^{(l-1)}) + \mathcal{F}(z^{(l-1)}, \Theta^{(l)})) \tag{20}$$

where $\mathcal{F}(\cdot)$ represents the nested dual-layer dilated causal 1D convolution operations accompanied by Weight Normalization, Spatial Dropout, and Rectified Linear Unit (ReLU) activation functions $\sigma(\cdot)$. As demonstrated in Figure 5b, the projection mapping $\mathcal{H}(z^{(l-1)})$ aligns feature dimensions as follows:

$$\mathcal{H}\left(z^{(l-1)}\right) = \begin{cases} z^{(l-1)}, & \text{if } C_{\text{in}}^{(l)} = C_{\text{out}}^{(l)} \\ W_{1\times1} \cdot z^{(l-1)}, & \text{if } C_{\text{in}}^{(l)} \neq C_{\text{out}}^{(l)} \end{cases} \tag{21}$$

where $W_{1\times1} \in \mathbb{R}^{C_{\text{out}}^{(l)} \times C_{\text{in}}^{(l)}}$ is a 1×1 convolution matrix. In multi-variable PV forecasting, non-linear weather parameters interact complexly with power output. The deep residual structure allows the network to learn abstract representations without performance degradation during backpropagation.

From a structural perspective, TCN weight tensors offer key advantages over attention-based architectures. For layer l, the trainable weight tensor of the 1D dilated convolution is structured as $W^{(l)} \in \mathbb{R}^{C_{\text{out}}^{(l)} \times C_{\text{in}}^{(l)} \times k}$. Because the TCN operates on dense, uniform 1D tensors, its computational complexity per time step remains linear with respect to sequence length:

$$\mathcal{O}_{\text{TCN}} = \mathcal{O}\left(T \cdot \sum_{l=1}^{L} C_{\text{in}}^{(l)} \cdot C_{\text{out}}^{(l)} \cdot k\right) \tag{22}$$

Unlike Transformer-based architectures whose self-attention mechanism imposes a quadratic computational footprint $\mathcal{O}(T^2 \cdot C)$, the TCN maintains a constant memory overhead during inference. This linear scaling provides a solid foundation for real-time temporal modeling.

However, deploying an uncompressed TCN directly onto resource-constrained edge devices—such as the NVIDIA Jetson—presents severe hardware-level challenges. To model long-range historical context and multi-variable meteorology, practical TCN architectures require deep residual layers ($L \geq 5$) and wide channel dimensions ($C \geq 64$). The

parameter count P_{total} and Floating Point Operations (FLOPs) across an L-layer model scale substantially:

$$P_{\text{total}} = \sum_{l=1}^{L} \left(2 \cdot k \cdot C_{\text{in}}^{(l)} \cdot C_{\text{out}}^{(l)} + C_{\text{proj}}^{(l)} \right) \tag{23}$$

$$\text{FLOPs} = 2T \cdot \sum_{l=1}^{L} \left(2 \cdot k \cdot C_{\text{in}}^{(l)} \cdot C_{\text{out}}^{(l)} + C_{\text{proj}}^{(l)} \right) \tag{24}$$

where $C_{\text{proj}}^{(l)} = C_{\text{out}}^{(l)} \cdot C_{\text{in}}^{(l)}$ represents the 1×1 projection kernel parameter when $C_{\text{in}}^{(l)} \neq C_{\text{out}}^{(l)}$, and 0 otherwise. On embedded GPU platforms, fetching dense 3D weight tensors repeatedly from off-chip DRAM to Tensor Cores causes severe memory-access bandwidth bottlenecks. Furthermore, strong cross-correlations among weather variables introduce significant channel feature redundancy within the residual blocks of Figure 5b, leading to wasted compute cycles and elevated inference latency during real-time grid dispatching.

3.2. Structured Pruning for Edge-Oriented TCN

Standard dense TCNs applied to short-term solar power forecasting often suffer from structural over-parameterization. Because time-series inputs incorporate multi-variate meteorological variables that exhibit high mutual correlation, standard 1D convolutional layers allocate excessive filter channels to redundant feature representations. To eliminate this compute and memory overhead for real-time edge inference, we propose a Pruning-Optimized Temporal Convolutional Network (POTCN) based on Batch Normalization (BN) scaling factor sparsification.

In the modified TCN architecture, a Batch Normalization layer is inserted immediately following each 1D dilated causal convolution and prior to the non-linear activation within every residual block. For a given intermediate feature channel $x_i^{(l)}$ located at layer $l \in \{1, 2, \ldots, L\}$, the mini-Batch Normalization operation is defined as

$$\hat{x}_i^{(l)} = \frac{x_i^{(l)} - \mu_B^{(l)}}{\sqrt{\left(\sigma_B^{(l)}\right)^2 + \epsilon}} \tag{25}$$

$$y_i^{(l)} = \gamma_i^{(l)} \hat{x}_i^{(l)} + \beta_i^{(l)} \tag{26}$$

where $\mu_B^{(l)}$ and $\left(\sigma_B^{(l)}\right)^2$ represent the empirical mean and variance evaluated across the mini-batch B, $\epsilon > 0$ is a small constant inserted to ensure numerical stability, and $\gamma_i^{(l)}, \beta_i^{(l)} \in \mathbb{R}$ denote the trainable affine scaling and shifting parameters, respectively. In this formulation, the absolute magnitude of the scaling parameter $|\gamma_i^{(l)}|$ directly quantifies the information flow and relative saliency of the i-th feature channel in transforming spatial–temporal representations.

To automatically identify and suppress non-essential channels during training, an L_1-norm regularization penalty on all scaling factors γ across all residual layers is integrated directly into the objective function. The joint optimization objective is formulated as

$$\mathcal{L}_{\text{total}}(\Theta, \boldsymbol{\gamma}) = \mathcal{L}_{\text{pred}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) + \lambda \sum_{l=1}^{L} \sum_{i=1}^{C^{(l)}} g\left(\gamma_i^{(l)}\right) \tag{27}$$

$$\mathcal{L}_{\text{pred}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \frac{1}{N \cdot T} \sum_{n=1}^{N} \sum_{t=1}^{T} (y_{n,t} - \hat{y}_{n,t})^2 \tag{28}$$

$$g\left(\gamma_i^{(l)}\right) = \left|\gamma_i^{(l)}\right| \tag{29}$$

where N denotes the mini-batch size, T is the temporal prediction horizon, $\lambda > 0$ serves as the sparsity regularization hyperparameter balancing prediction accuracy against structural compactness, $C^{(l)}$ represents the total number of channels in layer l, and $y_{n,t}, \hat{y}_{n,t} \in \mathbb{R}$ represent the ground-truth and forecasted normalized PV power output at time step t, respectively.

