PVFM: A SOLAR-AWARE, WEATHER-CONDITIONED FOUNDATION MODEL FOR ZERO-SHOT PHOTOVOLTAIC FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Rapid PV deployment across regions and installation types demands accurate and scalable power forecasting, particularly for newly commissioned plants. Existing forecasters are often site- and task-specific and transfer poorly to unseen sites, while general-purpose TSFMs show limited advantages in short-horizon and high-resolution PV forecasting, leaving unified forecasting across diverse PV scenarios an open challenge. We introduce PVFM, a 3.88M-parameter weather-conditioned foundation model for zero-shot PV power forecasting. The model separately models historical observations and future weather covariates through dedicated information streams, augmented with solar-aware embeddings that encode solar geometry and static identity information. It adopts a general-to-specific strategy, first pretraining on 59.2 million PV observations across global sites with weather covariates from multiple sources. The pretrained model is further evaluated under ramp events and extended through multimodal fine-tuning with satellite imagery for short-horizon forecasting. Across five zero-shot tasks spanning short- to long-term horizons at 1-hour and 15-minute resolutions, PVFM consistently outperforms both full-shot models and existing TSFMs, reducing MAE by 10.2–24.3% and CRPS by 10.0–29.5%. Satellite-image fine-tuning further reduces RMSE by 9.1% in ultra-short-term forecasting.

1 INTRODUCTION

Photovoltaic (PV) generation is expanding rapidly worldwide. Solar PV added a record 510 GW of capacity in 2025, accounting for nearly three quarters of global additions of renewable capacity International Renewable Energy Agency (2026). As PV supplies a growing share of electricity, its weather-driven variability creates increasing challenges for grid balancing and market operation. Accurate PV power forecasting is therefore essential for unit commitment, reserve allocation, market participation, and the reliable integration of solar energy.

PV forecasting spans diverse deployment settings and temporal scales (Antonanzas et al., 2016), from utility-scale plants to rooftop systems (Wang et al., 2023) and from short-term to multi-day prediction. Owing to the inherent characteristics of PV generation, its output is strongly affected by solar irradiance, cloud cover, and other atmospheric conditions. Early physics-based forecasting methods therefore relied on numerical weather prediction (NWP) and explicit physical models to translate future atmospheric conditions into expected PV output (Mayer & Grőf, 2021; Ahmed et al., 2020). Over time, PV forecasting has increasingly shifted toward data-driven models that learn these weather-to-power relationships from data while jointly exploiting historical power and meteorological covariates (Wang et al., 2025b; Kharlova et al., 2020). Combining improved weather forecasts with deep learning has further strengthened this weather-informed modeling paradigm (Zhang et al., 2026; Chen et al., 2025), while satellite and sky imagery provide complementary cloud information that is particularly valuable at shorter horizons (Zhang et al., 2018; Si et al., 2021; Ma et al., 2024; Boussif et al., 2023; Fan et al., 2026; Wan et al., 2025). However, most existing methods remain tailored to specific sites, forecasting tasks, or modality configurations, limiting the reuse of learned representations across PV datasets and operational settings.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

