

Article

A Cross-Variable Time-Series Transformer Architecture Integrating Physical Features for Photovoltaic Energy Forecasting

Chen Xie ^{1,2} , Mingju Chen ^{1,2,*} , Yuyan Wang ¹, Yangming Luo ^{1,2}, Xueyang Duan ^{1,2} and Zhihao Lin ^{1,2}

¹ School of Automation and Information Engineering, Sichuan University of Science and Engineering, Yibin 644005, China; 324085404528@stu.suse.edu.cn (C.X.); 324085404512@stu.suse.edu.cn (Z.L.)

² Intelligent Perception and Control Key Laboratory of Sichuan Province, Sichuan University of Science and Engineering, Yibin 644005, China

* Correspondence: chenmingju@suse.edu.cn

Abstract

Accurate photovoltaic (PV) energy forecasting is vital for grid stability and the global low-carbon transition. However, existing data-driven and channel-independent PV energy forecasting models struggle to capture nonlinear meteorological couplings, heterogeneous physical scales across stations, and high-frequency non-stationary fluctuations. To address these limitations, this study proposes a Physics-Guided Cross-Variable Temporal Transformer architecture. Building upon a channel-independent foundation, we introduce a Cross-Variable Attention mechanism to explicitly reconstruct nonlinear photothermal couplings via dynamic attention weights. To resolve multi-station physical scale discrepancies, a Physical Feature-wise Linear Modulation network utilizes installed capacity as a static prior for adaptive cross-station scale alignment. During optimization, a Time Dynamics-Aware Perceiving Loss jointly penalizes absolute errors and first-order time differences, constraining the network's tracking ability for transient ramping. Experiments demonstrate that the proposed architecture overcomes traditional channel-isolation limitations. The model achieves a 21.6% reduction in MSE compared to PatchTST, a 44.0% reduction compared to Autoformer, and a 2.7% improvement in R^2 over Informer. This provides an accurate, generalizable, and physically interpretable solution for collaborative multi-station distributed PV energy forecasting.

Keywords: photovoltaic energy forecasting; temporal transformer; cross-variable attention; nonlinear coupling; physical feature-wise linear modulation; time dynamics

1. Introduction

Amid global climate change and the accelerating low-carbon transition, building a power system dominated by renewable energy has become a broad consensus. Distributed photovoltaic (PV) generation is a major driver of this transition. Its proximity to load centers, low transmission losses, short construction cycles, and environmental benefits have enabled widespread integration into modern smart microgrids and urban energy networks. However, PV energy conversion is a highly complex system that couples natural environments and physical equipment. Its output characteristics depend heavily on surface solar radiation and real-time micro-meteorological conditions. High-frequency cloud shading, abrupt aerosol concentration changes, atmospheric pressure disturbances, and wind convection comprehensively affect generation. Consequently, PV output is inherently random, intermittent, and highly non-stationary in both space and time [1,2]. Directly



Academic Editor: Leijiao Ge

Received: 23 June 2026

Revised: 2 August 2026

Accepted: 3 August 2026

Published: 5 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

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

integrating large-scale, high-penetration PV systems severely disrupts the instantaneous power balance of traditional grids, where generation typically follows load. Denholm et al. [3] illustrated the well-known “duck curve” effect caused by high PV penetration. Subsequently, Zhan et al. [4] quantitatively demonstrated the impacts of deep daytime net load dips and extreme evening ramp events. These phenomena frequently trigger limited responses in Automatic Generation Control (AGC) systems, causing grid frequency oscillations and local voltage violations. Furthermore, they accelerate the degradation of grid-side Energy Storage Systems (ESS). Therefore, high-precision and high-resolution PV energy forecasting is essential. It provides a critical basis for day-ahead scheduling, dynamic spinning reserve allocation, and ESS charge–discharge strategies in microgrids. Accurate forecasting also supports renewable energy integration and stable grid operation [5].

Accurate PV forecasting also contributes to several United Nations Sustainable Development Goals (SDGs). It enables affordable and clean energy (SDG 7) and supports economic growth through optimized grid integration (SDG 8). Maximizing solar utilization is also directly relevant to climate action (SDG 13). As Rajput et al. [6] demonstrated, understanding the physical environmental factors that affect generation efficiency is important for the life-cycle sustainability of PV arrays. Therefore, accurate forecasting models are essential to maximize the environmental and economic benefits of solar assets.

Early PV forecasting relied on purely physical modeling and empirical equations, using first principles from astrophysics and semiconductor physics to establish explicit mathematical relationships between meteorological drivers and power output. For example, Shukla et al. [7] proposed an anisotropic sky model that used rigorous geometric calculations for solar trajectories and extraterrestrial radiation distributions. Subsequently, Stein et al. [8] refined the calculation of absolute diffuse and direct radiation within the Clear Sky Model. Considering the physical structure of PV panels, Akhsassi et al. [9] introduced heat loss coefficient equations based on ambient temperature and convective wind speed to depict the nonlinear degradation curve of photoelectric conversion efficiency. These physical models are valuable during initial commissioning when historical data are scarce, but their limitations become apparent over long-term operation. Yang et al. [10] noted that physical models depend heavily on numerical weather prediction (NWP) systems, which, constrained by coarse spatial grids and hourly resolutions, cannot accurately simulate sub-kilometer, minute-level transient cloud shading. Moreover, static equations cannot adaptively capture nonlinear aging, uneven soiling, or localized hot-spot effects, so forecasting errors grow substantially over time.

The widespread deployment of the power Internet of Things and microgrid SCADA systems has generated massive high-frequency operational data. Consequently, forecasting paradigms shifted to a second stage based on traditional statistics and machine learning. Bacher et al. [11] pioneered autoregressive moving average models to capture diurnal and seasonal periodicities, but these linear difference models gave large errors under sudden, intense nonlinear disturbances. To overcome linear limitations, Mellit et al. [12] showed that shallow artificial neural networks (ANNs) can effectively approximate complex nonlinear mappings. Subsequently, machine learning algorithms gained popularity. Pedro et al. [13] introduced kernel-based Support Vector Regression (SVR), while Voyant et al. [14] verified the robustness of ensemble random forests. Furthermore, Li et al. [15] proved that XGBoost could adaptively extract fluctuation patterns among complex variables. However, modern sensor networks have greatly increased input feature dimensions. Çayır et al. [16] noted that shallow models depend heavily on manual feature engineering and their Markovian assumptions lack long-term memory, making them prone to the curse of dimensionality and local optima in high-dimensional, long-span tasks.

To exploit heterogeneous big data and improve feature extraction, PV forecasting has entered a third phase based on deep temporal networks and Transformers. Gensler et al. [17] introduced Long Short-Term Memory (LSTM) networks, which capture long-range dependencies via internal states and gating mechanisms. Similarly, Zameer et al. [18] demonstrated the efficiency of Gated Recurrent Units (GRUs) for PV sequences. However, the sequential nature of Recurrent Neural Networks (RNNs) limits gradient stability and parallel efficiency on very long sequences [19]. Breakthroughs in natural language processing then revolutionized time-series analysis. The self-attention mechanism and Transformer architecture of Vaswani et al. [20], with global parallel computation and large receptive fields, set new benchmarks for long-term forecasting. To address the quadratic complexity of standard Transformers, Zhou et al. [21] proposed the Informer with ProbSparse attention. Additionally, Wu et al. [22] designed the Autoformer, integrating an auto-correlation mechanism with time-series decomposition. However, advanced deep learning research has revealed that the mixed mapping of high-dimensional multivariate inputs often dilutes attention matrices and introduces cross-channel gradient noise. Addressing this, Nie et al. [23] proposed the milestone PatchTST model, and Zeng et al. [24] confirmed these findings with DLinear. These models utilize a channel-independent mechanism to strictly isolate multidimensional input variables topologically. By implementing independent feature patching within the time domain, these methods reduce overfitting risks and improve the representation quality of isolated temporal signals.

When moving from general time-series benchmarks to practical PV microgrids, existing data-driven, channel-independent architectures show limitations in representing complex physical processes.

For feature extraction, capturing the coupled features of multidimensional meteorological factors remains restricted. From an engineering thermophysics perspective, PV output is not a simple linear function of meteorological factors but is governed by strong nonlinear photothermal interactions [25]. Studies indicate that operating temperatures and photoelectric conversion efficiencies are nonlinear outcomes of dynamically interwoven meteorological conditions. These include solar irradiance accumulation, ambient temperature fluctuations, and convective wind heat dissipation. Ono et al. [26] highlighted that during complex weather transitions, abrupt irradiance drops from clouds are often accompanied by simultaneous changes in temperature and humidity. This strong multivariate synergistic effect directly determines the transient physical boundaries of final power outputs. However, existing channel-independent architectures enforce absolute inter-variable isolation to suppress gradient noise from multivariate mixing [27]. Although this ensures temporal purity of each signal, it cuts off the intrinsic coupling among meteorological factors during feature representation. Consequently, under severe micro-meteorological shifts, this isolation creates a blind spot for physical synergy. The model cannot capture the chain reactions triggered by simultaneous changes in multiple drivers, leading to larger deviations at dynamic turning points.

For spatial generalization, physical scale discrepancies among heterogeneous stations weaken features. In distributed microgrid collaborative scheduling, the installed capacity of independent PV nodes exhibits inherent spatial heterogeneity. Under identical meteorological drivers, kilowatt-scale residential modules and megawatt-scale grid-connected arrays demonstrate significant cross-scale differences in absolute output magnitudes and dynamic response amplitudes. Conventional deep learning models rely on data normalization (e.g., Z-Score or Min–Max) during joint multi-station modeling to accelerate convergence and stabilize gradients [28,29]. However, Bhanja et al. [30] noted that while standard normalization homogenizes data distributions, it essentially strips away the absolute physical scales representing actual plant capacities. This weakening obscures equipment-specific

nonlinearities (e.g., inverter clipping), so the network learns only homogenized relative patterns. Without absolute spatial anchors, models struggle to generalize to nodes with unknown capacities, hindering adaptive reconstruction and precise cross-scale alignment.

For backward optimization, statistical objectives exert a smoothing effect on high-frequency non-stationary fluctuations. Even with effective forward feature extraction, the final optimization is still governed by the loss function. Most models use MSE (L2 norm). Kudrat et al. [31] showed that the point-wise squared penalty of MSE effectively estimates the conditional expectation of the target, which guides networks toward the local mean of the training distribution. Wellby et al. [32] noted that during high-frequency transient ramping events (caused by fast-moving clouds), MSE-optimized models produce overly smooth trajectories to avoid penalizing extreme values. Without first-order difference constraints, this leads to peak smoothing and phase lag, significantly reducing the network's ability to track non-stationary dynamics.