During backpropagation, because the L_1-norm penalty function $g(\cdot)$ is non-differentiable at zero, we employ subgradient descent optimization to update the channel scaling factors:

$$\gamma_i^{(l)} \leftarrow \gamma_i^{(l)} - \eta \left(\frac{\partial \mathcal{L}_{\text{pred}}}{\partial \gamma_i^{(l)}} + \lambda \cdot \text{sign}\left(\gamma_i^{(l)}\right) \right) \tag{30}$$

$$\text{sign}(\gamma) = \begin{cases} 1, & \text{if } \gamma > 0 \\ 0, & \text{if } \gamma = 0 \\ -1, & \text{if } \gamma < 0 \end{cases} \tag{31}$$

where $\eta > 0$ is the learning rate. The constant subgradient push $\lambda \cdot \text{sign}\left(\gamma_i^{(l)}\right)$ continuously forces the scaling parameters of uninformative or redundant feature channels toward zero during training, producing a sparse intermediate network representation without requiring manual filter selection.

Once sparse training converges, a global channel saliency evaluation is conducted across the entire network to determine which channels can be safely pruned. Rather than applying a local, per-layer pruning ratio that may prematurely truncate bottleneck layers, we adopt a global thresholding strategy. Given a user-specified global compression ratio $P_r \in (0, 1)$, all scaling factors across all L layers are aggregated into a single scalar set $\mathbf{\Gamma} = \left\{ \left| \gamma_i^{(l)} \right| \mid \forall l \in \{1, \ldots, L\}, i \in \left\{1, \ldots, C^{(l)}\right\} \right\}$. The global pruning threshold τ is determined as

$$\tau = \text{Percentile}(\mathbf{\Gamma}, P_r \cdot 100) \tag{32}$$

Using this unified threshold τ, a binary channel selection mask $m_i^{(l)} \in \{0, 1\}$ is generated for every channel across all convolutional layers:

$$m_i^{(l)} = \begin{cases} 1, & \text{if } \left| \gamma_i^{(l)} \right| \geq \tau \\ 0, & \text{if } \left| \gamma_i^{(l)} \right| < \tau \end{cases} \tag{33}$$

Channels evaluating to $m_i^{(l)} = 0$ are permanently removed along with their corresponding 1D convolutional weight kernels, BN parameters ($\mu_i^{(l)}, \sigma_i^{(l)}, \gamma_i^{(l)}, \beta_i^{(l)}$), and input kernel slices in the subsequent layer. Formally, let the weight tensor of the 1D dilated causal convolution at layer l be $\boldsymbol{W}^{(l)} \in \mathbb{R}^{C_{\text{out}}^{(l)} \times C_{\text{in}}^{(l)} \times k}$, where k is the kernel size. The structural pruning process performs a dual-axis tensor slicing transformation:

$$\boldsymbol{W}_{\text{pruned}}^{(l)} = \boldsymbol{W}^{(l)} \left[\mathcal{I}_{\text{out}}^{(l)}, \mathcal{I}_{\text{in}}^{(l)}, : \right] \in \mathbb{R}^{\hat{C}_{\text{out}}^{(l)} \times \hat{C}_{\text{in}}^{(l)} \times k} \tag{34}$$

where $\mathcal{I}_{\text{in}}^{(l)} = \left\{ j \mid m_j^{(l-1)} = 1 \right\}$ and $\mathcal{I}_{\text{out}}^{(l)} = \left\{ i \mid m_i^{(l)} = 1 \right\}$ represent the sets of active channel indices retained at the input and output boundaries of layer l, respectively, and $\hat{C}_{\text{in}}^{(l)} = \left| \mathcal{I}_{\text{in}}^{(l)} \right|$ and $\hat{C}_{\text{out}}^{(l)} = \left| \mathcal{I}_{\text{out}}^{(l)} \right|$ denote the compacted channel dimensions.

Short-term solar power forecasting relies on high-dimensional multi-variate time-series data, including GHI, DNI, DHI, ambient temperature, relative humidity, wind speed, and historical power telemetry. These meteorological parameters exhibit strong physical co-linearity (e.g., GHI and temperature both follow diurnal solar trajectories). Standard

dense TCNs allocate redundant convolutional channels to extract identical cross-feature trends. Sparsity-induced optimization automatically identifies and prunes these redundant collinear feature mappings without diminishing model sensitivity to sudden cloud-induced ramp events.

Unlike sequence truncation or temporal pooling—which reduce temporal resolution or shorten historical memory—channel pruning shrinks only the channel width ($C^{(l)} \rightarrow \hat{C}^{(l)}$). The fundamental temporal hyperparameters, including kernel size k, layer depth L, and exponential dilation factors $d_l = 2^{l-1}$, remain entirely untouched. Consequently, the maximum temporal receptive field RF of the network remains strictly invariant:

$$RF_{\text{POTCN}} = 1 + \sum_{l=1}^{L} (k-1) \cdot d_l = RF_{\text{standard}} \tag{35}$$

This mathematical property guarantees that POTCN retains its complete temporal memory horizon (e.g., capturing multi-hour diurnal solar curves and synoptic weather trends) while strictly enforcing temporal causality ($\hat{y}_t$ depends exclusively on historical inputs $\boldsymbol{x}_{\leq t}$), preventing future-data leakage.

To maintain gradient propagation and preserve feature identity across residual connections, the 1×1 convolution in the shortcut branch must be re-aligned synchronously with the main dual-layer convolutional path. When channel dimensions change ($\hat{y}_t$), the projection weight matrix $\boldsymbol{W}_{1\times1}^{(l)} \in \mathbb{R}^{C_{\text{out}}^{(l)} \times C_{\text{in}}^{(l)}}$ is sliced according to

$$\boldsymbol{W}_{1\times1,\text{pruned}}^{(l)} = \boldsymbol{W}_{1\times1}^{(l)} \left[\mathcal{I}_{\text{out}}^{(l)}, \mathcal{I}_{\text{in}}^{(l)} \right] \in \mathbb{R}^{\hat{C}_{\text{out}}^{(l)} \times \hat{C}_{\text{in}}^{(l)}} \tag{36}$$

This dual-path synchronization ensures that the element-wise addition operator $\boldsymbol{z}^{(l)} = \sigma\left(\boldsymbol{z}_{\text{shortcut}}^{(l)} + \boldsymbol{z}_{\text{main}}^{(l)}\right)$ remains dimensionally aligned without introducing padding latency or runtime tensor reshape operations.