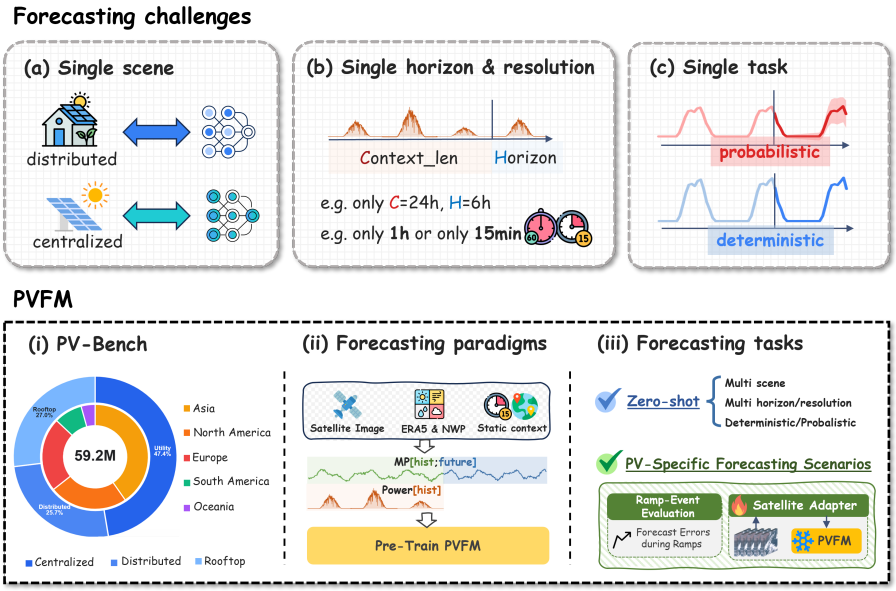

Figure 1: From fragmented PV forecasting to PVFM: unified pretraining across diverse PV scenes, temporal scales, and forecasting tasks.

Time-series foundation models (TSFMs) provide a promising paradigm for transferring knowledge from large, heterogeneous corpora to new forecasting tasks. Models such as Chronos-2 Ansari et al. (2025), TimesFM Das et al. (2024), and Moirai-2 Liu et al. (2025) have shown strong transfer performance, with recent TSFMs increasingly supporting covariates, including known future information. Their use in operational PV forecasting, however, remains limited, and the structured relationship between weather and PV generation may be underrepresented in general-purpose pretraining due to the limited presence of PV-specific data. This motivates a PV-specialized foundation model across diverse sites, horizons, and temporal resolutions.

In this paper, we propose PVFM, a foundation model for PV power forecasting. PVFM is designed to capture the interplay between historical PV generation and future weather conditions, while incorporating solar-aware temporal information for cross-site generalization. Beyond large-scale pretraining, we further study PVFM in PV-specific forecasting scenarios, including ramp-event evaluation and satellite-image fine-tuning. To the best of our knowledge, this is the first study to develop a foundation model specifically for photovoltaic power forecasting. Our contributions are fourfold:

- We introduce PVFM, a PV foundation model pretrained on 59.2 million valid PV time points across five continents, multiple temporal resolutions, and diverse installation types, enabling a unified model to transfer across heterogeneous PV forecasting settings.
- We propose a solar-aware architecture that separates historical representation learning from future weather conditioning and incorporates solar geometry to better characterize PV-relevant temporal patterns.
- PVFM demonstrates strong zero-shot generalization across five forecasting tasks at 1-hour and 15-minute resolutions, consistently improving over pretrained TSFMs and full-shot forecasting models trained on target-site data.
- We further demonstrate PVFM in PV-specific scenarios through ramp-event evaluation and satellite-image fine-tuning, showing its ability to handle challenging generation dynamics and incorporate additional weather observations for specialized forecasting needs.

2 Related Work

PV Power Forecasting. Data-driven PV forecasting models increasingly exploit weather information alongside historical power observations. Historical meteorological variables are commonly incorporated through recurrent or attention-based architectures (Kharlova et al., 2020; Hu et al., 2024), while recent methods further integrate future NWP to capture upcoming atmospheric conditions (Chen et al., 2025; Zhang et al., 2026). Beyond numerical weather variables, multimodal

approaches exploit satellite and cloud observations. CrossViViT combines satellite imagery with station time series (Boussif et al., 2023), FusionSF integrates heterogeneous inputs through vector-quantized representations (Ma et al., 2024), and recent methods further explore cross-modal, cloud-motion, and vision–language architectures (Wang et al., 2025a; Wan et al., 2025; Fan et al., 2026). However, PV forecasting is highly heterogeneous across installation types, resolutions, horizons, and regions. Existing methods are typically tailored to specific settings, with limited generalization to unseen sites and forecasting tasks.

