This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2026.3746466

Sky Conditions-Driven Architectures for Short-Term PV Power Forecasting

Saloni Dhingra, *Graduate Student Member, IEEE,* Giambattista Gruosso, *Senior Member, IEEE,* and Giancarlo Storti Gajani, *Senior Member, IEEE*

***Abstract*—Accurate short-term PV power forecasting remains challenging under rapidly varying sky conditions. This paper proposes a regime-routed mixture-of-experts framework in which PV power sequences are classified into clear, partly cloudy, overcast, and ramp regimes using causal clear-sky variability and ramp-rate indicators. CNN–LSTM and Time Series Transformer backbones are evaluated with Adam, Lookahead, AdaBelief, and Adam with Adaptive Gradient Clipping, with optimizer configurations independently selected for each atmospheric regime. The resulting regime-specific experts are trained exclusively on their corresponding operating conditions and activated through a causal hard-routing strategy at forecast issuance time. Experiments on two utility-scale, one-minute-resolution PV datasets for 60-minute-ahead forecasting demonstrate consistent improvements over global single-optimizer baselines. Test-set RMSE is reduced by approximately 14–16% for CNN–LSTM models and 15–18% for Transformer models, with the largest gains observed under ramp and partly cloudy conditions. The results demonstrate that physically interpretable regime-aware specialization and regime-specific optimizer selection improve forecasting accuracy and robustness across heterogeneous PV operating conditions.**

***Index Terms*—PV power forecasting, mixture of experts, regime-aware modeling, optimizer benchmarking, time-series Transformer.**

I. Introduction

THE high penetration of photovoltaic (PV) generation in modern power systems necessitates accurate short-term forecasting to support unit commitment, reserve allocation, and grid operation. At minute-scale horizons, deep learning (DL) models effectively capture the nonlinear and rapidly changing dynamics of PV generation. While considerable attention has focused on forecasting architectures, the influence of optimizer selection on convergence, stability, and generalization remains comparatively underexplored in PV forecasting.

This work extends our previous ICCEP 2025 study [1], which compared optimization algorithms across CNN, MLP, LSTM, and GRU models under conventional global training. Here, we introduce a regime-aware forecasting framework based on CNN–LSTM and Time Series Transformer backbones together with a regime-routed mixture-of-experts (MoE) architecture that jointly exploits regime-specific forecasting experts and optimizer specialization.

The proposed framework is evaluated on two utility-scale PV datasets with one-minute temporal resolution from contrasting climatic regions (Colorado and Nevada) for 60-minute-ahead forecasting. Four representative optimizers—Adam, Lookahead, AdaBelief, and Adam with Adaptive Gradient Clipping (AGC)—are benchmarked using hyperparameters selected through Bayesian optimization. Forecasting performance is evaluated using root mean squared error (RMSE) and mean absolute percentage error (MAPE) together with regime-specific analyses.

PV forecasting has evolved from statistical approaches toward machine learning and deep learning methods that exploit exogenous meteorological information and hybrid modeling pipelines [2]–[5]. CNN-, recurrent-, and attention-based architectures, including LSTM, GRU, Temporal Fusion Transformer (TFT) [6], Informer [7], Autoformer [8], FEDformer [9], and PatchTST [10], have demonstrated strong forecasting capability across a range of prediction horizons. Likewise, advances in optimization algorithms, including AdamW [11], RAdam [12], RMSprop [13], and SGD with Nesterov momentum [14], have improved training stability and generalization. Nevertheless, systematic investigations of optimizer behavior across different PV operating regimes remain scarce. Moreover, although regime stratification based on clear-sky or clearness-index statistics is widely used to analyze forecasting performance [15], [16], most studies still report aggregate metrics that obscure regime-dependent behavior.

The main contributions of this work are threefold:

1) A strictly causal, physically interpretable regime-identification rule based on clear-sky and ramp-rate statistics, formulated as an explicit Boolean routing criterion.
2) A regime-routed mixture-of-experts framework with regime-specific forecasting experts and independently optimized training configurations, where both the optimizer and its hyperparameters are selected for each atmospheric regime through Bayesian optimization.
3) Extensive ablation studies, including regime-only training, a capacity-matched baseline, and hard-versus-soft routing comparisons, demonstrating that the observed improvements arise from regime-aware specialization rather than increased model capacity or routing flexibility.

The remainder of this paper is organized as follows. Section II describes the datasets, regime identification procedure, forecasting models, and optimization strategies. Section III presents the proposed regime-routed MoE framework and experimental methodology. Section IV reports the forecasting and robustness results, while Section V concludes the paper.

S. Dhingra, G. Gruosso and G. Storti Gajani are with Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria, p.zza L.da Vinci 32, 20133, Milano, Italy (e-mail: saloni.dhingra@polimi.it, giancarlo.stortigajani@polimi.it, giambattista.gruosso@polimi.it).

Authorized licensed use limited to: Politecnico di Milano. Downloaded on November 10,2026 at 15:50:54 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2026.3746466

II. Data, Regime Stratification, and Forecasting Framework

A. Dataset Description and Feature Selection

This study uses three years of 1-minute data from two utility-scale PV plants [17]: a high-altitude semi-arid site in Colorado (Dataset 1) and a hot-desert site in Nevada (Dataset 2). Only DC power p_t and plane-of-array irradiance g_t are used, since these signals are widely available in PV plant monitoring systems and support causal online forecasting.

Power is normalized by the site rating,

$$\tilde{p}_t = \frac{p_t}{P_{\text{rated}}}, \tag{1}$$

and irradiance is filtered using physical range checks. The data are chronologically split into training (2014–2015), validation (first half of 2016), and testing (second half of 2016). Normalization parameters are estimated only from the training data. Sliding windows of length $T = 60$ are used to predict $\tilde{p}_{t+\tau}$ with $\tau = 60$ minutes.

Table I summarizes the characteristics of both datasets after preprocessing and daytime filtering, including the number of retained samples, irradiance statistics, and DC power operating ranges.

TABLE I: Statistical summary of the PVDAQ datasets after daytime filtering and preprocessing.

Statistic	**Dataset 1 (Golden, CO)**	**Dataset 2 (Las Vegas, NV)**
Years analyzed	2014–2016	2014–2016
Sampling interval	1-minute	1-minute
Total raw samples	$\approx 1.58 \times 10^6$	$\approx 1.58 \times 10^6$
Daytime samples retained	9.41×10^5	9.63×10^5
Nighttime samples removed	40.4%	39.1%
Mean irradiance (W/m^2)	412.6	468.3
Std. irradiance (W/m^2)	298.4	315.7
Minimum irradiance (W/m^2)	10	10
Maximum irradiance (W/m^2)	1087	1124
Mean DC power (kW)	286.4	331.8
Std. DC power (kW)	241.7	268.5
Minimum DC power (kW)	0	0
Maximum DC power (kW)	1000	1200

B. Regime Identification via Clear-Sky and Ramp Statistics

Each input window is assigned to one of four regimes: clear, partly cloudy, overcast, or ramp. Regime labels are computed causally using only samples available up to forecast issuance time t.