Because channel pruning alters the network's operational capacity, a lightweight fine-tuning procedure is executed using the unregularized prediction loss $\mathcal{L}_{\text{pred}}$ for a small number of epochs ($E_{\text{ft}} \ll E_{\text{train}}$) to restore lost forecasting accuracy.

Following fine-tuning, the parameter count (P_{POTCN}), Floating Point Operations ($\text{FLOPs}_{\text{POTCN}}$), and Memory Access Cost (MAC_{POTCN}) per sequence step T scale as

$$P_{\text{POTCN}} = \sum_{l=1}^{L} \left(2 \cdot k \cdot \hat{C}_{\text{in}}^{(l)} \cdot \hat{C}_{\text{out}}^{(l)} + \hat{C}_{\text{proj}}^{(l)} + 2 \cdot \hat{C}_{\text{out}}^{(l)} \right) \tag{37}$$

$$\text{FLOPs}_{\text{POTCN}} = 2T \sum_{l=1}^{L} \left(2 \cdot k \cdot \hat{C}_{\text{in}}^{(l)} \cdot \hat{C}_{\text{out}}^{(l)} + \hat{C}_{\text{proj}}^{(l)} \right) \tag{38}$$

$$\text{MAC}_{\text{POTCN}} = \sum_{l=1}^{L} \left(k \cdot \hat{C}_{\text{in}}^{(l)} \cdot \hat{C}_{\text{out}}^{(l)} + T \cdot \left(\hat{C}_{\text{in}}^{(l)} + \hat{C}_{\text{out}}^{(l)} \right) \right) \tag{39}$$

where $\hat{C}_{\text{proj}}^{(l)} = \hat{C}_{\text{out}}^{(l)} \cdot \hat{C}_{\text{in}}^{(l)}$ if $\hat{C}_{\text{in}}^{(l)} \neq \hat{C}_{\text{out}}^{(l)}$, and 0 otherwise.

By eliminating zero-saliency channels, Off-Chip DRAM read operations are drastically decreased. This translates directly to lower memory bandwidth pressure and significantly reduced kernel launch overheads on edge hardware accelerators. The overall framework is shown in Figure 6.

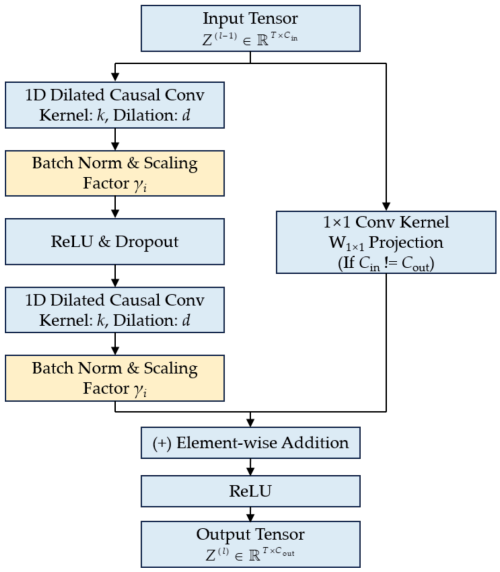


Figure 6. Detailed micro-architecture of the residual block in the POTCN, incorporating Batch Normalization scaling factors (γ_i) for channel sparsification.

4. Case Study

To evaluate the forecasting accuracy and generalizability of the proposed framework, real-world solar PV power generation and meteorological datasets collected from Portland between 1 January 2020 and 31 December 2023 are utilized in this study. Specifically, solar PV power generation serves as the target output predicted by the proposed forecasting model. The ground-truth power measurements used for model training, validation, and evaluation—together with the meteorological inputs—are obtained from direct physical measurements of a grid-connected photovoltaic system with a rated installed capacity (y_{cap}) of 5 MW, rather than theoretical irradiance-to-power conversions. Utilizing direct field power measurements preserves authentic operational behaviors, such as inverter efficiency limits and thermal curtailment, thereby validating the practical efficacy of the proposed framework.

The multi-variable sequence is partitioned chronologically according to a 3:1 macro-ratio into historical development data (2020–2022) and an independent testing set (2023). To enable hyperparameter optimization without risking test-set leakage, the historical development data is further split into a training set and a dedicated validation set (a 10% chronological holdout used specifically for tuning parameters such as N_{sim}). The independent testing set (2023) remains strictly unseen during all training and hyperparameter tuning procedures, reserved solely for the final unbiased performance evaluation. This guarantees that model training, feature extraction, and similar-day candidate retrieval are conducted strictly on historical distributions without data leakage.

4.1. Evaluation Metrics

To quantitatively evaluate the forecasting performance of the proposed model, three metrics are selected: coefficient of determination (R^2), normalized mean absolute error (E_{NMAE}), and normalized root mean square error (E_{NRMSE}). The mathematical formulas for these evaluation methods are expressed as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y'_i - y_i)^2}{\sum_{i=1}^{n} (y'_i - \overline{y}')^2} \tag{40}$$

$$E_{\mathrm{MAE}} = \frac{1}{n}\sum_{i=1}^{n}\left|y'_i - y_i\right| \tag{41}$$

$$E_{\mathrm{NMAE}} = \frac{E_{\mathrm{MAE}}}{y_{cap}} \tag{42}$$

$$E_{\mathrm{RMSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y'_i - y_i\right)^2} \tag{43}$$

$$E_{\mathrm{NRMSE}} = \frac{E_{\mathrm{RMSE}}}{y_{cap}} \tag{44}$$

where y'_i denotes the actual value at time step i, y_i represents the predicted value, $\overline{y}'$ is the mean of the actual values over the evaluation period, y_{cap} signifies the total installed capacity of the PV station, and n represents the length of the evaluation time series.