Recent PV forecasting methods increasingly use multi-source meteorological variables. To place our framework in context, we contrast current challenges with classical methods. Traditional physical equivalent models are deterministic and interpretable, but they lack adaptability to equipment degradation [33]. Classical machine learning methods extract shallow trends but often fail to capture very long-term dependencies. Modern deep learning architectures attempt to resolve these issues by processing multisource variables; however, they typically treat inputs as strictly independent channels, failing to capture intense multidimensional coupling effects and nonlinear photothermal synergies. Meanwhile, atmospheric volatility introduces high-frequency fluctuations, degrading accuracy during abrupt environmental changes. A research gap persists in explicitly modeling these dynamic synergies while suppressing stochastic noise [34]. The proposed architecture systematically bridges these historical and contemporary gaps. By embedding spatial anchor modulation into an attention mechanism, it retains the long-term tracking ability of modern deep learning while reintroducing the physical constraints typical of thermodynamic models. This hybrid paradigm ensures that the predictions remain mathematically optimal and physically realistic, distinguishing the current contribution from prior purely data-driven approaches.

To alleviate the limitations of existing models in multidimensional meteorological coupling, physical scale discrepancies, and non-stationary dynamics tracking, we propose a Physics-Guided Cross-Variable Temporal Transformer forecasting model. It utilizes the channel-independent PatchTST as the underlying temporal backbone for day-ahead photovoltaic forecasting over a 24-h horizon. This architecture establishes a third-order progressive feature decoupling and reconstruction logic spanning the temporal, variable, and spatial domains. The main theoretical and architectural contributions of this paper are as follows:

- Introducing a Cross-Variable Attention (CVA) mechanism based on multidimensional meteorological factors. After extracting pure temporal features via the channel-independent backbone, we design a cross-variable interaction module within the variable domain. This mechanism treats the evolving sequences of meteorological factors as independent interactive entities. By computing dynamic attention weights, it explicitly reconstructs the nonlinear synergistic responses between meteorological drivers and output power. This effectively mitigates topological restrictions caused by single-channel isolation, enhancing the model's representation of complex photothermal coupling dynamics;
- Designing an installed-capacity-guided Physical Feature-wise Linear Modulation (PFiLM) network. To address physical scale discrepancies in joint multi-station forecasting, this module utilizes the target station's installed capacity as a static prior

anchor. Through a dual-track Multilayer Perceptron (MLP), it smoothly maps this physical attribute into high-dimensional affine scaling and shifting operators, injecting them into the deep feature space. This mechanism enables adaptive alignment of absolute magnitudes across arrays of varying scales, thereby enhancing cross-domain generalization performance among heterogeneous stations;

- Formulating a Time Dynamics-aware Perceiving Loss (TDPL) function. To mitigate the temporal smoothing effect caused by singular statistical fitting objectives, we designed a dynamic penalty mechanism that jointly penalizes absolute magnitude errors and first-order time differences. By adding explicit constraints on transient ramping errors between adjacent time steps within the gradient propagation flow, this function compels the neural network to accurately track the physical turning points of high-frequency power mutations. This fundamentally optimizes the underlying logic, effectively alleviating peak-smoothing and phase-lag phenomena in traditional prediction trajectories.

2. Related Work

Accurate PV energy forecasting is a fundamental prerequisite for the secure, stable, and economic operation of modern microgrids. Driven by advancements in data acquisition and computational power, deep learning has recently achieved remarkable progress in this domain. This section systematically analyzes existing research advancements and current theoretical limitations across three dimensions: multidimensional meteorological coupling, spatial scale heterogeneity, and high-frequency non-stationary fluctuations.

2.1. Multi-Dimensional Meteorological Factor Coupling and Cross-Variable Feature Synergy

In photovoltaic forecasting with multidimensional meteorological features (e.g., irradiance, temperature, humidity), temporal Transformer architectures have evolved along two distinct paths. Wen et al. [35] noted that early architectures defaulted to cross-channel fusion, attempting to capture all variable interactions in the initial mapping layer. Zhang et al. [36] proposed Crossformer, which utilizes cross-dimensional dependencies for 3D array embedding of multivariate sequences. It constructs a routing mechanism to establish global attention maps, thereby strengthening explicit interactions. Conversely, Liu et al. [37] proposed iTransformer and related channel-independent architectures, arguing from an opposing perspective. They demonstrated that in real-world heterogeneous data, varying physical dimensions and evolution cycles make global channel mixing prone to distribution shifts and gradient noise propagation. Instead, topologically isolating variables yields more robust temporal representations.

However, migrating these cutting-edge architectures to PV microgrid forecasting reveals a mismatch with underlying physical logic. Du et al. [38] demonstrated experimentally in engineering thermophysics that PV generation results in nonlinear photothermal hysteresis. This effect introduces a temporal lag between irradiance and ambient temperature, which are two key meteorological drivers. When surging solar irradiance rapidly increases semiconductor carriers, accumulated panel heat intensifies lattice vibrations. This causes a nonlinear shrinkage of the monocrystalline silicon bandgap, ultimately resulting in thermal losses that degrade photoelectric conversion efficiency [39]. Given this physical reality, existing deep learning architectures face limitations. Channel-mixing architectures cannot effectively isolate high-frequency meteorological noise, often causing attention dispersion within deep networks. Meanwhile, channel-independent architectures like PatchTST guarantee the temporal purity of individual variables but sever vital photothermal synergies. They cut off the physical coupling where “irradiance drives temperature, and temperature inversely modulates efficiency.” Therefore, this study proposes the CVA

mechanism, which maintains temporal purity via a channel-independent backbone while dynamically reconstructing the nonlinear synergy between meteorology and power in the variable domain. Ultimately, this approach achieves an optimal balance between mathematical isolation and physical coupling.

2.2. Spatial Scale Heterogeneity and Physical Prior Modulation

As distributed rooftop PV and regional microgrids proliferate, developing global forecasting models for multiple stations has become a research hotspot. Traditional approaches predominantly rely on graph structure learning. For instance, Jiao et al. [40] utilized spatiotemporal graph routing to capture spatial correlations across multiple geographic stations via dynamic adjacency matrices. However, these purely data-driven models rely heavily on the statistical consistency of relative input data distributions. Consequently, they frequently encounter “cold start” problems when incorporating new nodes lacking historical data.

Recently, Karniadakis et al. [41] introduced Physics-Informed Neural Networks (PINNs), which constrain the deep learning solution space using partial differential equations or prior physical laws, thereby reducing reliance on large datasets. In solar forecasting, Gao et al. [42] injected atmospheric transmittance equations and solar zenith angle geometric constraints into deep residual networks as additional penalty terms. This improved the model’s temporal robustness under dynamic cloud cover variations.

However, existing data-driven and physics-informed efforts generally overlook the most intuitive heterogeneity source in distributed PV networks: absolute physical scale discrepancies determined by installed capacity. During joint multi-station model training, conventional models frequently employ standardizations like Z-Score or Min–Max to accelerate network convergence. Lima et al. [43] demonstrated that while standardizations homogenize time-series means and variances, they simultaneously strip away absolute scales and amplitude baselines. Without absolute magnitude priors, these networks struggle to reconstruct power scales accurately during cross-domain transfer.

In computer vision and multimodal reasoning, Perez et al. [44] proposed the Feature Modulation (FiLM) architecture. They proved the efficiency of utilizing dual-track mapping to losslessly inject external priors into deep networks. Inspired by this concept, we avoid direct data mixing and instead utilize the target station’s installed capacity as a high-dimensional affine operator, which is explicitly and adaptively injected into the PFILM network for spatial scale alignment.

2.3. Non-Stationary High-Frequency Fluctuations and Dynamical Objective Optimization

Beyond evolving forward feature extraction structures, backpropagation gradient guidance mechanisms also exhibit theoretical limitations when handling high-frequency PV fluctuations. Currently, mainstream temporal forecasting architectures primarily rely on L1 or L2 norms for loss functions. Statistically, MSE optimization essentially approximates the conditional expectation of a target distribution. Consequently, when processing PV sequences with high-frequency, non-stationary abrupt changes, point-wise absolute error penalties guide networks to generate overly smooth trajectories. This ultimately leads to a mean-reversion behavior in the predictions.

To address this issue, researchers in time-series forecasting and computer vision have explored shape-aware objective functions. For instance, the DILATE loss function proposed by Le Guen et al. [45] combines soft Dynamic Time Warping (DTW) with a temporal distortion penalty. This aligns predicted and true sequence shapes in non-Euclidean space, alleviating phase lag during abrupt transitions. However, the quadratic complexity of

the DTW algorithm imposes severe computational constraints during high-resolution, minute-level microgrid scheduling tasks.

In power system scheduling, these smoothing and lagging defects directly impact system stability. Studies emphasize that transient ramping events driven by fast-moving clouds are highly destructive to microgrid frequency stability [46,47]. If forecasting models output excessively smooth curves, Automatic Generation Control (AGC) systems struggle to proactively dispatch spinning reserves. Therefore, to avoid high computational overhead while accurately constraining the physical turning points of high-frequency PV ramping, we introduce a lightweight penalty mechanism. This mechanism, termed TDPL, targets the sequence's first-order time derivative. It reduces peak smoothing and phase lag within the gradient flow, improving the model's ability to capture high-frequency dynamic features.

Table 1 summarizes the proposed model against existing forecasting paradigms and highlights three core deficiencies in current architectures. Specifically, these approaches show an inherent trade-off between temporal signal fidelity and the ability to capture nonlinear meteorological interactions. Moreover, conventional normalization typically discards absolute magnitude information, which compromises spatial coherence across heterogeneous sensor arrays. Meanwhile, standard loss functions tend to under-penalize abrupt ramping events, manifesting as over-smoothed peaks and delayed phase responses. Our framework addresses these issues through three corresponding mechanisms, namely physics-guided interaction reconstruction, spatial-anchor-based magnitude calibration, and derivative-aware regularization, providing a physically grounded solution for cooperative PV forecasting.

Table 1. Comprehensive comparison of theoretical limitations in existing forecasting paradigms and the proposed structural solutions.