Time-series forecasting and foundation models. Time-series forecasting has evolved from task-specific models toward foundation models pretrained on large and heterogeneous time-series corpora. Chronos, Moirai, TimesFM, and Lag-Llama demonstrated that cross-domain pretraining can enable competitive zero-shot forecasting on previously unseen datasets, while Timer and MOMENT further established the transferability of pretrained representations across forecasting and other time-series analysis tasks (Ansari et al., 2024; Woo et al., 2024; Das et al., 2024; Rasul et al., 2023; Liu et al., 2024b; Goswami et al., 2024). Subsequent models have extended this paradigm toward multivariate and covariate-informed forecasting, including conditioning on known future covariates (Woo et al., 2024; Ansari et al., 2025). However, existing covariate-aware TSFMs primarily employ general-purpose conditioning mechanisms that are not tailored to PV forecasting, where predictions depend strongly on future weather conditions and may further benefit from satellite observations of evolving cloud fields.

3 PRELIMINARIES

We formulate PV power forecasting as a covariate-conditioned, multi-resolution probabilistic time-series forecasting problem. For a PV site s, the forecasting task is defined over a discrete time axis with sampling interval $\Delta \in \{15\,\mathrm{min}, 60\,\mathrm{min}\}$. Let C and H denote the context length and forecast horizon in time steps, respectively. We denote the historical PV capacity-factor series by $\mathbf{P}^{\mathrm{hist}} \in \mathbb{R}^{C\times 1}$, the historical meteorological covariates by $\mathbf{M}^{\mathrm{hist}} \in \mathbb{R}^{C\times d_m}$, and the future meteorological covariates by $\mathbf{M}^{\mathrm{fut}} \in \mathbb{R}^{H\times d_m}$. Temporal information over the context and forecast intervals is denoted by $\boldsymbol{\tau}$, and static site metadata by $\mathbf{S}_s$. The objective is to estimate conditional quantiles of the future PV capacity-factor sequence:

$$\widehat{\mathbf{Y}} = f_\theta\left(\mathbf{P}^{\mathrm{hist}}, \mathbf{M}^{\mathrm{hist}}, \mathbf{M}^{\mathrm{fut}}, \boldsymbol{\tau}, \mathbf{S}_s\right) \in \mathbb{R}^{H\times|\mathcal{Q}|}. \tag{1}$$

Here, $\mathcal{Q}$ denotes the set of quantile levels used for probabilistic forecasting, and $\widehat{Y}_{h,j}$ represents the q_j-quantile prediction at forecast step h, where $q_j \in \mathcal{Q}$. The median is used as the point forecast, while the full set of quantiles represents predictive uncertainty.

4 METHODOLOGY

PVFM is a foundation model for transferable PV forecasting across diverse sites, horizons, and temporal resolutions. As illustrated in Figure 2, it adopts a dual-branch architecture consisting of a historical encoder and a future NWP-conditioned decoder, with solar-aware temporal and site conditioning. When satellite imagery is available, a residual adapter further incorporates historical cloud information into the pretrained representation.

4.1 SOLAR-AWARE EMBEDDING

Site Conditioning. For station s, let ϕ_s, λ_s, and z_s denote latitude, longitude, and the UTC offset in hours. We encode the site information as

$$\mathbf{g}_s = \left[\frac{\phi_s}{90}, \sin\left(\frac{\pi\lambda_s}{180}\right), \cos\left(\frac{\pi\lambda_s}{180}\right), \frac{z_s}{14}\right]. \tag{2}$$

The periodic longitude representation avoids the geographic wrap-around discontinuity. The site descriptor is projected to the hidden dimension as $\mathbf{e}_s^{\mathrm{site}} = \mathrm{MLP}_{\mathrm{site}}(\mathbf{g}_s) \in \mathbb{R}^d$ and used as an auxiliary station-level conditioning signal.