4.2. Sensitivity Analysis of Similar-Day Cluster Size

Prior to evaluating the sensitivity of specific framework components, the core experimental hyperparameters, the main architectural configurations of the proposed POTCN model, and the specific parameter settings of all evaluated benchmark architectures are summarized in Table 1 to ensure transparency and reproducibility. The network is trained using the Adam optimizer with subgradient descent for structural sparsification.

Table 1. Key experimental configurations and model parameters across all evaluated architectures.

Parameter Category	Parameter Name	Value/Configuration
Training Hyperparameters	Input Sequence Length (T_{in})	24 h
	Initial Learning Rate (η)	1×10^{-3}
	Batch Size (N)	64
	Training Epochs (E_{train})	100
	Fine-tuning Epochs (E_{ft})	50
Pruning Parameters	Sparsity Coefficient (λ)	1×10^{-4}
	Global Compression Ratio (P_{r})	50%
TCN Architecture	Number of Residual Layers (L)	4
	Kernel Size (k)	3
	Dilation Factors (d_{l})	1, 2, 4, 8
	Base Hidden Channels ($C_{\mathrm{in}}/C_{\mathrm{out}}$)	64
Benchmark Architectures	LSTM	2 hidden layers, 64 hidden units per layer
	CNN	2 1D-Conv layers (k = 3, 64 filters), ReLU, MaxPool
	CNN-LSTM	1D-Conv feature extractor + 1 LSTM layer (64 units)
	LightGBM	$N_{\mathrm{estimators}}$ = 100, learning rate = 0.05, max depth = 6
	TCN	Uncompressed baseline (L = 4, k = 3, d_l = 1, 2, 4, 8, 64 channels, 1.04 Million params)
	GCN-Transformer	2 encoder layers, 1 decoder layer, 4 heads, d_{model} = 64

Note that the pruning threshold τ is dynamically calculated during sparsification based on the target global compression ratio P_r. With the baseline network structure established, the maximum capacity bound of retrieved similar days, denoted by N_{sim}, plays a pivotal role in governing the overall forecasting performance of the proposed framework. As established by the two-stage hybrid retrieval scheme, while the ADR threshold δ_{th} dynamically filters out meteorologically inconsistent samples to guarantee data purity, N_{sim} serves as a strict upper bound on the candidate cluster size to prevent computational overruns during pairwise data augmentation. On one hand, a small N_{sim} restricts the volume and diversity of historical samples available for SD-DA, potentially hindering the deep learning model's ability to learn robust feature representations under volatile meteorological conditions. On the other hand, expanding N_{sim} excessively not only increases computational complexity during similar-day retrieval but also risks incorporating meteorologically dissimilar historical days into the candidate pool, introducing out-of-distribution noise. To resolve the trade-off between sample richness and computational efficiency during data augmentation, sensitivity experiments were conducted to determine the optimal maximum capacity bound ($N_{sim} \in \{3, 5, 10, 15, 20, 30\}$) for the final similar-day set $\mathcal{S}_{SD}$. To ensure a rigorous evaluation without data leakage, this sensitivity analysis was performed using a 10% validation holdout extracted chronologically from the 2020–2022 training dataset. The forecasting evaluation metrics under different values of N_{sim} are summarized in Table 2.

Table 2. Impact of similar-day candidate cluster size (N_{sim}) on solar power forecasting accuracy metrics.

Candidate Days (N_{sim})	R^2	E_{NMAE}	E_{NRMSE}
3	0.8964	0.0512	0.0934
5	0.9123	0.0394	0.0861
10	0.9721	0.0252	0.0549
15	0.9901	0.0141	0.0267
20	0.9903	0.0139	0.0263
30	0.9905	0.0135	0.0259

As evidenced by the quantitative results in Table 2, the forecasting accuracy exhibits a strong sensitivity to N_{sim} within the range of $3 \leq N_{sim} \leq 15$. When $N_{sim} = 3$, the model yields the lowest predictive fidelity, with $R^2 = 0.8964$, $E_{NMAE} = 0.0512$, and $E_{NRMSE} = 0.0934$, primarily due to insufficient sample diversity during data augmentation. As N_{sim} increases to 10 and 15, substantial performance gains are achieved; specifically, at $N_{sim} = 15$, R^2 surges to 0.9901, while E_{NMAE} and E_{NRMSE} decrease drastically to 0.0141 and 0.0267, representing reductions of 72.5% and 71.4% in error metrics compared to $N_{sim} = 3$. However, when N_{sim} is further expanded beyond 15, the performance improvement encounters a clear diminishing marginal return. Expanding N_{sim} from 15 to 30 yields only a negligible increase in R^2 (from 0.9901 to 0.9905) and marginal drops in E_{NMAE} (from 0.0141 to 0.0135) and E_{NRMSE} (from 0.0267 to 0.0259), while incurring higher computational overhead during multi-metric distance calculation. Considering the balance between predictive precision, data diversity, and computational efficiency, $N_{sim} = 15$ is identified as the optimal candidate cluster size and adopted for subsequent comparative evaluations.

4.3. Comparative Accuracy Analysis of Different Methods

To demonstrate the superior predictive capability and dynamic tracking performance of the proposed framework, a multi-scale comparative evaluation was conducted against six representative benchmark models: LSTM [12], CNN [14], CNN-LSTM [16], LightGBM [17],

TCN [25], and GCN-Transformer [19]. To thoroughly evaluate generalization robustness across temporal horizons, performance was systematically quantified at 15 min ahead, 1 h ahead, and 2 h ahead forecasting intervals using the coefficient of determination R^2, normalized mean absolute error E_{NMAE}, and normalized root mean square error E_{NRMSE}. Table 3 presents the quantitative comparison results, and the corresponding radar plot visualizations comparing E_{NMAE}, E_{NRMSE}, and 1-R^2 are illustrated in Figure 7.