Research Dimension	Methodologies	Limitations	Proposed Method Solutions
Meteorological Coupling	Global channel mixing and channel-independent architectures	Mixing introduces gradient noise, while isolation severs nonlinear photothermal synergies	Introduces Cross Variable Attention to reconstruct physical interactions while preserving temporal purity
Spatial-Scale Heterogeneity	Graph structure learning and standard data normalization	Normalization removes absolute magnitude baselines, limiting generalization across multi-scale arrays	Uses a Physical Feature-wise Linear Modulation network to map installed capacity into spatial anchors for magnitude alignment
High-frequency Fluctuations	Optimization based on standard statistical loss functions	Point-wise penalties cause peak smoothing and temporal phase lag during transient events	Proposes a Time Dynamics-aware Perceiving Loss to penalize time differences and precisely track transient ramping

3. Methodology

Photovoltaic energy forecasting faces complex challenges, including multidimensional meteorological coupling, physical scale discrepancies, and high-frequency non-stationary fluctuations. To address these, we propose a Physics-Guided Cross-Variable Temporal Transformer forecasting model, utilizing the channel-independent PatchTST as its foundation. The overall architecture is illustrated in Figure 1. This design separates concerns that are entangled in conventional end-to-end models and establishes a progressive three-stage feature decoupling and reconstruction logic across the temporal, variable, and spatial domains.

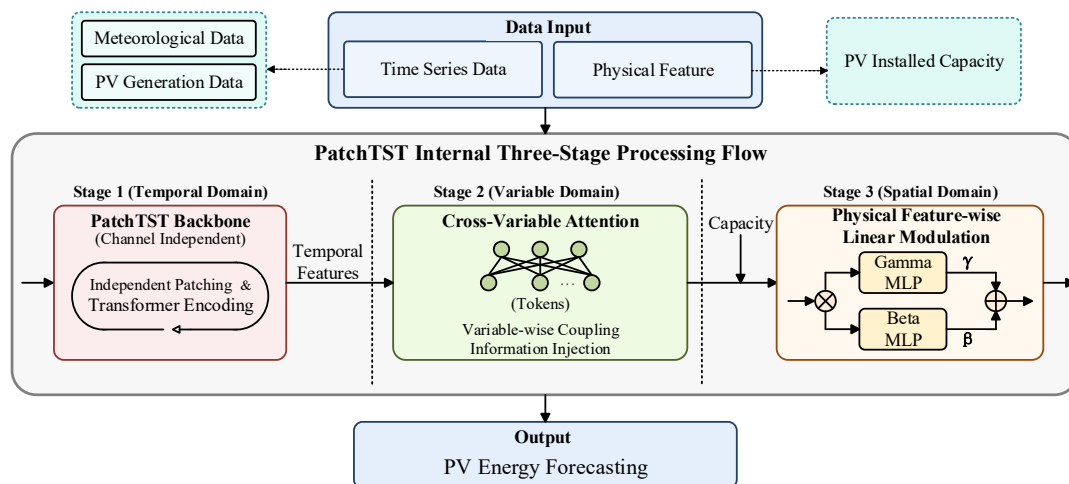


Figure 1. Comprehensive framework of the proposed Physics-Guided Cross-Variable Temporal Transformer. The framework operates in three decoupled stages: temporal extraction via a channel-independent backbone, physical interaction reconstruction via cross-variable attention, and spatial scaling via feature modulation. A dynamics-aware loss further suppresses ramp-induced smoothing.

First, in the temporal domain, the channel-independent backbone extracts pure temporal features, effectively suppressing gradient noise caused by multivariate mixing. Second, the data flow enters the variable domain to overcome the physical interaction severance caused by channel isolation. A Cross-Variable Attention mechanism computes dynamic attention weights to explicitly reconstruct the nonlinear photothermal coupling between meteorological drivers and output power. This effectively mitigates limitations in multidimensional meteorological synergistic representation. Subsequently, to address the absolute magnitude loss resulting from heterogeneous station standardization, a Physical Feature-wise Linear Modulation network is utilized. This network maps the target station's installed capacity, acting as a static physical prior, into high-dimensional affine operators to achieve adaptive cross-station scale alignment. Finally, traditional statistical objectives tend to smooth high-frequency non-stationary fluctuations. To counteract this, a Time Dynamics-aware Perceiving Loss is introduced during global optimization. By jointly penalizing absolute magnitude errors and first-order time differences, TDPL constrains the network's ability to track transient power ramps, correcting peak smoothing and phase lag in prediction trajectories.

3.1. Cross-Variable Attention Mechanism (CVA)

The original PatchTST model uses a channel-independent mechanism that effectively extracts pure univariate temporal features and prevents overfitting. However, it isolates inter-variable interactions at the network structure level. This defect prevents the model from representing nonlinear photothermal coupling among multidimensional meteorological factors. The CVA mechanism addresses this by dynamically reconstructing multidimensional meteorological synergy after the temporal backbone. Specifically, it models the nonlinear synergistic dynamics between meteorological drivers and power output within a high-dimensional hidden space. Figure 2 illustrates the overall CVA architecture.

The feature matrix from the preceding temporal network is transposed and reshaped into $X_C \in \mathbb{R}^{B \times M \times L}$ upon entering the CVA module. In this view, B represents the batch size. The variable channel count M acts as independent token sequences. The temporal trajectory window T serves as the initial feature embedding. To enhance representation capacity and nonlinear mapping for features with varying dimensions, linear mapping with a GELU

activation function is applied. This projects the temporal trajectory dimension into a high-dimensional hidden space d_{model} , yielding the hidden representation $H \in \mathbb{R}^{B \times M \times d_{model}}$.

$$H = GELU(W_{in}X_C + b_{in}) \quad (1)$$

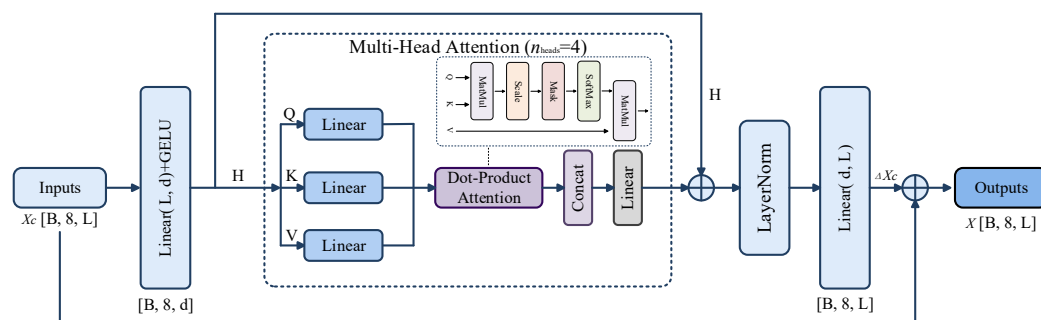


Figure 2. The overall CVA architecture. This module reshapes temporal features across variables, applies multi-head self-attention to compute driver-interaction weights, and adds a residual skip to preserve raw temporal signals.

Subsequently, this high-dimensional feature matrix enters a Multi-Head Self-Attention mechanism. Through linear transformations within the variable dimension, the model calculates the Query (Q), Key (K), and Value (V) matrices:

$$Q = HW_Q \quad (2)$$

$$K = HW_K \quad (3)$$

$$V = HW_V \quad (4)$$

By computing scaled dot-product attention, a dynamic correlation matrix among all variables is generated. This matrix essentially quantifies attention weights, such as the impact of “current temperature anomalies” on the “overall power output trajectory”:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

The self-attention weight matrix $A = \text{Softmax}(QK^T / \sqrt{d_k})$ forms an $M \times M$ dynamic relationship graph. This explicitly quantifies the transient impact of different meteorological fluctuations on the global evolutionary trajectory. The scaling factor $\sqrt{d_k}$ effectively prevents Softmax gradient vanishing caused by excessive inner products.

After multi-head fusion and Layer Norm, an output projection layer maps the high-dimensional features back to the original temporal trajectory length T . This extracts the coupled correction term $\Delta X_C \in \mathbb{R}^{B \times M \times T}$, which captures cross-variable physical interactions. To balance maintaining temporal purity and introducing variable coupling, the CVA module features a strict residual connection at the end:

$$X = X_C + \Delta X_C \quad (6)$$

By superimposing the coupled correction term ΔX_C , the model injects cross-variable synergistic information while retaining the univariate temporal features extracted by the backbone. This modification addresses the structural constraints of conventional channel isolation methods during meteorological variations. It reconstructs the photothermal synergistic network of multidimensional meteorological drivers at the underlying topology.

3.2. Physical Feature-Wise Linear Modulation (PFiLM) Network

The feature matrix from the CVA module achieves deep interweaving of meteorological and temporal features. However, conventional data standardization removes the absolute magnitude that represents the PV plant's true physical scale. In realistic distributed microgrid scheduling, differing installed capacities among heterogeneous stations cause significant physical scale spans in actual outputs. To eliminate this model generalization bottleneck, we design the PFiLM network. Inspired by the conditional mapping network proposed by Perez et al. [44], we extend semantic conditions to hard equipment parameters. This module transforms the static physical scale into dynamic spatial scaling anchors within deep networks. Figure 3 shows the overall PFiLM architecture.

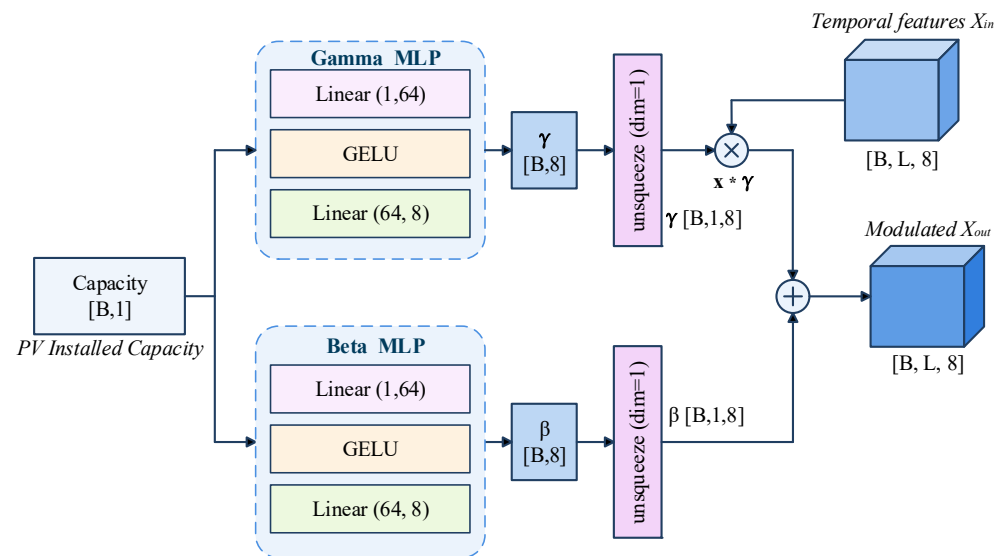


Figure 3. The overall PFiLM architecture. This component converts static station capacities into dynamic affine operators via a dual-branch MLP. The resulting scaling and shifting parameters align feature magnitudes across heterogeneous sites.