Sky state is represented by the clear-sky index

$$k_t = \frac{g_t}{g_t^{\text{cs}}}, \tag{2}$$

where g_t^{cs} is obtained from the Ineichen–Perez clear-sky model [18]. Short-term variability is measured by the trailing standard deviation $\sigma_k(t)$ over $W = 60$ minutes. Ramp events are detected using the backward normalized power gradient

$$r_t = \frac{\tilde{p}_t - \tilde{p}_{t-\Delta}}{\Delta}, \tag{3}$$

with $\Delta = 5$ minutes.

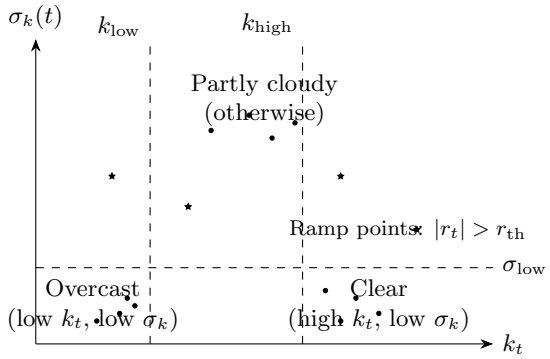


Fig. 1: Schematic representation of the regime stratification in the clear-sky index space. Clear and Overcast regimes are defined by the clear-sky index k_t together with the low-variability condition $\sigma_k(t) \leq \sigma_{\text{low}}$, while remaining non-ramp samples are assigned to Partly Cloudy. Ramp events are identified independently using the backward power gradient criterion $|r_t| > r_{\text{th}}$.

Clear, partly cloudy, and overcast regimes are defined using training-set quantile thresholds on k_t and $\sigma_k(t)$. Specifically, k_{low} and k_{high} are the 0.33 and 0.66 quantiles of k_t, while σ_{low} is the 0.66 quantile of $\sigma_k(t)$. Samples satisfying $|r_t| > r_{\text{th}}$ are assigned to the Ramp regime, where r_{th} is the 95th percentile of $|r_t|$ on the training set. Ramp identification takes precedence over sky-state classification. All thresholds are estimated using only the training set and then fixed for validation and testing, preventing data leakage.

Figure 1 illustrates the resulting regime stratification in the clear-sky index–variability space. Samples with $\sigma_k(t) \leq \sigma_{\text{low}}$ are classified as Clear or Overcast according to k_t, while the remaining non-ramp samples are assigned to Partly Cloudy. Ramp samples are identified independently using $|r_t| > r_{\text{th}}$. Applying this procedure to Dataset 1 yields approximately 27% Clear, 41% Partly Cloudy, 27% Overcast, and 5% Ramp samples, with similar proportions observed for Dataset 2.

1) Formal Regime Definition: The regime assignment is evaluated causally at each time t, with Ramp taking precedence over the sky-state classification:

$$\text{regime}(t) = \begin{cases} \textsc{Ramp}, & |r_t| > r_{th} \\ \text{sky-state}(t), & |r_t| \leq r_{th} \end{cases}$$

$$\text{sky-state}(t) = \begin{cases} \textsc{Clear}, & k_t \geq k_{\text{high}} \wedge \sigma_k(t) \leq \sigma_{\text{low}} \\ \textsc{Overcast}, & k_t \leq k_{\text{low}} \wedge \sigma_k(t) \leq \sigma_{\text{low}} \\ \textsc{PartlyCloudy}, & \text{otherwise} \end{cases}$$

C. Forecasting Models and Optimizers

Two forecasting backbones are evaluated. The CNN–LSTM extracts local temporal features from $\mathbf{X}_t \in \mathbb{R}^{T\times 2}$ using 1-D convolutions, followed by LSTM layers for temporal modeling:

$$\hat{\tilde{p}}_{t+\tau} = f_{\text{LSTM}}(f_{\text{CNN}}(\mathbf{X}_t)). \tag{4}$$

Authorized licensed use limited to: Politecnico di Milano. Downloaded on November 10,2026 at 15:50:54 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2026.3746466

TABLE II: Summary of optimization strategies used in this work.

Optimizer	Core mechanism	Main hyperparameters	Observed behavior in PV regimes
Adam	First- and second-moment adaptive learning-rate estimation	Learning rate α, β_1, β_2, ϵ	Fast convergence under smooth irradiance; reduced robustness under sharp ramp events
AdaBelief	Variance estimated from gradient deviation $(g_t - m_t)^2$	Learning rate α, β_1, β_2, ϵ	Improved stability and generalization under partly cloudy regimes
AGC (wrapper)	Gradient norm scaled relative to parameter norm	Clipping coefficient λ, base optimizer settings	Strong stabilization under extreme ramps and high-variability conditions
Lookahead	Periodic slow updates over fast inner optimizer	Sync steps k, interpolation factor β, base optimizer	Smoother loss trajectories; moderate benefit under mixed variability

The Transformer comprises a linear embedding, positional encoding, stacked self-attention encoder blocks, and global pooling:

$$\hat{\tilde{p}}_{t+\tau} = f_{\text{out}}(f_{\text{TR}}(\mathbf{X}_t)) \,. \tag{5}$$

Four optimizers are evaluated: Adam, Lookahead, AdaBelief, and Adam with Adaptive Gradient Clipping (AGC). Their optimization mechanisms, hyperparameters, and observed regime-wise behavior are summarized in Table II.

III. Regime-Routed Forecasting Framework and Experimental Protocol

A. Overall System Architecture

The proposed forecasting system comprises three components: (i) a regime identification module (Section II-B), (ii) four regime-specific forecasting experts, and (iii) a routing mechanism that assigns each input sequence to a single expert. CNN–LSTM and Time Series Transformer serve as alternative forecasting backbones, with each expert trained using its own optimizer configuration.

Each expert corresponds to one of four physically interpretable regimes: clear, partly cloudy, overcast, and ramp. The expert for regime $r \in \{1, \ldots, 4\}$ is denoted by $\mathcal{E}_r(\cdot; \boldsymbol{\theta}_r)$, where $\boldsymbol{\theta}_r$ denotes its trainable parameters.

For a given input sequence $\mathbf{X}_t$, the regime router produces a discrete routing decision

$$r_t = \mathcal{R}(\mathbf{X}_t), \tag{6}$$

where $\mathcal{R}(\cdot)$ implements the rule-based regime identification logic. The selected expert then produces the forecast

$$\hat{\tilde{p}}_{t+\tau} = \mathcal{E}_{r_t}(\mathbf{X}_t; \boldsymbol{\theta}_{r_t}). \tag{7}$$

Unlike conventional soft-gated MoE models, the proposed framework employs hard physical routing based on interpretable regime indicators. This eliminates trainable gating parameters, preserves strict causality, and provides a one-to-one correspondence between experts and operating regimes. Section IV-G compares this design with a soft-gated MoE, while the regime-only and capacity-matched ablations (Section IV-F and Table XII) isolate the contribution of regime-aware specialization.