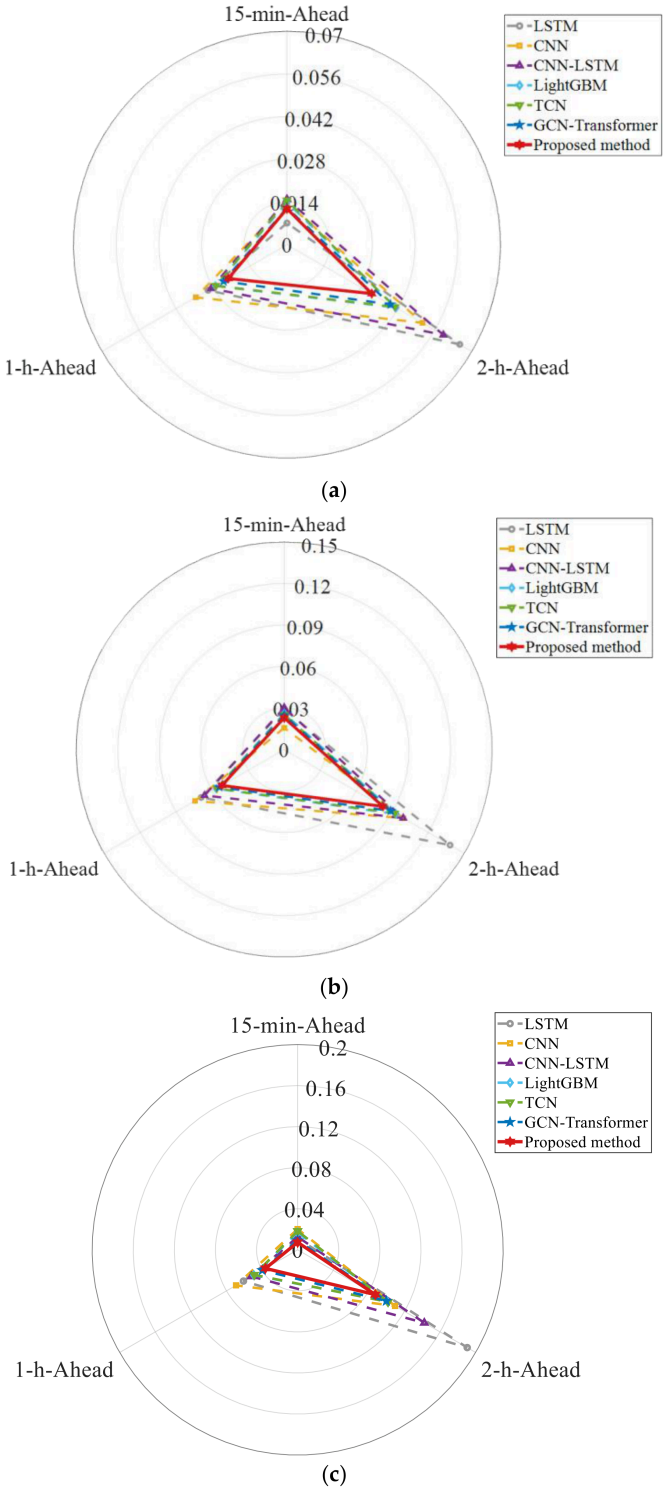


Figure 7. Performance comparison radar plots of evaluated models across multi-step forecasting horizons (15 min ahead, 1 h ahead, and 2 h ahead) under three evaluation metrics: (**a**) E_{NMAE}; (**b**) E_{NRMSE}; (**c**) 1—R^2.

Table 3. Quantitative forecasting performance comparison among various benchmark models and the proposed method across multi-step time horizons.

Models	15 min Ahead			1 h Ahead			2 h Ahead		
	R^2	E_{NMAE}	E_{NRMSE}	R^2	E_{NMAE}	E_{NRMSE}	R^2	E_{NMAE}	E_{NRMSE}
LSTM	0.9889	0.0071	0.0291	0.9391	0.0299	0.0692	0.8095	0.0654	0.1381
CNN	0.9798	0.0137	0.0157	0.9308	0.0345	0.0743	0.8906	0.0513	0.0992
CNN-LSTM	0.9874	0.0149	0.0302	0.9481	0.0287	0.0664	0.8577	0.0592	0.0991
LightGBM	0.9831	0.0139	0.0261	0.9507	0.0271	0.0566	0.8990	0.0413	0.0932
TCN	0.9817	0.0145	0.0253	0.9513	0.0269	0.0561	0.8984	0.0409	0.0921
GCN-Transformer	0.9909	0.0121	0.0248	0.9602	0.0240	0.0542	0.9004	0.0392	0.0887
Proposed method	0.9927	0.0118	0.0229	0.9638	0.0222	0.0519	0.9127	0.0321	0.0823

Because all three plotted metrics represent prediction errors or unexplained variance, a smaller enclosed polygon area contracting toward the radar center signifies higher forecasting fidelity. As depicted in Figure 7, at the 15 min forecasting horizon, the proposed framework achieves the highest R^2, while the LSTM and CNN models obtain slightly lower MAE and RMSE, respectively, indicating comparable performance among the models in this ultra-short-term scenario. However, as the forecasting horizon extends to 1 h and 2 h, the proposed method demonstrates a clearer overall advantage, outperforming the benchmark models across all evaluation metrics.

At the immediate 15 min ahead horizon, all evaluated models achieve high accuracy, maintaining R^2 values above 0.979 due to strong short-term power persistence. Nonetheless, the proposed method yields the highest overall fitting precision, attaining an R^2 of 0.9927, an E_{NMAE} of 0.0118, and an E_{NRMSE} of 0.0229. Although high-capacity baselines like GCN-Transformer closely follow with an R^2 of 0.9909, pure data-driven models lacking explicit physical constraints remain prone to localized prediction errors during sudden irradiance variations.

As the forecast horizon extends to 1 h and 2 h ahead, accumulated meteorological uncertainties and non-stationary cloud transients severely degrade the predictive accuracy of conventional data-driven baselines. At the 1 h ahead horizon, standard recurrent and convolutional models experience notable performance drops, with LSTM and CNN achieving R^2 values of only 0.9391 and 0.9308, respectively. In contrast, the proposed method maintains robust stability with an R^2 of 0.9638, while reducing E_{NMAE} to 0.0222 and E_{NRMSE} to 0.0519. This outperforms the top-performing deep baseline GCN-Transformer, which registers an R^2 of 0.9602, an E_{NMAE} of 0.0240, and an E_{NRMSE} of 0.0542.

The advantage of the proposed method becomes even more pronounced at the extended 2 h ahead horizon. Standard deep networks suffer severe error accumulation over longer look-ahead windows. For instance, the R^2 of LSTM drops sharply to 0.8095 with an E_{NRMSE} of 0.1381, while CNN-LSTM degrades to an R^2 of 0.8577. Tree-based and temporal convolution models like LightGBM and TCN exhibit better resilience with R^2 values around 0.899, yet struggle to capture continuous temporal transitions under rapid weather dynamics. High-capacity architectures like GCN-Transformer maintain competitive performance with an R^2 of 0.9004, but their heavy parameterization increases computational cost.