Let $C \in \mathbb{R}^1$ denote the physical installed capacity scalar of the PV array corresponding to the current input sample. To capture the nonlinear scaling laws between installed capacity and inverter output, PFiLM discards simple scalar multiplication. It utilizes a parallel dual-track MLP to smoothly map this scalar into a high-dimensional affine parameter space. Specifically, after a linear layer and GELU activation, the capacity scalar dynamically generates a scaling factor $\gamma \in \mathbb{R}^M$ and a shifting factor $\beta \in \mathbb{R}^M$. These align with the feature channel dimension M :

$$\gamma = W_{\gamma 2} \cdot \text{GELU}(W_{\gamma 1} \cdot C + b_{\gamma 1}) + b_{\gamma 2} \quad (7)$$

$$\beta = W_{\beta 2} \cdot \text{GELU}(W_{\beta 1} \cdot C + b_{\beta 1}) + b_{\beta 2} \quad (8)$$

This dual-track mapping enables the network to learn complex physical boundaries. The scaling factor γ reconstructs the dynamic extreme upper limits related to inverter clipping effects in high-dimensional space. Meanwhile, the shifting factor β adaptively fits the low-light startup thresholds and basic loss offsets of microinverters. After obtaining these spatial scale operators, the model broadcasts them along the temporal dimension T . It then performs element-wise affine modulation on the fused feature matrix X_{in} outputted by the CVA module:

$$X_{out} = X_{in} \odot \gamma + \beta \quad (9)$$

By using each station's static capacity as a physical prior, the MLP maps it into high-dimensional affine operators and injects them into the deep feature space. This resolves the cross-domain generalization issue of purely data-driven models and achieves physical magnitude alignment for multi-scale heterogeneous data.

3.3. Time Dynamics-Aware Perceiving Loss (TDPL)

This relies solely on the traditional MSE loss during optimization, which causes issues. Even with accurate forward feature extraction, the model tends toward mean reversion. Mathematically, the point-wise squared penalty of MSE estimates the conditional expectation of the target, which tends to guide the network toward overly smooth prediction trajectories.

When grids face high-frequency fluctuations from fast-moving clouds, this smoothing mechanism causes phase lag and peak flattening. To constrain the non-stationary dynamics of PV power at the derivative level, we propose the TDPL function. Its architecture is illustrated in Figure 4.

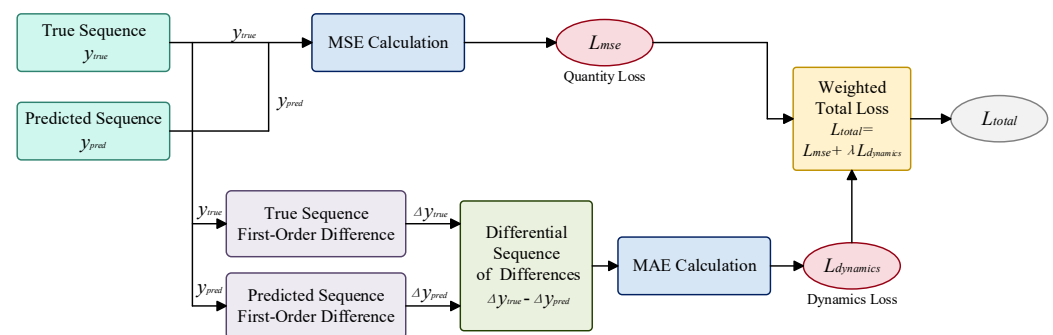


Figure 4. The overall TDPL architecture. This loss jointly penalizes magnitude errors and temporal derivatives, forcing the network to track sharp power turning points and reduce peak smoothing plus phase lag.

TDPL employs a dual-track synergistic constraint strategy. It jointly optimizes the absolute magnitude error and the first-order time derivative error of prediction trajectories:

$$\mathcal{L}_{\text{TDPL}} = \mathcal{L}_{\text{MSE}}(y, \hat{y}) + \lambda \mathcal{L}_{\text{dynamics}}(y, \hat{y}) \quad (10)$$

The first term, $\mathcal{L}_{\text{MSE}}$, measures the point-wise absolute error of the prediction trajectory:

$$\mathcal{L}_{\text{MSE}}(y, \hat{y}) = \frac{1}{B \times T} \sum_{b=1}^B \sum_{t=1}^T (y_{b,t} - \hat{y}_{b,t})^2 \quad (11)$$

The second term, $\mathcal{L}_{\text{dynamics}}$, explicitly penalizes the power change rate between adjacent time steps, acting as a temporal ramping derivative constraint:

$$\mathcal{L}_{\text{dynamics}}(y, \hat{y}) = \frac{1}{B \times (T-1)} \sum_{b=1}^B \sum_{t=2}^T |(y_{b,t} - y_{b,t-1}) - (\hat{y}_{b,t} - \hat{y}_{b,t-1})|^2 \quad (12)$$

Here, $\lambda > 0$ serves as the balancing coefficient. This optimization mechanism adjusts the direction of gradient propagation during model training.

During stable weather, the first-order difference is small, allowing the model to smoothly optimize MSE. However, when severe weather causes steep slopes in the real power curve, the first-order difference surges. Consequently, the weight of $\mathcal{L}_{\text{dynamics}}$ in the total loss amplifies. The resulting gradient encourages the network to avoid smooth transitions and instead fit the sharp turning points of power changes.

By jointly penalizing absolute errors and first-order time differences, TDPL constrains the network's ability to track transient power ramps. This corrects the peak-smoothing and phase-lag issues inherent in MSE-based optimization, shifting the network from static fitting to high-frequency transient tracking.

4. Experimental Settings and Datasets

4.1. Datasets and Data Preprocessing

To evaluate the generalization capability and spatiotemporal feature extraction performance of the proposed architecture, we use two real-world distributed PV datasets with distinctly different physical topologies and climatic conditions. The Hong Kong dataset [48] serves as the primary benchmark, while the Australian Desert Knowledge Precinct (DKASC) dataset is used for cross-domain validation to test robustness under extreme arid and highly heterogeneous conditions.

4.1.1. Hong Kong Distributed Rooftop PV Dataset

The primary dataset comes from distributed rooftop PV systems in Hong Kong, collected from sixty grid-connected rooftop stations and one on-site automatic weather station at the Hong Kong University of Science and Technology. This dataset was chosen for its complex subtropical monsoon microgrid characteristics, which provide a demanding testbed for evaluating model robustness under strong convective weather and sudden meteorological shifts. Summer temperatures in this region can exceed 30 °C, and annual average relative humidity generally exceeds 75%. The microgrid includes 6085 PV modules with a total installed capacity of 2230.8 kWp and recorded continuous operation and meteorological data from 2021 to 2023. The feature space includes seven meteorological variables collected by the on-site station: global horizontal irradiance, rainfall, relative humidity, sea level pressure, temperature, visibility, and wind speed. Table 2 summarizes the core physical and meteorological metadata extracted from the original observation framework. A key strength of this dataset is its high temporal granularity, with inverter-level electrical measurements at five-minute intervals and localized meteorological data at one-minute intervals.

Table 2. Comprehensive metadata summary of the Hong Kong distributed photovoltaic dataset based on the original observation framework.

Feature Category	Observation Variable	Physical Unit	Temporal Resolution
Site Level Generation (Without Optimizer)	Active Power and Generation	W and kWh	1 h
Site Level Generation (With Optimizer)	Active Power and Generation	W and kWh	15 min
Solar Resource	Irradiance	W/m ²	1 min
Thermal Dynamics	Ambient Temperature	Celsius	1 min
Atmospheric Moisture	Relative Humidity	Percent	1 min
Atmospheric Pressure	Sea Level Pressure	hPa	1 min
Environmental Visibility	Visibility	km	1 min
Convective Factors	Wind Speed	m/s	1 min
Precipitation	Rainfall	mm	1 min

To process the diverse data sources, a structured preprocessing pipeline was implemented. Meteorological variables recorded at one-minute intervals were resampled to

hourly resolution using a right-closed interval to reduce high-frequency noise. During downsampling, rainfall was aggregated via summation, while continuous features such as irradiance, pressure, and temperature were averaged. Missing values were filled using linear interpolation and bidirectional smoothing. Inverter-level power data, sampled at five-minute intervals, were truncated to remove invalid pre-connection segments and resampled to hourly granularity. This step synchronized the meteorological drivers and power responses. The meteorological matrix was then merged with each power station's data using a left join on the temporal axis. Rated installed capacity for each station was extracted from standardized equipment metadata and integrated into the feature stream as a static physical prior for the modulation layer.

4.1.2. DKASC Heterogeneous Array Dataset

To test spatial generalization under extreme arid conditions and highly heterogeneous arrays, we use the DKASC dataset from the Desert Knowledge Australia Solar Centre in Alice Springs as a cross-domain benchmark. Unlike Hong Kong's subtropical climate, this site is in an extreme arid desert zone with intense solar radiation and large diurnal temperature swings. DKASC is a large-scale PV platform that has been operating since 2008 and includes 49 independent arrays with diverse cell technologies. These include monocrystalline silicon, polycrystalline silicon, cadmium telluride thin-film, and copper indium gallium selenide. Furthermore, it incorporates different mechanical structures, such as fixed ground mounts, single-axis, and dual-axis solar tracking systems. This heterogeneity in technologies, tracking designs, and module degradation provides a stringent test of the model's spatial generalization.

The raw DKASC data (2021–2023) required structural reorganization: the wide-table format was converted into a long-table format grouped by site and time. A five-dimensional meteorological feature subset including irradiance, rainfall, relative humidity, ambient temperature, and wind speed was extracted. The original five-minute samples were resampled to hourly resolution using a right-closed interval. Rainfall was aggregated by maximum, and other variables were averaged. These features were mapped to the arrays along the temporal axis. Since nameplate capacity was missing, the effective capacity for each site was estimated based on its maximum global peak output over the three-year period. These capacity scalars were integrated as physical priors for the modulation layer.

4.1.3. Data Features and Cross-Domain Correlation Analysis

After data cleaning and temporal alignment, we performed a Pearson correlation analysis on the global meteorological driving features during daytime hours for both datasets. Figure 5 illustrates the correlation coefficients between meteorological factors and normalized PV power for both datasets. Notably, wind speed is omitted from the DKASC correlation heatmap because all recorded wind speed values in this dataset are zero.

As Figure 5 shows, irradiance remains the dominant driving force for power output. This holds true for both Hong Kong's subtropical monsoon microgrid (0.74) and DKASC's arid desert arrays (0.87). Strong nonlinear photothermal cross-coupling is also evident. The irradiance–temperature correlation reaches 0.51 and 0.58 in the two datasets. This covariance confirms the photothermal hysteresis effect: sudden irradiance spikes cause nonlinear temperature increases, which trigger thermal losses in conversion efficiency. Furthermore, inter-domain climate differences cause significant driver weight drifts. In Hong Kong, relative humidity (−0.43) and visibility (0.27) exert a stronger modulating effect on final power. This ensures consistency with the region's frequent aerosol shading and strong convective weather.