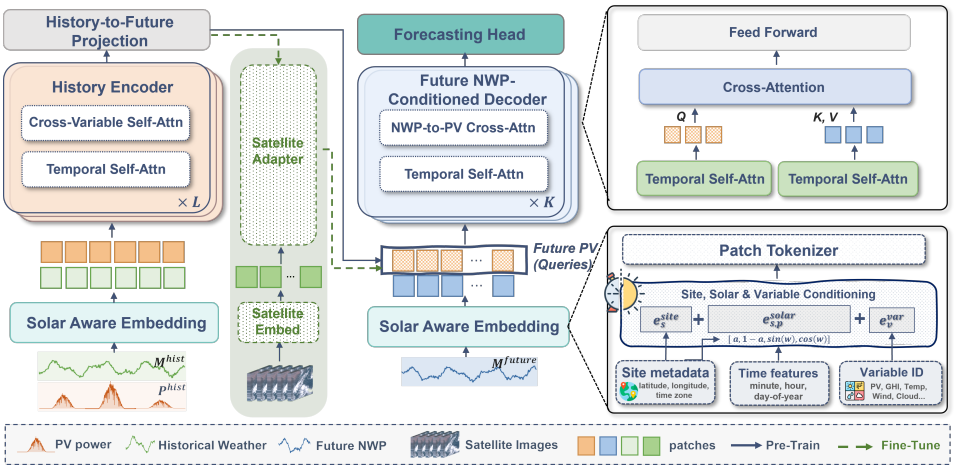

Figure 2: The overall framework of PVFM.

Solar-Aware Temporal Embedding. For each station s and time step k, PVFM computes the solar declination δ_k, hour angle $\omega_{s,k}$, and a nonnegative solar-elevation proxy $a_{s,k}$ from timestamps and geographic information using standard solar-position equations (Appendix B.2.1), and constructs

$$\mathbf{q}_{s,k} = [a_{s,k},\ 1 - a_{s,k},\ \sin\omega_{s,k},\ \cos\omega_{s,k}]\,. \tag{3}$$

Within each patch p, the solar features are concatenated and projected to the hidden dimension as $\mathbf{e}^{\text{solar}}_{s,p} = \text{MLP}_{\text{solar}}([\mathbf{q}_{s,k}]_{k\in p}) \in \mathbb{R}^d$, yielding a patch-level solar-aware temporal embedding.

Patch Tokenizer. Following patch-based time-series modeling (Nie et al., 2023; Zhang & Yan, 2023), each variable is divided into overlapping patches with patch size $P = 12$ and stride $S = 6$. Let $\mathbf{x}_{v,p} \in \mathbb{R}^P$ denote patch p of variable v. The initial patch token is constructed as

$$\mathbf{h}^{(0)}_{v,p} = \underbrace{\text{Linear}_{\text{patch}}(\mathbf{x}_{v,p})}_{\mathbf{z}_{v,p}} + \mathbf{e}^{\text{var}}_v + \mathbf{e}^{\text{site}}_s + \mathbf{e}^{\text{solar}}_{s,p}, \tag{4}$$

where $\mathbf{e}^{\text{var}}_v \in \mathbb{R}^d$ is a learnable variable identity embedding shared across all patches of variable v. Historical and future sequences use separate patch and solar projection modules.

4.2 Historical Encoder

Historical PV power and meteorological variables are jointly observed and are therefore modeled symmetrically in the historical encoder. Let $\mathbf{H}^{(0)} \in \mathbb{R}^{V\times N_{\text{h}}\times d}$ denote the historical patch representations, where $V = 7$ consists of the target PV series and six meteorological variables.

Temporal Self-Attention. Following the patch-wise modeling paradigm Zhang & Yan (2023), PVFM factorizes historical representation learning along the temporal and variable axes. Given historical patch representations $\mathbf{H}^{(\ell-1)} \in \mathbb{R}^{V\times N_h\times d}$, temporal self-attention is applied independently to the patch sequence of each variable:

$$\widetilde{\mathbf{H}}^{(\ell)}_{v,:,:} = \text{TSA}^{(\ell)}\left(\mathbf{H}^{(\ell-1)}_{v,:,:}\right), \qquad v = 1,\dots,V, \tag{5}$$

where $\text{TSA}^{(\ell)}$ denotes the temporal-attention sublayer, including its normalization and residual update. Rotary positional encoding is applied along the temporal axis.

Cross-Variable Self-Attention. At each aligned patch position, the temporally contextualized representations of PV and the six meteorological variables are jointly processed across the variable axis:

$$\overline{\mathbf{H}}^{(\ell)}_{:,p,:} = \text{VSA}^{(\ell)}\left(\widetilde{\mathbf{H}}^{(\ell)}_{:,p,:}\right), \qquad p = 1,\dots,N_{\text{h}}. \tag{6}$$

Cross-variable self-attention enables information exchange among PV and all historical meteorological variables at each aligned patch position. The resulting representations are subsequently processed by the feed-forward sublayer to obtain $\mathbf{H}^{(\ell)}$.