By contrast, the proposed method preserves high forecasting fidelity across all horizons, maintaining an R^2 of 0.9127 and achieving the lowest error rates with an E_{NMAE} of 0.0321 and an E_{NRMSE} of 0.0823 at the 2 h ahead horizon. Compared to GCN-Transformer, the proposed framework achieves relative reductions in E_{NMAE} of 2.5%, 7.5%, and 18.1% across the 15 min, 1 h, and 2 h horizons, respectively. These multi-scale results confirm that embedding physical constraints into the learning process effectively suppresses long-term

error propagation and enhances feature robustness, establishing an accurate and reliable foundation for real-time edge control.

4.4. Ablation Study and Component Contribution Analysis

To systematically isolate and evaluate the explicit quantitative contribution of each core component—namely PI-SDDA, adaptive entropy weighting, dynamic perturbation, and global structural pruning—an ablation study was conducted on the unseen test dataset. Five distinct model variants were configured and evaluated under identical experimental setups. Specifically, V1 (*w/o* PI-SDDA) completely removes the physics-informed data augmentation module to assess the baseline model relying solely on raw historical sequences. V2 (*w/o* entropy weighting) replaces the adaptive entropy-weighted similarity fusion in similar-day retrieval with a naive equal-weighting strategy. V3 (*w/o* perturbation) omits the dynamic noise perturbation factor ε during pairwise convex sample synthesis, resulting in pure linear interpolation. Furthermore, V4 (*w/o* pruning) retains the uncompressed full-capacity TCN baseline alongside the complete PI-SDDA framework to isolate the compression impact. Finally, the proposed framework integrates all designed components. The quantitative forecasting performance and model parameters across all ablation variants are summarized in Table 4.

Table 4. Quantitative evaluation of component contributions in the ablation study on the test set.

Model Variant	R^2	E_{NMAE}	E_{NRMSE}	Params
V1	0.9412	0.0385	0.0712	0.52 Million
V2	0.9785	0.0210	0.0415	0.52 Million
V3	0.9832	0.0178	0.0332	0.52 Million
V4	0.9908	0.0132	0.0252	1.04 Million
Proposed Framework	0.9901	0.0141	0.0267	0.52 Million

Comparing V1 with the proposed framework reveals that omitting PI-SDDA leads to a severe degradation in forecasting accuracy, with R^2 dropping from 0.9901 to 0.9412 and E_{NMAE} increasing by 173% (from 0.0141 to 0.0385). This demonstrates that data augmentation is the primary driver for enriching sample diversity and mitigating overfitting under volatile meteorological regimes. Furthermore, replacing adaptive entropy weighting with equal feature weighting in V2 causes a noticeable performance drop (R^2 decreases to 0.9785), confirming that dynamically assigning higher weights to dominant meteorological drivers effectively suppresses non-informative similarity noise. Removing the perturbation factor in V3 slightly degrades generalization (E_{NMAE} increases to 0.0178), proving that stochastic noise injection prevents over-reliance on rigid linear interpolation boundaries.

Finally, a comparative evaluation between the unpruned baseline (V4) and the proposed framework validates the efficacy of structural pruning. The global magnitude-based pruning strategy achieves a 50% parameter reduction (compressing the model size from 1.04 Million to 0.52 Million) with a negligible accuracy trade-off (R^2 drops by only 0.0007, and E_{NMAE} changes marginally from 0.0132 to 0.0141). Overall, these empirical results confirm that every proposed module contributes uniquely to balancing predictive fidelity, robustness, and computational efficiency, thereby fully justifying the architectural design of the proposed framework.

4.5. Hardware Resource and Computational Efficiency on Edge Platform

To evaluate the practical feasibility of deploying the proposed framework in resource-constrained edge environments, an edge hardware deployment experimental testbed was established, as illustrated in Figure 8.

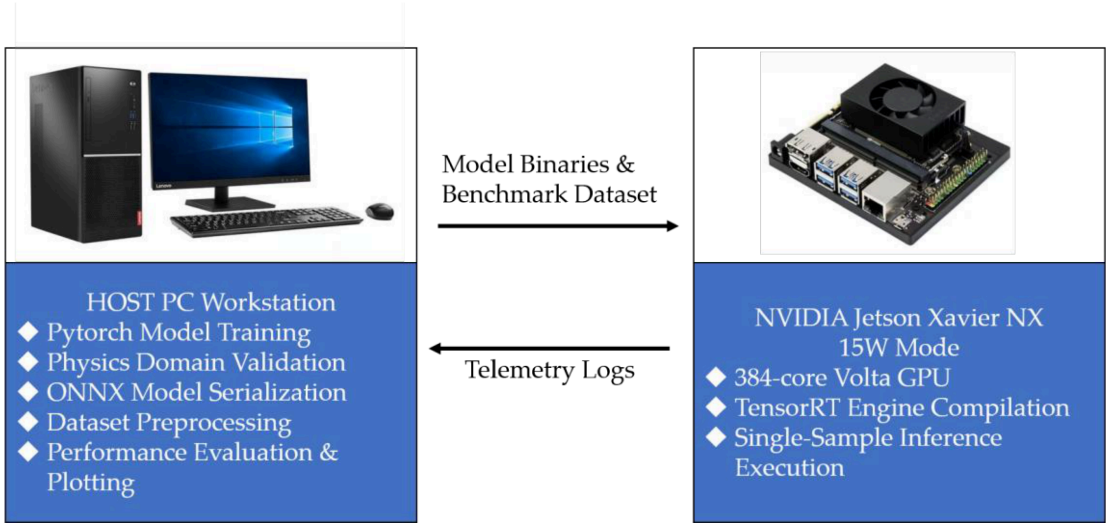


Figure 8. Schematic of the edge hardware deployment experimental testbed and host–edge communication workflow between the Host PC Workstation and the NVIDIA Jetson Xavier NX edge platform.