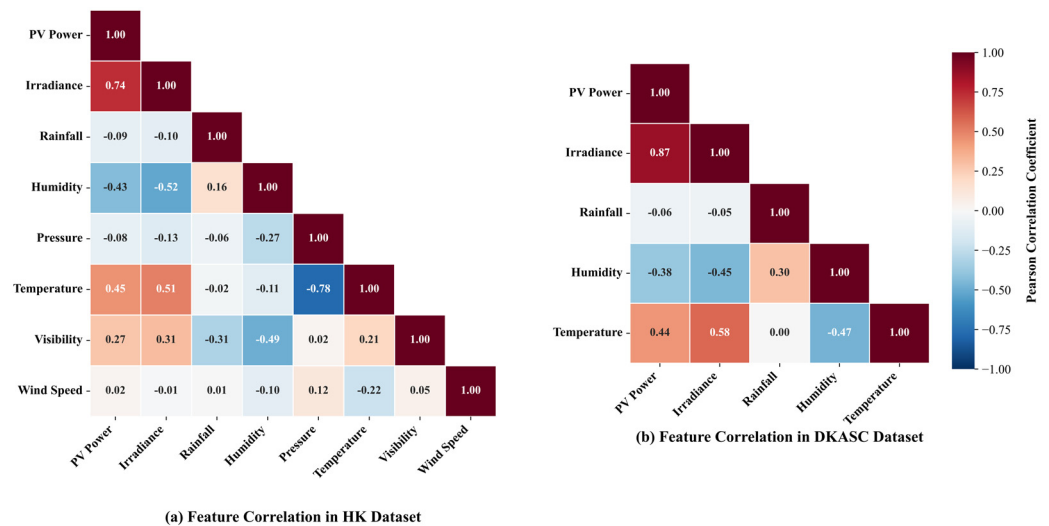


Figure 5. A correlation heatmap of the Hong Kong and DKASC datasets. Irradiance shows the strongest positive correlation with power output across all climate types. The irradiance–temperature covariance further confirms the photothermal hysteresis effect.

This analysis highlights the limitations of channel-independent models: strict topological isolation prevents them from capturing cross-variable photothermal synergies. This physical coupling provides a strong motivation for introducing the CVA mechanism and PFILM in our architecture.

4.1.4. Dataset Complexity and Scenario Definition

To characterize dataset complexity and support rigorous evaluation, we examined the non-stationary behavior of the data distributions. Both datasets display distinct high-frequency volatility. In Hong Kong, complexity arises from subtropical convective weather, which produces severe ramping events with irradiance shifts of over 40% within hourly intervals. Conversely, the DKASC dataset shows extreme diurnal temperature swings and intense clear-sky radiation peaks, providing a high-variance setting with substantial spatial heterogeneity.

To avoid masking local failures with global averages, the evaluation pipeline included explicit scenario definitions. The test set was split into three categories: Clear Sky, Continuous Overcast, and Rapid Cloud Shading. These categories test model tracking capability under both stable and severe weather conditions, ensuring that the evaluation reflects real grid operational complexity.

4.2. Experimental Environment and Parameter Configuration

Model training and evaluation were conducted on a computing platform equipped with an NVIDIA GeForce RTX 3060 GPU. The software environment utilized Python 3.11.0 and the PyTorch 2.6.0 framework. Additionally, CUDA 12.6 was configured to accelerate tensor computations.

To ensure the rigor and reproducibility of comparative experiments, a random seed of 42 was established, and all experimental groups utilized strictly aligned hyperparameter configurations. Given the intra-day periodicity of the PV output, the input sequence and prediction lengths were set to 288 (12-day historical window) and 24 (1-day-ahead prediction). The patch length and stride were configured to 24 and 12, respectively.

The training batch size was uniformly set to 64. Parameters were updated using an optimizer with a learning rate of 0.0001, incorporating a 0.2 dropout mechanism. The

maximum training epochs were capped at 60, with an early stopping patience of 10. Finally, all evaluation error metrics were calculated using inverse-transformed data.

4.3. Evaluation Metrics

To quantitatively evaluate the forecasting performance of the models, four standard statistical metrics were employed: Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). The formulations are defined as follows:

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (13)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (14)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (15)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (16)$$

where y_i and $\hat{y}_i$ represent the actual and predicted PV power values at the i -th step, respectively, and $\bar{y}$ is the global mean of the actual values.

Among these, MSE and R^2 serve as the primary evaluation metrics in this study. By squaring the errors, MSE heavily penalizes larger forecasting deviations, making it the definitive standard for evaluating the model's accuracy in capturing extreme meteorological abruptions. Meanwhile, R^2 directly reflects the model's overall capability to explain the temporal fluctuation trends of PV output (a value closer to 1 indicates superior goodness of fit). Conversely, MAE and RMSE are utilized strictly as auxiliary reference metrics to assess the baseline average error and original-magnitude deviations.

5. Experimental Results and Analysis

This section evaluates model performance on the day-ahead forecasting task (predicting PV output for the next 24 h). This lead time provides the grid with a sufficient margin for generation scheduling, spinning reserve allocation, and cross-day ESS dispatch. All baseline comparisons and ablation experiments therefore use a 24-h forecasting horizon.

5.1. Comparative Experiments

To evaluate the proposed physics-guided architecture's advancement, we conducted benchmark tests. These utilized the Hong Kong and DKASC datasets, covering multi-site heterogeneous features. Representative models were tested, including PatchTST, Informer, Autoformer, Transformer, TiDE [49], and DLinear.

5.1.1. Performance Analysis on the Hong Kong PV Datasets

Table 3 reports the results. Our model achieves the lowest MSE (6.2493) and the highest R^2 (0.7945) among all baselines. Compared to PatchTST (7.9735 MSE), the improvement is 21.6%. This gain can be attributed to two factors: the CVA module provides the missing cross-variable information, and the TDPL prevents the oversmoothing that affects the MSE-optimized PatchTST.

Classical attention models with global mixed mapping (Informer and Transformer) showed acceptable R^2 values (0.7945 and 0.7943), but their internal cross-mapping introduces gradient noise, resulting in higher MSE (6.2505 and 6.2542). Meanwhile, the Autoformer, integrating a time-series decomposition mechanism, faced severe feature overfitting risks on this dataset. Its MSE surged to 11.1564, and R^2 plummeted to 0.6331.

These quantitative metric differences are physically corroborated by the three typical weather prediction curves in Figure 6. Under stable clear sky conditions (Figure 6a), most deep learning models fit the basic envelope well. The exception is Autoformer, which exhibits significant peak underestimation. However, under continuous overcast (Figure 6b) and extreme rapid cloud shading (Figure 6c) scenarios, the theoretical defects of each baseline model amplify.

Table 3. Comparison results of different models on the Hong Kong PV dataset. Bold values denote the best performance.

Model	MSE	MAE	RMSE	R^2
PatchTST	7.9735	1.5904	2.8237	0.7378
Informer	6.2505	1.4055	2.5001	0.7945
Autoformer	11.1564	1.9702	3.3401	0.6331
DLinear	22.0419	1.7560	4.6949	0.2752
Transformer	6.2542	1.3967	2.5008	0.7943
TiDE	17.0808	1.7942	4.1329	0.4383
Ours	6.2493	1.3960	2.4999	0.7945

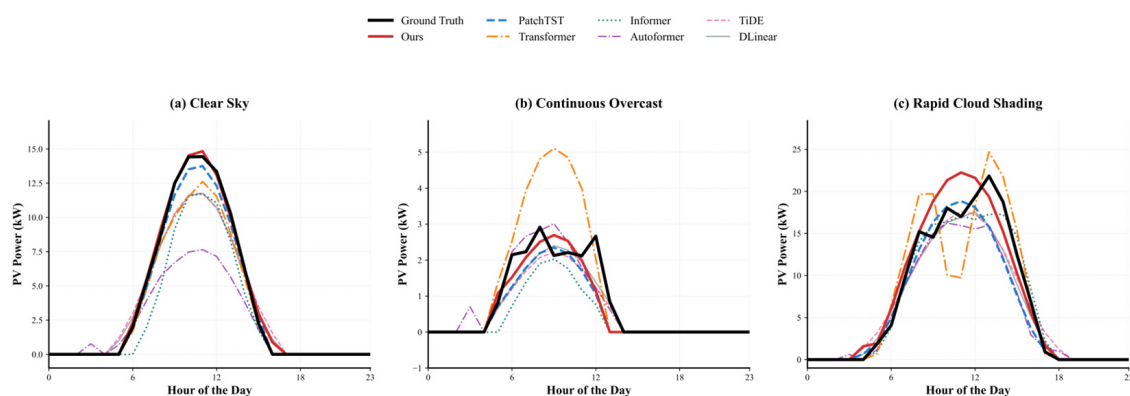


Figure 6. Prediction results of three typical weather conditions using different models on the Hong Kong photovoltaic dataset. Across clear-sky (a), overcast (b), and rapid-clouding (c) conditions, the proposed method consistently outpaces baselines in tracking transient fluctuations and recovering peak extremes.

Facing high-frequency non-stationary fluctuations caused by cloud movements (Figure 6c), Autoformer's temporal decomposition module fails. It generates numerous false peaks deviating from physical laws. Informer and Transformer barely track macro-fluctuations but exhibit persistent phase-lag phenomena. Their responses to sudden power drops and rebounds are delayed. The channel-independent PatchTST demonstrates peak-smoothing tendencies at the crests, frequently underestimating the actual extreme generation values during transient events.

To execute rigorous uncertainty quantification and evaluate the statistical robustness of the proposed architecture, a bootstrapping procedure with 1000 resampling iterations was executed on the HK dataset. The framework achieved a Mean Squared Error of 6.2493 with a standard deviation of 0.0687, corresponding to a 95 percent confidence interval of [6.1146, 6.3840]. Additionally, the Coefficient of Determination recorded a mean of 0.7945

with a confidence interval ranging from 0.7921 to 0.7969. These statistical boundaries objectively quantify the variance and prediction uncertainty of the model during different meteorological conditions.

Our model, benefiting from the CVA-reconstructed photothermal synergies and TDPL derivative constraints, tracks transient fluctuations closely and achieves accurate peak recovery, yielding the lowest MSE. This overcomes the dynamic limitations of traditional models under complex micro-meteorological conditions.