B. Regime-Specific Expert Specialization

Each expert $\mathcal{E}_r$ is trained exclusively on samples belonging to its assigned regime. Let

$$\mathcal{D}_r = \{(\mathbf{X}_t, \tilde{p}_{t+\tau}) : r_t = r\} \tag{8}$$

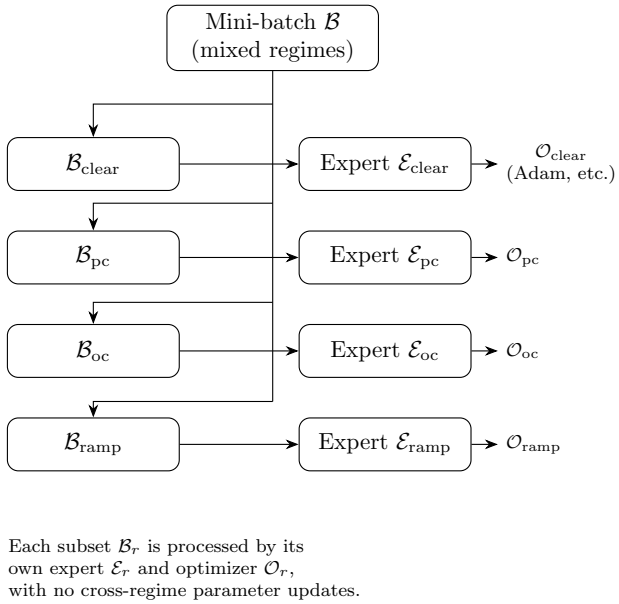


Fig. 2: Regime-partitioned training. Each mini-batch is divided into regime-specific subsets, and each subset updates only its corresponding expert and optimizer.

denote the subset of the training data associated with regime r. The empirical loss minimized by expert r is

$$\mathcal{L}_r(\boldsymbol{\theta}_r) = \frac{1}{|\mathcal{D}_r|} \sum_{(\mathbf{X}_t, \tilde{p}_{t+\tau}) \in \mathcal{D}_r} \ell(\mathcal{E}_r(\mathbf{X}_t; \boldsymbol{\theta}_r), \tilde{p}_{t+\tau}) \,, \tag{9}$$

where $\ell(\cdot)$ denotes the mean squared error.

This training strategy allows each expert to specialize to the statistical characteristics of its assigned atmospheric regime while preventing cross-regime parameter interference. Figure 2 illustrates the regime-partitioned mini-batch training procedure, where each subset is processed only by its corresponding expert and optimizer. Figure 3 shows a representative test day for Dataset 1, highlighting the causal regime assignments used for expert routing.

Expert selection is performed strictly at forecast issuance using only information available up to time t. The routing decision is based on historical clear-sky index statistics and backward ramp gradients, without access to future observations. Consequently, the selected expert reflects the operating conditions at forecast issuance rather than those that may occur during the prediction horizon, preserving strict causality.

Authorized licensed use limited to: Politecnico di Milano. Downloaded on November 10,2026 at 15:50:54 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2026.3746466

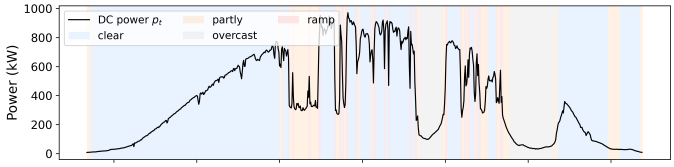


Fig. 3: Representative regime assignments for Dataset 1 (Colorado) on 2016–08–15. Measured DC power is shown with background shading indicating the causally identified atmospheric regimes used for expert routing.

C. Optimizer Specialization and Training Procedure

Each regime-specific expert is trained using its own optimizer instance and hyperparameter configuration. Let $\mathcal{O}_r$ denote the optimizer assigned to expert $\mathcal{E}_r$, where the candidate optimizers are Adam, Lookahead, AdaBelief, and AGC.

During training, each mini-batch is partitioned according to regime labels. For each regime r present in the batch, the corresponding subset $\mathcal{B}_r \subset \mathcal{B}$ is forwarded through expert $\mathcal{E}_r$, the loss is computed over $\mathcal{B}_r$, and parameters are updated using optimizer $\mathcal{O}_r$.

This regime-specific optimization ensures that each expert is updated exclusively with samples from its assigned regime, preventing cross-regime parameter interference while enabling specialization of both model parameters and optimizer dynamics.

D. Training Configuration and Hyperparameter Selection

All models are implemented in Python using PyTorch and trained with shuffled mini-batches. Input windows of length T are constructed from the normalized power and irradiance sequences $(\tilde{p}_t, g_t)$ to predict $\tilde{p}_{t+\tau}$.

Both CNN–LSTM and Transformer models minimize the mean squared error (MSE) loss

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{p}_{t_i+\tau} - \hat{\tilde{p}}_{t_i+\tau}\right)^2. \tag{10}$$

Early stopping based on validation loss is used to mitigate overfitting. In the proposed MoE framework, each expert is trained only on samples from its assigned regime, whereas global baselines are trained on the complete dataset.

Optimizer and architectural hyperparameters are selected using Bayesian optimization on the validation set. The search space includes learning rate, Lookahead synchronization parameters, AGC clipping coefficient, and architecture-specific parameters summarized in Table III. The optimization objective is validation RMSE, computed over the full validation set for global models and over the corresponding validation subset for each regime-specific expert.

Bayesian optimization is implemented in Optuna using the Tree-structured Parzen Estimator (TPE) surrogate with expected improvement. Each architecture–optimizer–regime combination is allocated 50 trials with median pruning, and identical search spaces and stopping criteria are used across all datasets and models to ensure fair comparison.

E. Evaluation Protocol and Baseline Models

Forecasting performance is evaluated using root mean squared error (RMSE), and mean absolute percentage error (MAPE), computed over daylight test samples after quality-control filtering.

Two global baselines are considered: a single CNN–LSTM and a single Time Series Transformer, each trained on the full dataset using a globally selected optimizer. Comparisons with these models isolate the benefits of regime specialization and optimizer adaptation.

Robustness is assessed across multiple independent training runs with different random seeds, emphasizing reproducibility and stability rather than per-sample hypothesis testing, which is inappropriate for temporally correlated minute-resolution PV time series.

The proposed MoE requires more model instances and hyperparameter optimization than global baselines because each regime-specific expert is tuned independently. However, the same Bayesian optimization procedure, search space, and trial budget are applied to every model, ensuring fair comparison while reflecting the additional complexity inherent to regime-specialized forecasting.

F. Causality and Leakage Prevention

Because regime labels are used for both regime-stratified evaluation and expert routing, all regime identification is performed strictly causally using only information available at forecast issuance.

At time t, every model receives the trailing input window $\mathbf{X}_t = \{(\tilde{p}_{t-T+1}, g_{t-T+1}), \ldots, (\tilde{p}_t, g_t)\}$, and predicts $\tilde{p}_{t+\tau}$. The router assigns a regime $r_t \in \{\text{clear}, \text{partly cloudy}, \text{overcast}, \text{ramp}\}$ using only past and present observations. Specifically, the clear-sky index $k_t = g_t/g_t^{\text{cs}}$, the trailing variability $\sigma_k(t)$, and the backward power gradient $r_t = (\tilde{p}_t - \tilde{p}_{t-\Delta})/\Delta$ are all computed exclusively from measurements available up to time t, ensuring that routing is implementable in real time without future information.