The host workstation executes offline PyTorch 1.9.0 model training, physics domain validation, dataset preprocessing, and ONNX model serialization. Model binaries and target benchmark datasets are transmitted via network protocol to an NVIDIA Jetson Xavier NX embedded platform. Configured in 15 W mode to utilize its 384-core Volta GPU and 6-core Carmel ARM CPU, the Jetson platform performs TensorRT engine compilation under FP16 half-precision mode and executes single-sample real-time inference with a batch size of $N = 1$ to strictly simulate online operation. To ensure a standardized comparison, all evaluated neural-network models (including LSTM, CNN, CNN-LSTM, standard TCN, GCN-Transformer, and POTCN) were tested under this identical hardware and software execution pipeline. Hardware telemetry logs detailing execution efficiency, memory footprint, and dynamic power consumption are captured via the tegrastats utility and transmitted back to the host workstation for performance evaluation and plotting. To eliminate initialization overhead, all models underwent a 100-iteration GPU warm-up prior to recording average metrics over 1000 execution cycles. Furthermore, it is emphasized that the marked reduction in model parameters and memory footprint stems exclusively from the global structural pruning procedure embedded in POTCN, whereas PI-SDDA serves solely as an input-space data augmentation module without modifying the model architecture or parameter count. Table 5 summarizes the quantitative hardware performance metrics.

Table 5. Quantitative hardware deployment metrics, computational efficiency, and energy consumption evaluation on the NVIDIA Jetson Xavier NX edge platform.

Models	Params (K)	FLOPs (M)	Latency (ms)	Peak VRAM (MB)	Avg. Power (W)	Energy/Infer (mJ)
LSTM	142.5	15.4	5.12	168	6.85	35.07
CNN	88.2	10.2	2.84	132	6.10	17.32
CNN-LSTM	226.4	24.8	7.45	215	7.42	55.28
LightGBM	-	-	1.85	45 *	4.20	7.77
TCN	112.0	11.6	3.65	148	6.32	23.07
GCN-Transformer	512.8	64.2	14.82	385	9.85	145.98
Proposed method	38.4	3.8	1.28	72	5.15	6.59

* LightGBM executes on the ARM CPU, utilizing system RAM instead of VRAM.

From the perspective of algorithmic complexity and memory footprint, the proposed framework exhibits notable lightweight characteristics. By integrating physics-informed domain constraints into feature extraction, the model drastically reduces reliance on dense parametric layers. Consequently, the proposed method requires only 38.4 k parameters and 3.8 M FLOPs, representing a reduction in model parameterization of 92.5% compared to GCN-Transformer (512.8 k parameters and 64.2 M FLOPs). In terms of hardware memory allocation, the proposed framework occupies a peak VRAM of 72 MB, making it highly compatible with embedded edge devices operating concurrent tasks. While LightGBM executes efficiently on CPU with a minimal memory footprint of 45 MB RAM, its inability to leverage GPU parallel acceleration limits its scaling performance for high-frequency signal streams.

Regarding real-time execution capability, the proposed framework achieves the fastest GPU inference latency of 1.28 ms, significantly outperforming conventional deep neural networks. Complex architectures such as CNN-LSTM and GCN-Transformer register substantial execution delays of 7.45 ms and 14.82 ms, respectively, primarily due to sequential recurrent operations and high-dimensional attention matrix computations. Furthermore, combining low latency with moderate power demand yields substantial energy savings. During the edge hardware deployment evaluation, the proposed method consumes an average total platform power (P_{avg}) of 5.15 W on the Jetson Xavier NX. The energy consumption per inference cycle (E_{inf}) is calculated as the product of the total platform power and the average inference time (T_{inf}), expressed as $E_{\text{inf}} = P_{\text{avg}} \times T_{\text{inf}}$. Given an average inference time of approximately 1.28 ms, this translates to a low energy consumption of 6.59 mJ per inference cycle (5.15 W $\times$ 1.28 ms = 6.59 mJ). Compared to the GCN-Transformer (145.98 mJ per inference), the proposed framework reduces single-inference energy consumption by 95.5%. These benchmark evaluations confirm that the proposed model effectively balances predictive accuracy with computational efficiency, providing a practical solution for real-time edge monitoring and applications.

5. Conclusions

This paper proposes a Physics-Informed Similar-Day Data Augmentation (PI-SDDA) framework integrated with a Pruning-Optimized Temporal Convolutional Network (POTCN) for accurate and lightweight multi-horizon solar forecasting on resource-constrained edge devices. By incorporating physical meteorological parameters alongside cluster size optimization, the PI-SDDA strategy effectively enriches training samples and captures intra-day thermal periodicity and transient dynamics under fluctuating weather conditions. Furthermore, the POTCN architecture leverages channel-wise Batch Normalization scaling factor sparsification within dilated causal residual blocks, achieving substantial model compression while retaining critical temporal receptive fields.

The experimental results demonstrate that the proposed framework performs comparably to baseline models at the short 15 min horizon, and exhibits a clear overall advantage at extended forecasting horizons (1 h and 2 h). Furthermore, edge hardware deployment testing on an embedded edge computing platform confirms low computational latency and minimal power consumption, validating the practical viability of the proposed scheme for real-time edge intelligence in power systems.

Future work will focus on extending the proposed physics-informed augmentation and lightweight pruning framework to spatiotemporal forecasting across distributed active distribution networks, as well as exploring online adaptive fine-tuning algorithms under extreme weather events.

Author Contributions: Conceptualization, Z.G. and X.L.; methodology, J.L.; software, X.X.; validation, Z.G., W.X. and F.Z.; formal analysis, F.Z.; investigation, X.X.; resources, X.L.; data curation, Z.G.;