5.1.2. Performance Analysis on the DKASC Dataset

On the DKASC dataset, which features arid climate, large diurnal swings, and heterogeneous capacities, the performance gap between our model and the baselines widens. The PFILM module is particularly beneficial here, as capacity varies widely across the 49 arrays. Models without capacity normalization (Informer, Transformer) yield MSE above 29, indicating that relative features alone cannot reconcile kW-scale and tens-of-kW-scale arrays. As Table 4 shows, the MSE of the Transformer deteriorated to 30.1808. Informer regressed to 29.5556, and Autoformer remained at 25.6917 with a low overall goodness-of-fit. In comparison, the proposed PFILM framework demonstrated consistent predictive stability across the evaluated metrics. Its MSE stabilized at 20.9808, while its R^2 reached 0.9376.

Table 4. Comparison results of different models on the DKASC dataset. Bold values denote the best performance.

Model	MSE	MAE	RMSE	R^2
PatchTST	22.4796	0.9283	4.7413	0.9331
Informer	29.5556	1.5335	5.4365	0.9121
Autoformer	25.6917	1.1977	5.0687	0.9236
DLinear	22.6804	0.9400	4.7624	0.9325
Transformer	30.1808	1.4851	5.4937	0.9102
TiDE	22.6795	0.9400	4.7623	0.9325
Ours	20.9808	0.9029	4.5805	0.9376

Figure 7 shows the 72-h continuous weather transition prediction. Informer and Transformer exhibit significant deviations during daytime hours, with constrained response amplitudes that fall short of actual peaks. Autoformer reaches peak amplitudes but produces irregular distortions. Even PatchTST, relatively robust here, shows underfitting in daytime peak regions.

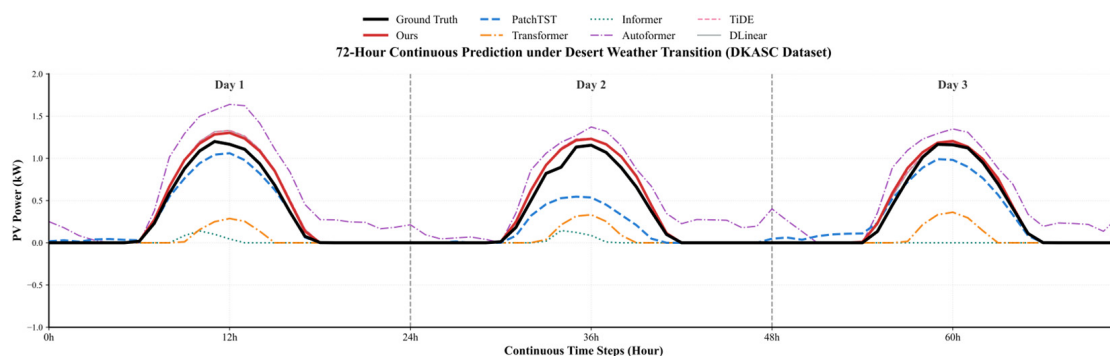


Figure 7. Continuous three-day prediction results of different models on the DKASC dataset. Under extreme radiation swings, baselines suffer from clipped amplitudes and distorted midday trajectories. Our framework delivers precise magnitude tracking and consistent spatial coverage.

Evaluation on the DKASC dataset similarly quantified the intrinsic uncertainty of the prediction framework. The procedure yielded a Mean Squared Error of 20.9808 and

a 95 percent confidence interval of [20.5396, 21.4220]. Furthermore, the Coefficient of Determination metric established an interval of [0.9362, 0.9390]. The provision of these bounded intervals alongside rigorous uncertainty quantification demonstrates the stable forecasting capability of the proposed method under complex environmental transitions.

In contrast, our model's prediction curves closely match actual values in macroscopic periodic phases (reflecting the high R^2 of 0.9376). Furthermore, they achieved strong magnitude coverage and envelopment at the absolute power peaks of each independent heterogeneous array. These results indicate that the PFiLM module maps static installed capacities into spatial anchors. Consequently, it enables the deep network to perform physical magnitude alignment across different spatial scales within diverse desert environments.

5.2. Ablation Study

The generalization performance of our model across multi-site datasets originates from the physical synergy of three core modules: CVA, PFiLM, and TDPL. This synergy is quantitatively confirmed via an ablation study on the Hong Kong dataset. Table 5 and the radar charts in Figure 8 illustrate the specific performance degradation following the removal of each component.

Table 5. Ablation experiment results on the Hong Kong dataset. Bold values denote the best performance.

Model	MSE	MAE	RMSE	R^2
PatchTST	7.9735	1.5904	2.8237	0.7378
Only PFiLM	7.8845	1.5704	2.8079	0.7407
Only CVA	8.0151	1.6292	2.8311	0.7364
Only TDPL	8.1497	1.6159	2.8548	0.7320
w/o CVA	7.9217	1.5837	2.8145	0.7395
w/o TDPL	6.2509	1.3968	2.5002	0.7944
w/o PFiLM	8.1891	1.6609	2.8617	0.7307
Ours	6.2493	1.3960	2.4999	0.7945

As Table 5 shows, the largest degradation occurs when the PFiLM module is removed: MSE increases by 31.0% (from 6.2493 to 8.1891) and R^2 drops by 8.0%. This drop indicates that, in multi-site training, standard data normalization alone is insufficient to reconcile output from stations with different installed capacities. The network presented without PFiLM lacks the spatial anchor needed to map relative features to absolute power magnitudes.

Removing CVA also degrades performance (MSE +26.8%, R^2 −6.9%), with the most noticeable impact on waveform shape. The model without CVA misses sharp turning points during abrupt irradiance changes, consistent with Figure 6c.

TDPL shows a modest effect on aggregate metrics (MSE +0.03%, R^2 −0.01%), but its impact is localized to transient events. During rapid ramps, the model without TDPL produces smoother transitions and lower peaks than the full model.

The three modules address distinct failure modes—magnitude misalignment (PFiLM), missing interactions (CVA), and oversmoothing (TDPL)—and their effects are largely additive. Addressing any single limitation in isolation yields limited gains, confirming that PV forecasting requires simultaneous handling of coupled meteorological, spatial, and temporal non-stationarities.

These phenomena verify that the physical limitations of PV forecasting are multidimensionally intertwined. Algorithmic patches targeting a single dimension struggle to achieve global optimization. A systematic architectural fusion encompassing the temporal, variable, and spatial domains stimulates the synergistic gains of deep neural networks. This enhances overall forecasting performance.

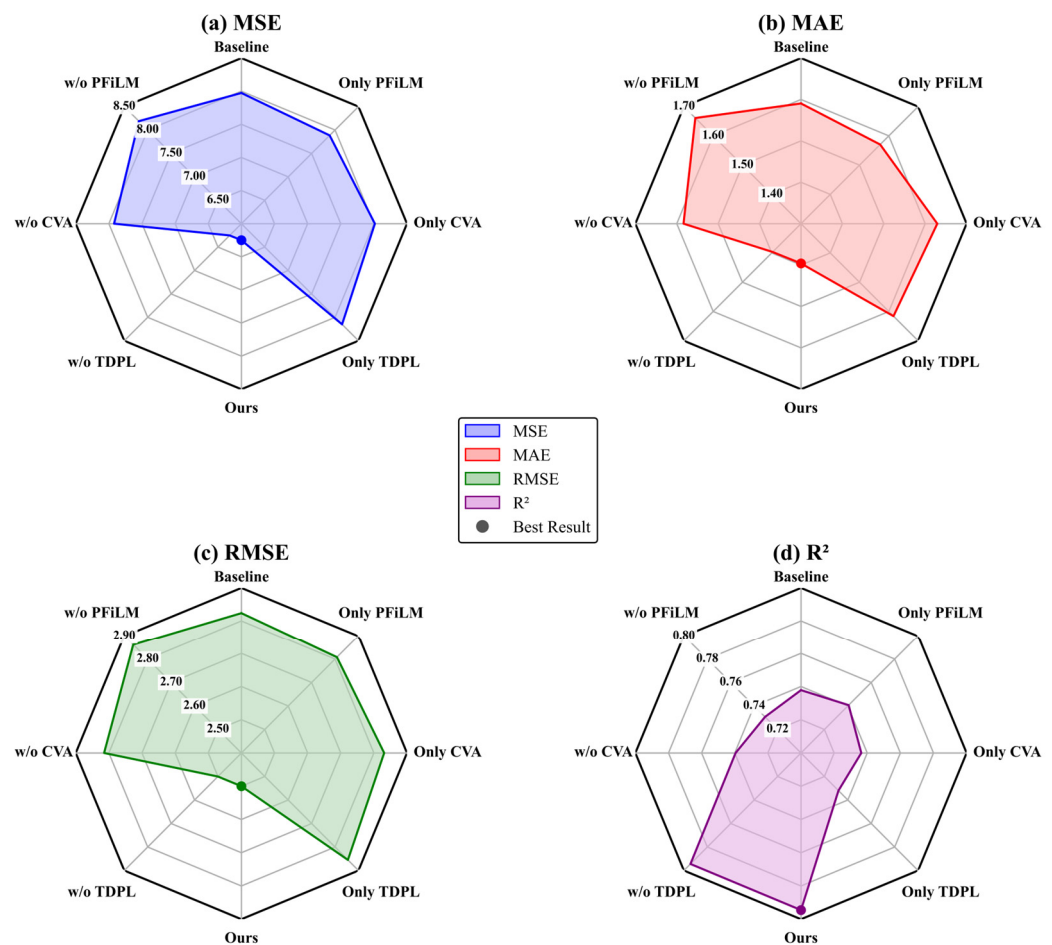


Figure 8. Radar image of ablation experiment on Hong Kong PV dataset. Panels (a–d) plot MSE, MAE, RMSE, and R^2 . Wider dispersion marks performance loss from ablating each module, confirming that full integration yields the best accuracy.

5.3. Parameter Sensitivity Experiment of the Historical Observation Window

The model's look-back length for historical sequences directly determines its receptive field scale in the temporal domain. Excessively short sequences tend to miss long-term trends and periodic patterns. Conversely, overly long sequences may introduce redundant climate noise and dilute core attention weights. As Table 6 shows, the parameter sensitivity test on the Hong Kong dataset reveals the quantitative metric evolutions across four sequence lengths: 72, 144, 288, and 336. For lengths of 72 and 144, MSE is 8.9004 and 8.2377, and R^2 is 0.7073 and 0.7291. At 288, accuracy peaks (MSE = 6.2493, R^2 = 0.7945). Increasing to 336 degrades performance (MSE = 6.5766, R^2 = 0.7837).

Table 6. Experimental results of parameter sensitivity on the Hong Kong dataset. Bold values denote the best performance.