If irradiance samples required for computing $\sigma_k(t)$ are unavailable, missing values are imputed using only past observations. When the current POA irradiance measurement is unavailable, the ramp criterion is still evaluated from the power sequence; otherwise, the router retains the most recent valid sky-state assignment. Consequently, routing remains causal during temporary irradiance outages.

To assess robustness, irradiance measurements are perturbed with zero-mean Gaussian noise (5% and 10% standard deviation) and random five-minute daytime data gaps. Under these perturbations, global test RMSE increases by only 2–6%, while fewer than 5% of routing assignments change, indicating stable regime identification under moderate sensor imperfections.

All regime thresholds are estimated exclusively from the training data and then fixed for validation and testing. Specifically, k_{low}, k_{high}, σ_{low}, and r_{th} are derived from training-set quantiles as described in Section II-B; no validation or test information is used for threshold estimation or post-hoc adjustment.

Authorized licensed use limited to: Politecnico di Milano. Downloaded on November 10,2026 at 15:50:54 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2026.3746466

TABLE III: Hyperparameter search space used during Bayesian optimization for all architectures and optimizers.

Category	**CNN–LSTM**	**Transformer**	**Optimizers**
Learning rate	$[10^{-5}, 10^{-2}]$	$[10^{-5}, 10^{-2}]$	$[10^{-5}, 10^{-2}]$
Hidden units / Filters	Conv filters: $\{16, 32, 64\}$ LSTM units: $\{32, 64, 128\}$	Embedding dim: $\{32, 64, 128\}$ FFN dim: $\{64, 128, 256\}$	–
Depth (layers)	Conv layers: $\{1, 2, 3\}$ LSTM layers: $\{1, 2\}$	Encoder blocks: $\{2, 3, 4\}$	–
Dropout rate	$\{0.1, 0.2, 0.3\}$	$\{0.1, 0.2, 0.3\}$	–
Lookahead parameters	–	–	Sync steps $k \in [5, 10]$ Interpolation $\beta \in [0.1, 0.5]$
AGC parameters	–	–	Clipping $\lambda \in [10^{-3}, 10^{-1}]$
Adam/AdaBelief parameters	–	–	$\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ (fixed)

Finally, all global baselines use the same input features, forecasting horizon, train/validation/test split, and Bayesian hyperparameter-selection protocol as the proposed MoE framework. Performance improvements are therefore attributed to regime specialization and regime-specific optimizer selection rather than differences in data access or experimental protocol.

IV. Results and Discussion

This section evaluates the proposed regime-routed MoE framework on both datasets using CNN–LSTM and Transformer backbones with four optimizers (Adam, Lookahead, AdaBelief, and AGC). The analysis considers: (i) global forecasting performance without regime specialization, (ii) regime-wise optimizer behavior, and (iii) MoE performance using the best optimizer for each regime.

A. Global Forecasting Performance Without Regime Experts

Table IV summarizes the performance of the globally trained CNN–LSTM and Transformer models. MAPE is computed only for daylight samples with normalized power above $\epsilon = 10^{-3}$ to avoid instability near zero output.

Across both datasets and all optimizers, CNN–LSTM consistently outperforms the Transformer, reflecting its ability to capture local temporal dependencies and short-term irradiance fluctuations characteristic of minute-resolution PV forecasting.

Among the optimizers, AdaBelief achieves the best overall performance for both architectures, although differences between optimizers remain relatively small when evaluated over the complete operating range. These aggregate results mask substantial regime-dependent behavior, where optimizer performance varies across clear, partly cloudy, overcast, and ramp conditions. The following subsection therefore presents a regime-wise analysis to identify the optimizer most suited to each atmospheric regime.

B. Regime-Wise Performance Across Optimizers and Architectures

Regime labels are assigned strictly causally using only measurements available at forecast issuance, ensuring that the reported regime-wise results do not benefit from future information (Section III-F).

TABLE IV: Global deterministic forecasting performance without regime specialization.

Arch.	Opt.	D1		D2	
		RMSE	MAPE	RMSE	MAPE
CNN–LSTM	Adam	0.0656	2.9	0.0714	3.4
	Lookahead	0.0686	3.1	0.0728	3.6
	AdaBelief	**0.0640**	**2.6**	**0.0678**	**3.1**
	AGC	0.0656	2.8	0.0700	3.3
Transformer	Adam	0.0735	3.5	0.0787	4.0
	Lookahead	0.0755	3.8	0.0800	4.3
	AdaBelief	0.0728	3.4	0.0768	3.8
	AGC	0.0768	3.9	0.0812	4.4

Table V presents the detailed regime-wise forecasting performance for Dataset 1. Dataset 2 exhibits the same qualitative behavior, including the consistent ordering of forecasting difficulty, together with the same optimizer preferences within each regime. Therefore, the complete regime-wise results for Dataset 2 are omitted for brevity.

First, forecasting difficulty follows a consistent ordering across both datasets. Clear-sky conditions are the easiest because of their smooth and predictable irradiance profiles, whereas ramp regimes remain the most challenging owing to rapid irradiance transitions. Partly cloudy conditions exhibit intermediate difficulty because of intermittent cloud passages, while overcast conditions remain challenging due to persistently low irradiance levels and reduced signal strength.

The increase in forecasting error under ramp conditions is moderate rather than extreme. For a fixed architecture, dataset, and optimizer, the Ramp-regime RMSE (Table V) is typically about 1.3–1.4$\times$ the corresponding global RMSE (Table IV), indicating that the aggregate performance reflects a mixture of relatively easy Clear conditions (approximately 27% of samples) and substantially more challenging Ramp conditions (approximately 5%).

Using the site-rated capacities ($P_{\text{rated}} = 1\,\text{MW}$ for Dataset 1 and $1.2\,\text{MW}$ for Dataset 2), the best global CNN–LSTM baseline corresponds to approximately 64–81 kW RMSE, while the corresponding Transformer baselines range from approx-

Authorized licensed use limited to: Politecnico di Milano. Downloaded on November 10,2026 at 15:50:54 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2026.3746466

TABLE V: Regime-wise deterministic forecasting performance for Dataset 1 (Colorado) across architectures and optimizers.