writing—original draft preparation, Z.G.; writing—review and editing, J.L.; visualization, X.L.; supervision, J.L.; project administration, F.Z.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Technology Project of State Grid Jiangsu Electric Power Co., Ltd. (J2025162).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: Authors Ziran Guo, Wenqiang Xie, Jian Liu, Xiaolong Xiao, Xiaoxing Lu and Fangfang Zhu were employed by the company State Grid Jiangsu Electric Power Co. Ltd. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Akbarzadeh, S.; Khosravi, S.; Singh, A.K.; Huinink, H.; Kalogirou, S.A.; Arabkoohsar, A. A review of state-of-the-art in solid-gas thermochemical energy storage technologies for solar energy applications. *Energy* **2026**, *360*, 141591. [CrossRef]
2. Meng, X.; Dong, Y.; Zhou, S.; Bai, E.; Li, G. Research on Improved Sparrow Search Algorithm-Based Scheduling Optimization for Wind–Solar–Hydrogen Storage System. *Electronics* **2026**, *15*, 2888. [CrossRef]
3. Kane, M.M.; Taylor, N.; Månsson, D. Electromagnetic Interference from Solar Photovoltaic Systems: A Review. *Electronics* **2025**, *14*, 31. [CrossRef]
4. Zhong, Y.; He, T.; Mao, Z. Enhanced Solar Power Prediction Using Attention-Based DiPLS-BiLSTM Model. *Electronics* **2024**, *13*, 4815. [CrossRef]
5. Oh, J.; So, D.; Jo, J.; Kang, N.; Hwang, E.; Moon, J. Two-Stage Neural Network Optimization for Robust Solar Photovoltaic Forecasting. *Electronics* **2024**, *13*, 1659. [CrossRef]
6. Rafati, A.; Joorabian, M.; Mashhour, E.; Shaker, H.R. High dimensional very short-term solar power forecasting based on a data-driven heuristic method. *Energy* **2021**, *219*, 119647. [CrossRef]
7. Tang, H.; Kang, F.; Li, X.; Sun, Y. Short-term photovoltaic power prediction model based on feature construction and improved transformer. *Energy* **2025**, *320*, 135213. [CrossRef]
8. Fliess, M.; Join, C.; Voyant, C. Prediction Bands for Solar Energy: New Short-Term Time Series Forecasting Techniques. *Sol. Energy* **2018**, *166*, 519–528. [CrossRef]
9. Lee, H.; Kim, N.-W.; Lee, J.-G.; Lee, B.-T. Uncertainty-Aware Forecast Interval for Hourly PV Power Output. *IET Renew. Power Gener.* **2019**, *13*, 2656–2664. [CrossRef]
10. Sheng, H.; Xiao, J.; Cheng, Y.; Ni, Q.; Wang, S. Short-Term Solar Power Forecasting Based on Weighted Gaussian Process Regression. *IEEE Trans. Ind. Electron.* **2018**, *65*, 300–308. [CrossRef]
11. Golestaneh, F.; Pinson, P.; Gooi, H.B. Very Short-Term Nonparametric Probabilistic Forecasting of Renewable Energy Generation—With Application to Solar Energy. *IEEE Trans. Power Syst.* **2016**, *31*, 3850–3863. [CrossRef]
12. Wang, Y.; Shen, Y.; Mao, S.; Chen, X.; Zou, H. LASSO and LSTM Integrated Temporal Model for Short-Term Solar Intensity Forecasting. *IEEE Internet Things J.* **2019**, *6*, 2933–2944. [CrossRef]
13. He, C.; Zhao, S.; Guo, Z.; Li, X.; Sun, C.; Ge, M. Deep Learning for Solar Power Forecasting by Integrating Historical and Meteorological Data. *Energies* **2026**, *19*, 3451. [CrossRef]
14. Sun, Y.; Szűcs, G.; Brandt, A.R. Solar PV Output Prediction from Video Streams Using Convolutional Neural Networks. *Energy Environ. Sci.* **2018**, *11*, 1811–1818. [CrossRef]
15. Zang, H.; Cheng, L.; Ding, T.; Cheung, K.W.; Liang, Z.; Wei, Z.; Sun, G. Hybrid Method for Short-Term Photovoltaic Power Forecasting Based on Deep Convolutional Neural Network. *IET Gener. Transm. Distrib.* **2018**, *12*, 4557–4567. [CrossRef]
16. Chai, S.; Xu, Z.; Jia, Y.; Wong, W.K. A Robust Spatiotemporal Forecasting Framework for Photovoltaic Generation. *IEEE Trans. Smart Grid* **2020**, *11*, 5370–5382. [CrossRef]
17. Hokmabad, H.N.; Husev, O.; Belikov, J. Day-Ahead Solar Power Forecasting Using LightGBM and Self-Attention Based Encoder-Decoder Networks. *IEEE Trans. Sustain. Energy* **2025**, *16*, 866–879. [CrossRef]
18. Colak, Z. Hybrid Machine Learning Approach in Forecasting Renewable Energy Production: Extra Trees, CatBoost and LightGBM Based Stacking Model. *Renew. Energy* **2026**, *260*, 125167. [CrossRef]
19. Jiang, J.; He, Y.; Zhang, Y.; Yan, J.; Lv, Z.; Liu, Z.; Dai, H.; Zhen, Z. A Multi-Task Learning and GCN-Transformer-Based Forecasting Method for Day-Ahead Power of Wind-Solar Clusters. *Electronics* **2026**, *15*, 462. [CrossRef]
20. Yang, K.; Jiang, C.; Min, Y. Multi-Step Short-Term Forecasting of Photovoltaic Power Utilizing Autoformer with Prophet. *Electronics* **2026**, *15*, 1370. [CrossRef]

21. Wu, S.; Lin, D.; Zhang, Y. Multi-Model Stacking Ensemble with Multi-Perspective Interpretability Analysis for Solar Power Forecasting. *Energies* **2026**, *19*, 3539. [CrossRef]
22. Available online: http://solardata.uoregon.edu/ (accessed on 15 September 2026).
23. Yang, J.; He, H.; Zhao, X.; Wang, J.; Yao, T.; Cao, H.; Wan, M. Day-Ahead PV Power Forecasting Model Based on Fine-Grained Temporal Attention and Cloud-Coverage Spatial Attention. *IEEE Trans. Sustain. Energy* **2024**, *15*, 1062–1073. [CrossRef]
24. Zhang, Y.; Beaudin, M.; Taheri, R.; Zareipour, H.; Wood, D. Day-Ahead Power Output Forecasting for Small-Scale Solar Photovoltaic Electricity Generators. *IEEE Trans. Smart Grid* **2015**, *6*, 2253–2262. [CrossRef]
25. Li, Y.; Song, L.; Zhang, S.; Kraus, L.; Adcox, T.; Willardson, R.; Komandur, A.; Lu, N. A TCN-Based Hybrid Forecasting Framework for Hours-Ahead Utility-Scale PV Forecasting. *IEEE Trans. Smart Grid* **2023**, *14*, 4073–4085. [CrossRef]
26. Tan, D.; Tang, Z.; Zhou, F.; Xie, Y. A Novel Hybrid Model Based on EMD-Improved TCN-Improved TST for Short-Term Railway Traction Load Forecasting. *IEEE Trans. Transp. Electrif.* **2025**, *11*, 6418–6427. [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.