Sequence Length	MSE	MAE	RMSE	R^2
Seq72	8.9004	1.6389	2.9834	0.7073
Seq144	8.2377	1.5884	2.8701	0.7291
Seq336	6.5766	1.4390	2.5645	0.7837
Seq288(Ours)	6.2493	1.3960	2.4999	0.7945

Figure 9 shows a U-shaped relationship between sequence length and error (and an inverted U-shape for R^2). Short sequences lack sufficient historical information to capture diurnal cycles and heat accumulation. At 288, the model captures recent photothermal

dynamics effectively. At 336, weakly correlated fluctuations and distant noise dilute attention weights and introduce interference. Therefore, we select 288 as the optimal look-back window for this task.

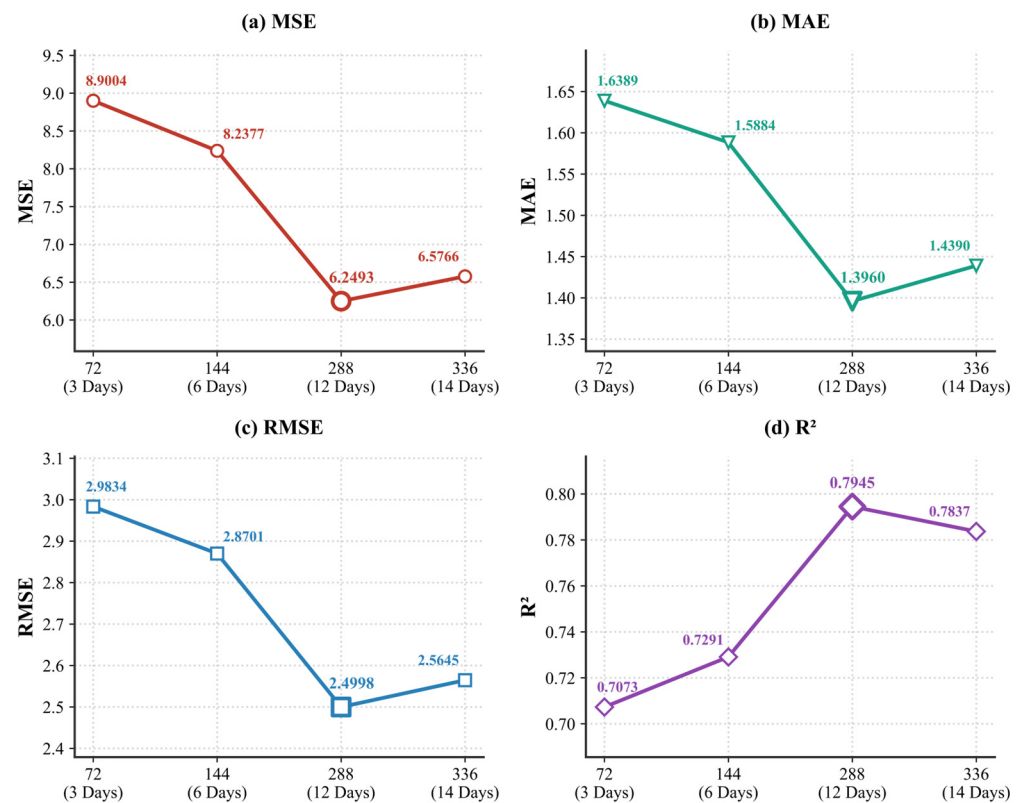


Figure 9. Experimental results of parameter sensitivity on the Hong Kong dataset. The four subplots show error trends over varying sequence lengths. Errors dip nonlinearly and bottom out at 288 steps—the optimal window that captures long-range dynamics without overfitting climate noise.

5.4. Model Complexity Analysis

To assess deployment feasibility on edge devices, we evaluated model complexity in terms of parameter count and FLOPs. Table 7 presents the results.

Table 7. Comparison of Complexity and Parameter Quantity of Different Models.

Model	FLOPs (M)	Parameters (K)
PatchTST	155.32	474.39
Informer	211.88	703.11
Autoformer	273.96	600.58
DLinear	0.23	13.87
Transformer	274.02	604.03
TiDE	12.58	332.18
Ours	164.04	551.36

Data comparisons in Table 7 reveal that linear or fully connected architectures, such as DLinear and TiDE, possess inherent baseline advantages in complexity. DLinear requires only 0.23 MFLOPs. However, its lack of modeling capability for complex nonlinear dynamics leads to higher forecasting errors. Among deep architectures, traditional heavy encoder–decoder models face computational overhead. The total computational complexities of Autoformer and Transformer reach up to 273.96 MFLOPs and 274.02 MFLOPs, respectively. Benefiting from the patching mechanism’s compression of lengthy time series,

our model maintains a total computational complexity of 164.04 MFLOPs. It improves forecasting accuracy while reducing computational overhead by approximately 40% compared to heavy architectures. Compared to PatchTST (155.32 MFLOPs, 474.39 K parameters), our model adds the CVA and PFiLM modules, increasing parameters to 551.36 K and FLOPs by 8.7 MFLOPs.

The bubble chart for model complexity analysis in Figure 10 provides a clearer observation of the relative relationship between accuracy and computational cost. The horizontal axis represents the parameter count, while the vertical axis represents the Mean Squared Error (MSE). The bubble size intuitively reflects the relative scale of computational complexity. As illustrated, our model is located in the lower-right area of the chart, indicating low error and high efficiency. Compared to the larger bubbles representing Autoformer and Informer, our model achieves lower errors along the vertical axis. Simultaneously, it exhibits a reduction in bubble area. This indicates that our study achieves multidimensional feature decoupling and physical prior knowledge injection with limited hardware overhead increments.

Overall, the proposed model offers a good balance between accuracy and computational efficiency, making it suitable for real-time deployment on low-cost edge devices.

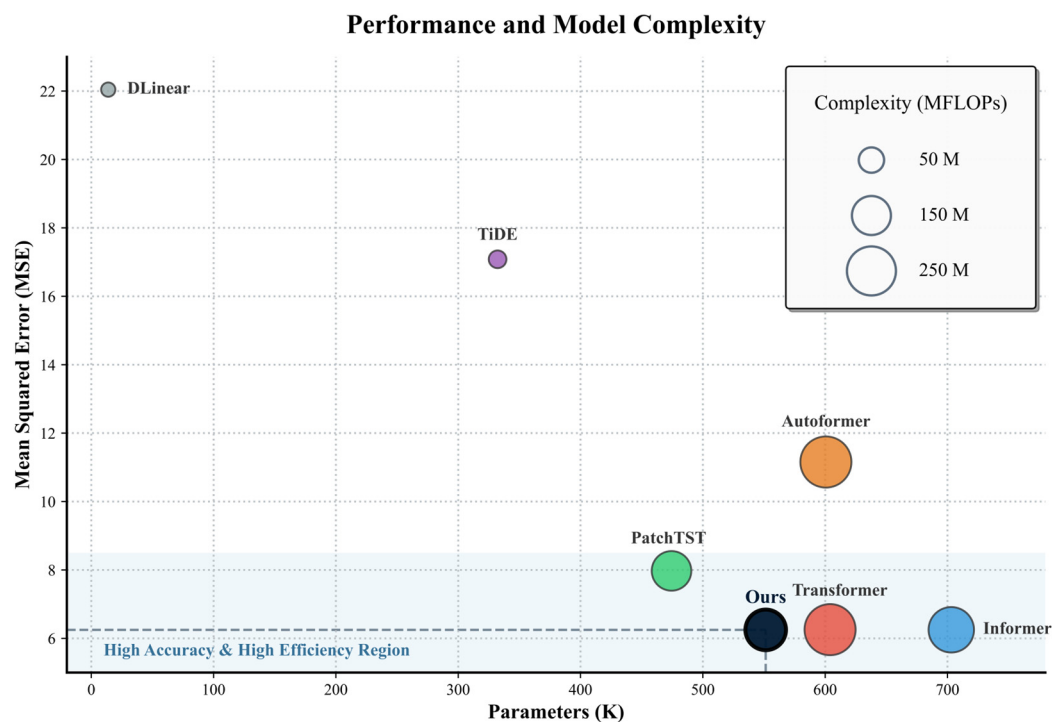


Figure 10. Bubble Chart Comparing Complexity and Parameter Quantity of Different Models. X-axis: parameter count; Y-axis: MSE; bubble size: FLOPs. Our model sits in the sweet spot—low error, moderate parameters, and modest FLOPs.

6. Conclusions

Addressing the inherent limitations of existing day-ahead photovoltaic energy forecasting models, this study proposes a Physics-Guided Cross-Variable Temporal Transformer that systematically decouples features across temporal, variable, and spatial domains. By integrating a Cross-Variable Attention mechanism with a Physical Feature-wise Linear Modulation network, the architecture successfully captures nonlinear photothermal synergistic dynamics and adaptively aligns physical magnitudes across heterogeneous stations using static installed capacity priors. The subsequent implementation of a Time Dynamics-aware Perceiving Loss function provides a gradient constraint, enabling the network to accurately

track high-frequency transient power ramping events. Extensive empirical evaluations on the Hong Kong and DKASC datasets confirm that this synergistic approach significantly outperforms mainstream baselines in both absolute accuracy and cross-domain robustness. Furthermore, the model maintains a highly efficient computational footprint of 164.04 MFLOPs, validating its practical feasibility for real-time edge computing deployments in microgrids. Moving forward, future research will focus on dynamically embedding refined equipment features, such as continuous PV module degradation patterns and non-linear inverter clipping constraints, to further enhance forecasting reliability over extended operational lifespans in complex micro-meteorological environments.