Regime	Architecture	Optimizer	RMSE	MAPE (%)
Clear	CNN–LSTM	Adam	0.0529	2.0
Clear	CNN–LSTM	Lookahead	0.0548	2.2
Clear	CNN–LSTM	AdaBelief	**0.0490**	**1.7**
Clear	CNN–LSTM	AGC	0.0510	1.9
Clear	Transformer	Adam	0.0600	2.5
Clear	Transformer	Lookahead	0.0608	2.7
Clear	Transformer	AdaBelief	**0.0574**	**2.3**
Clear	Transformer	AGC	0.0616	2.8
Partly cloudy	CNN–LSTM	Adam	0.0707	3.2
Partly cloudy	CNN–LSTM	Lookahead	0.0728	3.4
Partly cloudy	CNN–LSTM	AdaBelief	**0.0678**	**2.8**
Partly cloudy	CNN–LSTM	AGC	0.0693	3.0
Partly cloudy	Transformer	Adam	**0.0775**	**3.7**
Partly cloudy	Transformer	Lookahead	0.0794	3.9
Partly cloudy	Transformer	AdaBelief	0.0800	4.0
Partly cloudy	Transformer	AGC	0.0812	4.2
Overcast	CNN–LSTM	Adam	0.0735	3.9
Overcast	CNN–LSTM	Lookahead	**0.0707**	**3.5**
Overcast	CNN–LSTM	AdaBelief	0.0728	3.8
Overcast	CNN–LSTM	AGC	0.0742	3.9
Overcast	Transformer	Adam	**0.0819**	**4.3**
Overcast	Transformer	Lookahead	0.0831	4.5
Overcast	Transformer	AdaBelief	0.0837	4.6
Overcast	Transformer	AGC	0.0849	4.8
Ramp	CNN–LSTM	Adam	0.0860	4.5
Ramp	CNN–LSTM	Lookahead	0.0877	4.7
Ramp	CNN–LSTM	AdaBelief	**0.0831**	**4.2**
Ramp	CNN–LSTM	AGC	0.0837	4.3
Ramp	Transformer	Adam	0.0943	5.3
Ramp	Transformer	Lookahead	0.0964	5.5
Ramp	Transformer	AdaBelief	0.0954	5.4
Ramp	Transformer	AGC	**0.0927**	**5.1**

imately 73–97 kW. These values provide a direct physical interpretation of the normalized forecasting errors.

Optimizer performance is strongly regime dependent. AdaBelief performs best under Clear and Partly Cloudy conditions, Lookahead is most effective under Overcast conditions, and AGC provides the best Transformer performance during Ramp events. For the CNN–LSTM architecture, AdaBelief consistently achieves the lowest RMSE across most regimes, including Ramp conditions.

Although Transformers capture long-range temporal dependencies, CNN–LSTM consistently outperforms them across all operating regimes for the 60-minute forecasting horizon, suggesting that local temporal feature extraction is more effective for this forecasting task. Transformer performance is also more sensitive to optimizer selection, particularly under highly variable ramp conditions.

Overall, forecasting accuracy depends jointly on atmospheric regime, network architecture, and optimizer, motivating the proposed regime-routed mixture-of-experts framework with regime-specific specialization.

C. Global Performance Under Regime-Routed Optimizer Specialization

Table VI summarizes the optimizer assignments selected for each atmospheric regime together with the resulting global

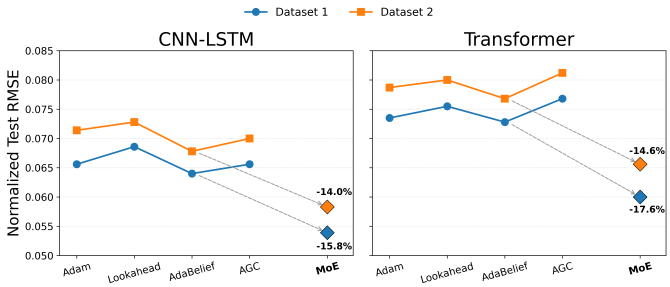


Fig. 4: Global forecasting performance of the single-optimizer baselines and the proposed regime-routed mixture-of-experts (MoE). Curves show the RMSE obtained when a single optimizer is used across all operating conditions, whereas the MoE marker represents the proposed regime-specific optimizer specialization.

forecasting performance obtained by aggregating predictions across all regimes.

For the CNN–LSTM architecture, AdaBelief is consistently selected for the Clear, Partly Cloudy, and Ramp regimes, whereas Lookahead performs best under Overcast conditions. In contrast, the Transformer favors AdaBelief under Clear conditions, Adam under Partly Cloudy and Overcast conditions, and AGC during Ramp events, reflecting the greater sensitivity of attention-based architectures to highly variable operating conditions.

These regime-specific optimizer assignments consistently outperform the corresponding single-optimizer baselines across both datasets. For the CNN–LSTM, the proposed MoE reduces the global RMSE from 0.0640 to 0.0539 on Dataset 1 (15.8%) and from 0.0678 to 0.0583 on Dataset 2 (14.0%). Similarly, the Transformer-based MoE reduces the RMSE from 0.0728 to 0.0600 (17.6%) on Dataset 1 and from 0.0768 to 0.0656 (14.6%) on Dataset 2. These consistent improvements indicate that optimizer effectiveness depends jointly on atmospheric regime and forecasting architecture, and that regime-aware optimizer specialization provides a practical means of improving forecasting accuracy without increasing model complexity or sacrificing interpretability. Figure 4 provides a visual comparison of the global forecasting performance obtained using each optimizer and the proposed regime-routed mixture-of-experts (MoE). The MoE consistently achieves the lowest RMSE on both datasets for both forecasting architectures, highlighting the benefit of regime-specific optimizer specialization over any single global optimizer.

Figure 5 illustrates a representative high-variability, mixed-regime test day. Compared with the global CNN–LSTM baseline, the proposed MoE more accurately tracks rapid power fluctuations and adapts to transitions between atmospheric regimes, consistent with the reductions in both regime-wise and global RMSE reported in Table VI.

D. Performance under In-Horizon Regime Transitions

Because the regime router uses only information available at forecast issuance time t (Section III-F), it cannot anticipate ramp events that begin within the prediction horizon $(t, t+$

Authorized licensed use limited to: Politecnico di Milano. Downloaded on November 10,2026 at 15:50:54 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2026.3746466

TABLE VI: Comparison of global (single-optimizer) and regime-routed (MoE) strategies. Regime columns list the optimizer used in each regime; performance columns report global test-set RMSE for each dataset.

		Optimizer used per regime				Dataset 1 RMSE	Dataset 2 RMSE
Architecture	Strategy	Clear	Partly-Cloudy	Overcast	Ramp	RMSE-D1	RMSE-D2
CNN–LSTM	Global Adam	Adam	Adam	Adam	Adam	0.0656	0.0714
CNN–LSTM	Global Lookahead	Lookahead	Lookahead	Lookahead	Lookahead	0.0686	0.0728
CNN–LSTM	Global AdaBelief	AdaBelief	AdaBelief	AdaBelief	AdaBelief	**0.0640**	**0.0678**
CNN–LSTM	Global AGC	AGC	AGC	AGC	AGC	0.0656	0.0700
CNN–LSTM	MoE (best per regime)	AdaBelief	AdaBelief	Lookahead	AdaBelief	**0.0539**	**0.0583**
Transformer	Global Adam	Adam	Adam	Adam	Adam	0.0735	0.0787
Transformer	Global Lookahead	Lookahead	Lookahead	Lookahead	Lookahead	0.0755	0.0800
Transformer	Global AdaBelief	AdaBelief	AdaBelief	AdaBelief	AdaBelief	0.0728	0.0768
Transformer	Global AGC	AGC	AGC	AGC	AGC	0.0768	0.0812
Transformer	MoE (best per regime)	AdaBelief	Adam	Adam	AGC	**0.0600**	**0.0656**

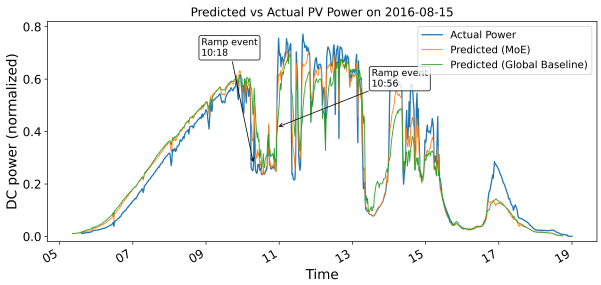


Fig. 5: Actual versus predicted normalized PV power on 2016-08-15, a day characterized by multiple ramp events. The proposed regime-aware CNN-LSTM mixture-of-experts (MoE) demonstrates improved tracking of rapid power drops and recoveries compared to a global CNN-LSTM baseline, especially around annotated ramp events.

TABLE VII: Test RMSE for non-ramp-issued samples, with and without an in-horizon ramp transition.

Arch.	**Data**	**Global (No-Trans/In-Horiz.)**	**MoE (No-Trans/In-Horiz.)**
CNN–LSTM	D1	0.0616 / 0.0900	0.0510 / 0.0837
CNN–LSTM	D2	0.0656 / 0.0943	0.0548 / 0.0877
Transformer	D1	0.0700 / 0.0975	0.0566 / 0.0917
Transformer	D2	0.0742 / 0.1010	0.0616 / 0.0949

$\tau]$. To quantify this limitation, test samples issued under non-ramp conditions (Clear, Partly Cloudy, or Overcast) are divided into two subsets: (i) no in-horizon ramp event ($|r_{t'}| \leq r_{\text{th}}$ for all $t' \in (t, t+\tau]$), and (ii) at least one in-horizon ramp event ($|r_{t'}| > r_{\text{th}}$ for some $t' \in (t, t+\tau]$). This analysis uses future information only for post-hoc evaluation; routing remains strictly causal.

Both approaches exhibit higher RMSE when a ramp develops within the forecast horizon (Table VII), since neither has causal access to the impending transition. Nevertheless, the proposed MoE consistently outperforms the corresponding global baseline across both datasets and architectures. The improvement is most pronounced for samples without an in-horizon transition, where RMSE is reduced by approximately 16–19%, indicating that regime-specialized experts are most effective when the operating regime remains unchanged throughout the prediction horizon. When an unexpected ramp occurs, the improvement decreases to approximately 6–7%, reflecting the inherent difficulty of forecasting transitions that are not observable at forecast issuance.

This behavior illustrates an inherent limitation of strictly causal regime specialization rather than a weakness of the proposed framework. The Clear, Partly Cloudy, and Overcast experts are trained exclusively on their respective operating conditions (Section III-B) and therefore have limited exposure to the onset of ramp events, whereas the global model encounters such transitions during pooled training. Even under these challenging conditions, the MoE consistently achieves lower RMSE than the global baseline, demonstrating that regime-aware specialization remains beneficial despite unforeseen changes in atmospheric conditions.

These observations suggest two promising directions for future work: (i) incorporating a short-horizon nowcasting indicator based on recent trend or acceleration features (e.g., dr_t/dt) to anticipate emerging ramp events before they occur, and (ii) replacing hard routing with probabilistic expert blending based on learned regime-transition probabilities.

E. Robustness Across Random Initializations

Each optimizer–architecture configuration was trained using five random seeds to assess robustness with respect to random initialization. Table VIII reports the mean and standard deviation of the test RMSE. The optimizer ranking remains consistent across seeds, with AdaBelief achieving the lowest mean RMSE, while AGC exhibits comparable performance with stable convergence characteristics under high-variability conditions.

To further evaluate robustness, the same five-seed protocol was applied to the best global baseline and the proposed MoE across all four architecture–dataset combinations. A one-tailed paired Wilcoxon signed-rank test was used to evaluate the hypothesis that the MoE reduces RMSE. The proposed MoE achieved lower RMSE than the corresponding global baseline for every seed in all four configurations, yielding $p = 0.03125$ in each case, which is the smallest attainable one-tailed p-value for five paired observations. These results indicate that the observed improvements are consistent across

Authorized licensed use limited to: Politecnico di Milano. Downloaded on November 10,2026 at 15:50:54 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2026.3746466

TABLE VIII: Multi-seed test-set RMSE statistics across five random seeds (CNN–LSTM, Dataset 1).

Optimizer	Mean RMSE	Std. Dev. RMSE
AdaBelief	**0.0640**	0.00047
Adam	0.0678	0.00052
Lookahead	0.0686	0.00058
AGC + Adam	0.0663	0.00053

TABLE IX: Ablation study for CNN–LSTM architecture.

Configuration	**D1 RMSE**	**D2 RMSE**
Global model (AdaBelief)	0.0640	0.0678
Regime-only (AdaBelief in all regimes)	0.0583	0.0624
Full MoE (regime-specific optimizers)	**0.0539**	**0.0583**

random initializations, although the limited number of seeds necessarily restricts the statistical power of the analysis. Full per-seed results are omitted for brevity.

F. Effect of Regime and Optimizer Specialization

To isolate the contributions of regime routing and optimizer specialization, three configurations were evaluated: (i) a global model trained with a single optimizer, (ii) regime-specific experts sharing the same optimizer (AdaBelief), and (iii) the full MoE with regime-specific optimizer selection.

Tables IX and X report the resulting global test-set RMSE for the CNN–LSTM and Transformer architectures. For both architectures and datasets, regime-specific training alone consistently reduces RMSE relative to the corresponding global baseline, indicating that partitioning the training data according to atmospheric regimes mitigates cross-regime interference. Introducing regime-specific optimizer selection yields a further reduction in RMSE, demonstrating that optimizer specialization provides additional benefits beyond regime partitioning alone. The improvement is particularly evident for the CNN–LSTM architecture, although consistent gains are also observed for the Transformer. Overall, these results show that the performance improvements arise from the complementary effects of regime-aware specialization and optimizer selection rather than from increased model capacity alone.

G. Hard versus Soft Routing

To validate the hard-routing design, we compare it with a soft-gated MoE in which a lightweight two-layer MLP, using the same causal routing features (k_t, $\sigma_k(t)$, and r_t), produces softmax weights over the four regime-specific experts. The forecast is given by

$$\hat{\tilde{p}}_{t+\tau} = \sum_{r=1}^{4} w_r(t)\, E_r(X_t; \theta_r), \tag{11}$$

$$w(t) = \mathrm{softmax}(g(k_t, \sigma_k(t), r_t))\,. \tag{12}$$

Here, $g(\cdot)$ denotes the gating network, while the expert parameters θ_r are fixed to those obtained in Section III-B, ensuring that only the routing strategy differs from the proposed hard-routed MoE. The gate is trained on the training set using the same MSE loss. Table XI compares the global test RMSE of the hard- and soft-routed variants for both architectures and datasets. Forecast continuity is assessed using the mean forecast-jump magnitude, $|\hat{\tilde{p}}_{t+\tau} - \hat{\tilde{p}}_{(t-1)+\tau}|$, computed over all regime transitions in the test set.

TABLE X: Ablation study for Transformer architecture.

Configuration	**D1 RMSE**	**D2 RMSE**
Global model (AdaBelief)	0.0728	0.0768
Regime-only (AdaBelief in all regimes)	0.0648	0.0700
Full MoE (regime-specific optimizers)	**0.0600**	**0.0656**

TABLE XI: Hard-routed vs. soft-gated MoE: global test RMSE and mean forecast-jump magnitude at regime transitions.