Author Contributions: Conceptualization, C.X. and M.C.; methodology, C.X.; validation, C.X., Y.W. and X.D.; data curation, C.X.; writing—original draft preparation, C.X.; writing—review and editing, C.X. and M.C.; visualization, C.X., Y.L. and Z.L.; supervision, M.C.; project administration, M.C.; funding acquisition, M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Sichuan Provincial Natural Science Foundation under Grant Nos. 2025ZNSFSC0477 and 2024NSFSC2042, the Opening Fund of the Intelligent Perception and Control Key Laboratory of Sichuan Province under Grant No. 2025IPCY02, and the Postgraduate Innovation Fund of Sichuan University of Science and Engineering under Grant No. Y2025090.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

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

References

1. Gaboitaolelwe, J.; Zungeru, A.M.; Yahya, A.; Lebekwe, C.K.; Vinod, D.N.; Salau, A.O. Machine learning based solar photovoltaic power forecasting: A review and comparison. *IEEE Access* **2023**, *11*, 40820–40845. [[CrossRef](#)]
2. Iheanetu, K.J. Solar photovoltaic power forecasting: A review. *Sustainability* **2022**, *14*, 17005. [[CrossRef](#)]
3. Denholm, P.; O’Connell, M.; Brinkman, G.; Jorgenson, J. *Overgeneration from Solar Energy in California. A Field Guide to the Duck Chart*; National Renewable Energy Lab. (NREL): Golden, CO, USA, 2015.
4. Zhan, Y.; Xie, X.; Liu, H.; Li, Y. Frequency-domain modal analysis of the oscillatory stability of power systems with high-penetration renewables. *IEEE Trans. Sustain. Energy* **2019**, *10*, 1534–1543. [[CrossRef](#)]
5. Agoua, X.G.; Girard, R.; Kariniotakis, G. Short-term spatio-temporal forecasting of photovoltaic power production. *IEEE Trans. Sustain. Energy* **2017**, *9*, 538–546.
6. Rajput, P.; Malvoni, M.; Manoj Kumar, N.; Sastry, O.S.; Jayakumar, A. Operational performance and degradation influenced life cycle environmental–economic metrics of mc-Si, a-Si and HIT photovoltaic arrays in hot semi-arid climates. *Sustainability* **2020**, *12*, 1075. [[CrossRef](#)]
7. Shukla, K.N.; Rangnekar, S.; Sudhakar, K. Comparative study of isotropic and anisotropic sky models to estimate solar radiation incident on tilted surface: A case study for Bhopal, India. *Energy Rep.* **2015**, *1*, 96–103. [[CrossRef](#)]
8. Stein, J.S.; Hansen, C.W.; Reno, M.J. *Global Horizontal Irradiance Clear Sky Models: Implementation and Analysis*; Sandia National Laboratories (SNL-NM): Albuquerque, NM, USA, 2012.
9. Akhsassi, M.; Erraissi, N.; Raoufi, M.; EL Fathi, A.; Zouine, M.; Ouboubker, L.; Outzourhit, A.; Bennouna, A. Experimental investigation and thermo-electrical performance modeling of two PV plants in arid climates. *Energy Convers. Manag.* **2024**, *321*, 119025. [[CrossRef](#)]
10. Yang, J.; Sengupta, M.; Jiménez, P.A.; Kim, J.-H.; Xie, Y. Evaluating WRF-Solar EPS cloud mask forecast using the NSRDB. *Sol. Energy* **2022**, *243*, 348–360. [[CrossRef](#)]
11. Bacher, P.; Madsen, H.; Nielsen, H.A. Online short-term solar power forecasting. *Sol. Energy* **2009**, *83*, 1772–1783. [[CrossRef](#)]
12. Mellit, A.; Kalogirou, S.A. Artificial intelligence techniques for photovoltaic applications: A review. *Prog. Energy Combust. Sci.* **2008**, *34*, 574–632. [[CrossRef](#)]

13. Pedro, H.T.C.; Coimbra, C.F.M. Assessment of forecasting techniques for solar power production with no exogenous inputs. *Sol. Energy* **2012**, *86*, 2017–2028. [\[CrossRef\]](#)
14. Voyant, C.; Notton, G.; Kalogirou, S.; Nivet, M.-L.; Paoli, C.; Motte, F.; Fouilloy, A. Machine learning methods for solar radiation forecasting: A review. *Renew. Energy* **2017**, *105*, 569–582. [\[CrossRef\]](#)
15. Li, X.; Ma, L.; Chen, P.; Xu, H.; Xing, Q.; Yan, J.; Lu, S.; Fan, H.; Yang, L.; Cheng, Y. Probabilistic solar irradiance forecasting based on XGBoost. *Energy Rep.* **2022**, *8*, 1087–1095. [\[CrossRef\]](#)
16. Çayır, A.; Yenidoğan, I.; Dağ, H. Feature extraction based on deep learning for some traditional machine learning methods. In *Proceedings of the 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, Bosnia and Herzegovina, 20–23 September 2018*; IEEE: New York, NY, USA, 2018; pp. 494–497.
17. Gensler, A.; Henze, J.; Sick, B.; Raabe, N. Deep Learning for solar power forecasting—An approach using AutoEncoder and LSTM Neural Networks. In *Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, Hungary, 9–12 October 2016*; IEEE: New York, NY, USA, 2016; pp. 002858–002865.
18. Zameer, A.; Jaffar, F.; Shahid, F.; Muneeb, M.; Khan, R.; Nasir, R. Short-term solar energy forecasting: Integrated computational intelligence of LSTMs and GRU. *PLoS ONE* **2023**, *18*, e0285410. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Sherstinsky, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys. D Nonlinear Phenom.* **2020**, *404*, 132306. [\[CrossRef\]](#)
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; NeurIPS Proceedings: La Jolla, CA, USA, 2017; Volume 30.
21. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 11106–11115. [\[CrossRef\]](#)
22. Wu, H.; Xu, J.; Wang, J.; Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 22419–22430.
23. Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv* **2022**, arXiv:2211.14730.
24. Zeng, A.; Chen, M.; Zhang, L.; Xu, Q. Are transformers effective for time series forecasting? *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 11121–11128. [\[CrossRef\]](#)
25. Badi, N.; Laatar, A.H.; Ben Mabrouk, A.; Alsharari, A.M.; Alghamdi, S.A.; Albalawi, H. Effects of nonlinear thermal radiation on the efficiency of building integrated photovoltaic systems with nanofluid cooling. *PLoS ONE* **2024**, *19*, e0304685. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Ono, L.K.; Raga, S.R.; Wang, S.; Kato, Y.; Qi, Y. Temperature-dependent hysteresis effects in perovskite-based solar cells. *J. Mater. Chem. A* **2015**, *3*, 9074–9080. [\[CrossRef\]](#)
27. Han, L.; Ye, H.J.; Zhan, D.C. The capacity and robustness trade-off: Revisiting the channel independent strategy for multivariate time series forecasting. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 7129–7142. [\[CrossRef\]](#)
28. Simeunović, J.; Schubnel, B.; Alet, P.J.; Carrillo, R.E. Spatio-temporal graph neural networks for multi-site PV power forecasting. *IEEE Trans. Sustain. Energy* **2021**, *13*, 1210–1220. [\[CrossRef\]](#)
29. Jeong, J.; Kim, H. Multi-site photovoltaic forecasting exploiting space-time convolutional neural network. *Energies* **2019**, *12*, 4490. [\[CrossRef\]](#)
30. Bhanja, S.; Das, A. Impact of data normalization on deep neural network for time series forecasting. *arXiv* **2018**, arXiv:1812.05519.
31. Kudrat, D.; Xie, Z.; Sun, Y.; Jia, T.; Hu, Q. Patch-wise structural loss for time series forecasting. *arXiv* **2025**, arXiv:2503.00877.
32. Wellby, S.J.; Engerer, N.A. Categorizing the meteorological origins of critical ramp events in collective photovoltaic array output. *J. Appl. Meteorol. Climatol.* **2016**, *55*, 1323–1344. [\[CrossRef\]](#)
33. Antonanzas, J.; Osorio, N.; Escobar, R.; Urraca, R.; Martinez-de-Pison, F.J.; Antonanzas-Torres, F. Review of photovoltaic power forecasting. *Sol. Energy* **2016**, *136*, 78–111. [\[CrossRef\]](#)
34. Di Leo, P.; Ciocia, A.; Malgaroli, G.; Spertino, F. Advancements and challenges in photovoltaic power forecasting: A comprehensive review. *Energies* **2025**, *18*, 2108. [\[CrossRef\]](#)
35. Wen, Q.; Zhou, T.; Zhang, C.; Chen, W.; Ma, Z.; Yan, J.; Sun, L. Transformers in time series: A survey. *arXiv* **2022**, arXiv:2202.07125.
36. Zhang, Y.; Yan, J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023*.
37. Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; Long, M. Itransformer: Inverted transformers are effective for time series forecasting. *Int. Conf. Learn. Represent.* **2024**, *2024*, 11116–11140.
38. Du, Y.; Fell, C.J.; Duck, B.; Chen, D.; Liffman, K.; Zhang, Y.; Gu, M.; Zhu, Y. Evaluation of photovoltaic panel temperature in realistic scenarios. *Energy Convers. Manag.* **2016**, *108*, 60–67. [\[CrossRef\]](#)
39. Başaran Filik, Ü.; Filik, T.; Gerek, Ö.N. A hysteresis model for fixed and sun tracking solar PV power generation systems. *Energies* **2018**, *11*, 603. [\[CrossRef\]](#)

40. Jiao, X.; Li, X.; Lin, D.; Xiao, W. A graph neural network based deep learning predictor for spatio-temporal group solar irradiance forecasting. *IEEE Trans. Ind. Inform.* **2021**, *18*, 6142–6149. [[CrossRef](#)]
41. Karniadakis, G.E.; Kevrekidis, I.G.; Lu, L.; Perdikaris, P.; Wang, S.; Yang, L. Physics-informed machine learning. *Nat. Rev. Phys.* **2021**, *3*, 422–440. [[CrossRef](#)]
42. Gao, X.; Zang, Y.; Ma, Q.; Liu, M.; Cui, Y.; Dang, D. A physics-constrained deep learning framework enhanced with signal decomposition for accurate short-term photovoltaic power generation forecasting. *Energy* **2025**, *326*, 136220. [[CrossRef](#)]
43. Lima, F.T.; Souza, V.M.A. A large comparison of normalization methods on time series. *Big Data Res.* **2023**, *34*, 100407. [[CrossRef](#)]
44. Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; Courville, A. Film: Visual reasoning with a general conditioning layer. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
45. Le Guen, V.; Thome, N. Shape and time distortion loss for training deep time series forecasting models. In *Advances in Neural Information Processing Systems; NeurIPS Proceedings*; La Jolla, CA, USA, 2019; Volume 32.
46. Zhang, J.; Zhu, X.; Xie, Y.; Chen, G.; Liu, S. Detection and prediction of wind and solar photovoltaic power ramp events based on data-driven methods: A critical review. *Energies* **2025**, *18*, 3290. [[CrossRef](#)]
47. Abuella, M.; Chowdhury, B. Forecasting solar power ramp events using machine learning classification techniques. In *Proceedings of the 2018 9th IEEE International Symposium on Power Electronics for Distributed Generation Systems (PEDG)*, Charlotte, NC, USA, 25–28 June 2018; IEEE: New York, NY, USA, 2018; pp. 1–6.
48. Lin, Z.; Zhou, Q.; Wang, Z.; Wang, C.; Bookhart, D.B.; Leung-Shea, M. A high-resolution three-year dataset supporting rooftop photovoltaics (PV) generation analytics. *Sci. Data* **2025**, *12*, 63. [[CrossRef](#)] [[PubMed](#)]
49. Das, A.; Kong, W.; Leach, A.; Mathur, S.; Sen, R.; Yu, R. Long-term forecasting with tide: Time-series dense encoder. *arXiv* **2023**, arXiv:2304.08424.

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.