Architecture	**Dataset**	**RMSE (Hard)**	**RMSE (Soft)**	**Mean Jump (Hard/Soft)**
CNN–LSTM	D1	**0.0539**	0.0574	0.0138 / 0.0101
CNN–LSTM	D2	**0.0583**	0.0600	0.0159 / 0.0122
Transformer	D1	**0.0600**	0.0632	0.0151 / 0.0103
Transformer	D2	**0.0656**	0.0693	0.0169 / 0.0139

The soft-gated variant is consistently less accurate than the proposed hard-routed MoE, with global RMSE increasing by approximately 3–7% across all architecture–dataset combinations, indicating that learned gating provides no forecasting advantage in this setting. The penalty is slightly larger for the Transformer than for the CNN–LSTM, consistent with its greater sensitivity to the routing strategy. Soft gating reduces the mean forecast-jump magnitude at regime transitions by 18–32%, but the hard-routed forecasts remain sufficiently smooth, with a mean jump not exceeding 0.017 (less than 2% of the normalized power range). Since the modest improvement in forecast continuity does not compensate for the loss in forecasting accuracy, and hard routing preserves a simple, causal, and physically interpretable expert assignment (Section III-A), the hard-routed design is adopted in the proposed framework.

H. Capacity-Matched Baseline Analysis

To determine whether the observed performance gains arise primarily from increased model capacity rather than regime-aware specialization, a capacity-matched global CNN–LSTM baseline with approximately the same number of trainable parameters as the proposed MoE was constructed.

As shown in Table XII, increasing the capacity of the global CNN–LSTM reduces the test RMSE relative to the standard baseline, confirming that additional model capacity provides a measurable benefit. However, the capacity-matched model remains consistently less accurate than the proposed MoE on both datasets. Specifically, the capacity-matched model achieves RMSE values of 0.0608 and 0.0648 on Datasets 1 and 2, respectively, compared with 0.0539 and 0.0583 for the proposed MoE. These results indicate that the observed improvements are primarily attributable to regime-aware specialization and optimizer specialization, rather than to increased parameter count alone.

Authorized licensed use limited to: Politecnico di Milano. Downloaded on November 10,2026 at 15:50:54 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2026.3746466

TABLE XII: Effect of model capacity versus regime-aware specialization for CNN–LSTM architecture.

Model	D1 RMSE	D2 RMSE
Global CNN–LSTM (standard)	0.0640	0.0678
Global CNN–LSTM (capacity-matched)	0.0608	0.0648
MoE CNN–LSTM (regime-specialized)	**0.0539**	**0.0583**

TABLE XIII: Comparison with conventional and recent forecasting methods.

Method	D1 RMSE	D2 RMSE	Inf. (ms)
Naive persistence	0.0959	0.0990	0.02
Smart persistence	0.0781	0.0831	0.03
SARIMA	0.0866	0.0900	0.80
CNN–LSTM	0.0640	0.0678	0.28
Transformer	0.0728	0.0768	0.35
PatchTST	0.0574	0.0640	0.51
PatchTST + ACI	0.0583	0.0632	0.57
NGBoost + CNN–BiLSTM–Attn.	0.0529	0.0583	1.18
CNN–LSTM MoE	**0.0539**	**0.0583**	**0.22**
Transformer MoE	**0.0600**	**0.0656**	**0.29**

I. Comparison with Conventional and Recent Forecasting Baselines

The proposed framework is compared with conventional operational baselines and recent machine-learning and deep-learning methods using the same held-out test set and a 60-minute forecasting horizon. Learning-based methods use identical input variables, preprocessing, lookback window, and validation-based hyperparameter tuning. For PatchTST, tuning includes the patch length, embedding dimension, Transformer depth, and dropout rate. For PatchTST+ACI, only the central predictive estimate is used when computing deterministic RMSE, since this study focuses on point forecasting.

The benchmark includes naive persistence, smart clear-sky persistence, site-specific SARIMA, PatchTST [19], PatchTST with Adaptive Conformal Inference (ACI) [20], and NGBoost + CNN–BiLSTM–Attention [21].

Table XIII summarizes the test-set RMSE and per-sample inference latency. As expected, all deep-learning approaches outperform the conventional persistence and SARIMA baselines. Among the recent learning-based methods, the proposed CNN–LSTM MoE achieves competitive forecasting accuracy while requiring substantially lower inference time than more computationally intensive architectures such as NGBoost + CNN–BiLSTM–Attention. Although the latter achieves a marginally lower RMSE on Dataset 1 and identical RMSE on Dataset 2, the proposed MoE offers a favorable trade-off between forecasting accuracy and computational efficiency. This efficiency is enabled by the hard-routing strategy, which activates only a single regime-specific expert during inference. Consequently, the additional computational cost is incurred primarily during offline expert training and Bayesian hyperparameter optimization rather than during online deployment.

J. Threshold Sensitivity Analysis

To assess robustness to the regime-definition thresholds, the training-set thresholds from Section III-F are perturbed while keeping the trained experts and optimizer assignments fixed. Specifically, k_{low} and k_{high} are jointly varied by ± 0.10, σ_{low} by ± 0.10, and r_{th} from the 90th to the 99th percentile, one parameter at a time. Table XIV reports the resulting RMSE range for the CNN–LSTM MoE on Dataset 1.

TABLE XIV: Threshold sensitivity: RMSE range versus the baseline configuration (CNN–LSTM, Dataset 1, MoE).

Perturbation	RMSE range
Baseline (0.33/0.66/95th)	0.0539
k-thresholds (± 0.10)	0.0548–0.0566
σ_{low} (± 0.10)	0.0557–0.0574
r_{th} (90th–99th pct.)	0.0557–0.0592

Across all perturbations, RMSE increases by only 2–10%, indicating that the proposed routing strategy is not unduly sensitive to the exact threshold values. The largest degradation occurs when r_{th} is increased to the 99th percentile, reducing the Ramp class from approximately 5% to 2% of samples and worsening the same blindsided failure mode discussed in Section IV-D. Overall, the training-set quantiles provide a robust operating point without requiring precise threshold tuning.

K. Computational Cost and Deployment Feasibility

To assess deployment feasibility, the computational cost of the proposed framework was benchmarked on a CPU-only Google Colab instance (2× Intel® Xeon @ 2.20 GHz, 12.7 GB RAM, no GPU), which was used throughout this study. Table XV reports the parameter count, total training time (including Bayesian hyperparameter optimization), peak training memory, and per-sample inference latency for the global baseline and the proposed four-expert MoE.

The complete experimental protocol comprises up to 3,200 Bayesian-optimization trials for the regime-specific experts and 800 trials for the global baselines (4,000 trials across both datasets). Because the four experts are optimized independently using regime-specific training subsets, MoE training requires approximately 2.4–2.5× the training time of the corresponding global model, substantially less than the naive fourfold increase expected from training four full models. Peak memory usage increases by only 8–8.5% because the experts are trained sequentially rather than simultaneously. During deployment, inference latency is reduced by 17–21% relative to the global baseline because hard routing activates only a single expert for each input sample. Consequently, the additional computational cost is incurred primarily during offline training and Bayesian hyperparameter optimization, while online inference remains computationally lightweight.

L. Limitations and Trade-offs

The consistent results across two climatically distinct sites (Colorado and Nevada), including similar regime-difficulty ordering and optimizer preferences (Table V), suggest that the proposed framework captures general regime-dependent behavior rather than site-specific characteristics. Nevertheless, several limitations should be acknowledged.

Authorized licensed use limited to: Politecnico di Milano. Downloaded on November 10,2026 at 15:50:54 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Industry Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIA.2026.3746466

TABLE XV: Computational cost: global baseline vs. 4-expert MoE (CPU-only, Google Colab: 2× Intel Xeon @ 2.20 GHz, 12.7 GB RAM).

Arch.	Model	Params	Training (incl. HPO)	Peak mem.	Infer. latency
CNN–LSTM	Global	243K	4.5 h	180 MB	0.28 ms
CNN–LSTM	MoE*	303K	10.7 h	195 MB	0.22 ms
Transformer	Global	397K	6.0 h	240 MB	0.35 ms
Transformer	MoE*	581K	14.8 h	260 MB	0.29 ms

*Sum across the four regime experts.

The regime-identification thresholds (k_{low}, k_{high}, σ_{low}, and r_{th}) are estimated separately for each site and must be recalibrated for new installations or when sensor characteristics or the clear-sky model change. Although the required measurements (PV power and plane-of-array irradiance) are routinely available in utility-scale PV monitoring systems, the transferability of the proposed framework to distributed or rooftop PV installations remains to be investigated.

The framework is validated only for a 60-minute forecasting horizon and focuses on deterministic point forecasting rather than probabilistic prediction. Hard routing may introduce small forecast discontinuities at regime transitions, although Section IV-G shows that these discontinuities remain modest compared with the accuracy degradation observed for soft gating. Furthermore, because routing decisions are based exclusively on information available at forecast issuance, the framework cannot anticipate atmospheric regime transitions that occur within the prediction horizon. Finally, the proposed MoE requires greater offline computational effort for expert training and Bayesian hyperparameter optimization than a conventional global model, although its online inference cost remains low, as demonstrated in Section IV-K.

V. Conclusion

This paper presented a regime-routed mixture-of-experts framework for 60-minute-ahead PV power forecasting that combines regime-specific forecasting experts with optimizer specialization under a strictly causal and interpretable routing strategy. Using clear-sky variability and backward ramp-rate indicators, input windows are routed to CNN–LSTM or Transformer experts trained independently with regime-specific optimizer selection.

Experiments on two utility-scale PV datasets with one-minute temporal resolution demonstrated consistent improvements over conventional global training. Relative to the best global single-optimizer baselines, the proposed framework reduced test-set RMSE by approximately 14–16% for CNN–LSTM models and 15–18% for Transformer models. Ablation studies, capacity-matched comparisons, and robustness analyses further showed that these improvements arise primarily from regime-aware specialization rather than increased model capacity or optimizer selection alone. Future work will extend the proposed framework to additional climatic regions, multiple forecasting horizons, and probabilistic forecasting, while investigating adaptive regime identification and transition-aware routing strategies for improved performance during rapidly evolving atmospheric conditions.

References

[1] S. Dhingra, G. Gruosso, and G. S. Gajani, "Short-term solar pv power forecasting: A comparative analysis of neural network optimization techniques," in *2025 International Conference on Clean Electrical Power (ICCEP)*, 2025, pp. 181–186.

[2] R. H. Inman, H. T. C. Pedro, and C. F. M. Coimbra, "Solar forecasting methods for renewable energy integration," *Progress in Energy and Combustion Science*, vol. 39, pp. 535–576, 2013.

[3] C. Voyant, G. Notton, S. Kalogirou, M. Nivet, C. Paoli, F. Motte, and A. Fouilloy, "Machine learning methods for solar radiation forecasting: A review," *Renewable Energy*, vol. 105, pp. 569–582, 2017.

[4] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de Pison, and F. Antonanzas-Torres, "Review of photovoltaic power forecasting," *Solar Energy*, vol. 136, pp. 78–111, 2016.

[5] D. K. Dhaked, V. L. Narayanan, R. Gopal, O. Sharma, S. Bhattarai, and S. K. Dwivedy, "Exploring deep learning methods for solar photovoltaic power output forecasting: A review," *Renewable Energy Focus*, vol. 53, 2025.

[6] B. Lim, S. O. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, pp. 1748–1764, 2021.

[7] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting." in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 11 106–11 115.

[8] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," in *35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021, pp. 1–12.

[9] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting," *Proceedings of the 39 th International Conference on Machine Learning*, 2022.

[10] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," *arXiv preprint*, vol. arXiv:2211.14730, 2023.

[11] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint*, vol. arXiv:1711.05101, 2017.

[12] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," *arXiv preprint*, vol. arXiv:1908.03265, 2019.

[13] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint*, vol. arXiv:1609.04747, 2017.

[14] Z. Luo, S. Chen, Y. Qian, and Y. Hou, "Multi-stage stochastic gradient method with momentum acceleration," in *Signal Processing*, vol. 188, 2021.

[15] P. Lauret, R. Perez, L. Mazorra Aguiar, E. Tapachès, H. M. Diagne, and M. David, "Characterization of the intraday variability regime of solar irradiation of climatically distinct locations," *Solar Energy*, vol. 125, pp. 99–110, 2016.

[16] P. Lauret, R. Alonso-Suárez, J. Le Gal La Salle, and M. David, "Solar forecasts based on the clear sky index or the clearness index: Which is better?" *Solar*, vol. 2, pp. 432–444, 2022.

[17] C. Deline, K. Perry, M. Muller, W. Sekulic, and D. Jordan, "Photovoltaic data acquisition (pvdaq) public datasets, url = https://data.openei.org/submissions/4568," 2021.

[18] F. E. Atencio Espejo, S. Grillo, and L. Luini, "Photovoltaic power production estimation based on numerical weather predictions," in *2019 IEEE Milan PowerTech*, 2019, pp. 1–6.

[19] H. Wang, H. Shen, Y. Song, B. Feng, Y. Wu, Y. Zhang, P. Tao, J. Zhao, and R. Wan, "Enhanced short-term photovoltaic forecasting using patchtst and metaheuristic optimization," *IEEE Access*, vol. 13, pp. 147 467–147 481, 2025.

[20] V. Suresh, "Benchmarking transformer variants for hour-ahead pv forecasting: Patchtst with adaptive conformal inference," *Energies*, vol. 18, no. 18, p. 5000, 2025. [Online]. Available: https://doi.org/10.3390/en18185000

[21] Z. Song, F. Xiao, Z. Chen, and H. Madsen, "Probabilistic ultra-short-term solar photovoltaic power forecasting using natural gradient boosting with attention-enhanced neural networks," *Energy and AI*, vol. 20, p. 100496, 2025. [Online]. Available: https://doi.org/10.1016/j.egyai.2025.100496

Authorized licensed use limited to: Politecnico di Milano. Downloaded on November 10,2026 at 15:50:54 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.