Historical PV Representation. After L encoder layers, we retain the PV-channel representation

$$\mathbf{Z}_{\mathrm{PV}} = \mathbf{H}^{(L)}_{\mathrm{PV},:,:} \in \mathbb{R}^{N_{\mathrm{h}} \times d}. \tag{7}$$

Although only the PV channel is passed to the subsequent history-to-future projection, it has interacted with all historical meteorological variables through cross-variable self-attention and therefore encodes both historical PV dynamics and meteorological context.

4.3 Future NWP-Conditioned Decoder

Unlike the historical encoder, the future decoder assigns distinct roles to the target and exogenous variables. Future NWP is available over the forecast horizon, whereas future PV is unobserved and must be predicted. We therefore construct latent future PV states and condition them asymmetrically on the future NWP representations.

Future PV Initialization. Let $\mathbf{Z}_{\mathrm{PV}} \in \mathbb{R}^{N_{\mathrm{h}} \times d}$ denote the historical PV representation produced by the encoder. The initial future PV states are obtained by projecting this representation onto the future patch axis,

$$\mathbf{D}^{(0)} = \mathcal{R}_{N_{\mathrm{f}}}\big(\mathcal{L}_{\mathrm{hist}\to\mathrm{future}}\big(\mathcal{R}_{\bar{N}_{\mathrm{h}}}(\mathbf{Z}_{\mathrm{PV}})\big)\big) \in \mathbb{R}^{N_{\mathrm{f}} \times d}, \tag{8}$$

where $\mathcal{L}_{\mathrm{hist}\to\mathrm{future}}$ is a learned linear projection along the patch axis and $\mathcal{R}$ denotes latent-space interpolation used to accommodate variable patch counts. No future PV observations are used in this initialization.

Temporal Self-Attention. At decoder layer k, temporal dependencies are first modeled separately within the PV and NWP branches. The latent PV sequence is updated as

$$\widetilde{\mathbf{D}}^{(k)} = \mathrm{TSA}^{(k)}_{\mathrm{PV}}\left(\mathbf{D}^{(k-1)}\right), \tag{9}$$

while temporal self-attention is independently applied to each NWP variable,

$$\widetilde{\mathbf{F}}^{(k)}_{v,:,:} = \mathrm{TSA}^{(k)}_{\mathrm{NWP}}\left(\mathbf{F}^{(k-1)}_{v,:,:}\right), \qquad v = 1, \ldots, V_m. \tag{10}$$

Here, TSA denotes the pre-normalized temporal-attention sublayer with residual connection. RoPE is applied along the future patch axis.

NWP-to-PV Cross-Attention. After temporal modeling, cross-variable interaction is performed asymmetrically at each future patch. Specifically, the latent PV state serves as the query, while the temporally contextualized NWP variables provide the keys and values. For patch p,

$$\widehat{\mathbf{d}}^{(k)}_{p} = \widetilde{\mathbf{d}}^{(k)}_{p} + \mathrm{MHA}\Big(\mathrm{LN}\Big(\widetilde{\mathbf{d}}^{(k)}_{p}\Big), \mathrm{LN}\Big(\widetilde{\mathbf{F}}^{(k)}_{:,p,:}\Big), \mathrm{LN}\Big(\widetilde{\mathbf{F}}^{(k)}_{:,p,:}\Big)\Big), \tag{11}$$

4.4 Probabilistic Forecasting Head

Future PV patch tokens are projected to step-level quantile outputs, with overlapping predictions averaged at each time step. The median is predicted directly, while the remaining quantiles are constructed through positive $\mathrm{softplus}$ increments, ensuring non-crossing forecasts by construction. Training minimizes the masked multi-quantile pinball loss,

$$\mathcal{L}_{\mathrm{q}} = \frac{\sum_{b,h} m_{b,h} \sum_{j=1}^{11} \rho_{q_j}\Big(y_{b,h} - \widehat{Y}_{b,h,j}\Big)}{11 \sum_{b,h} m_{b,h}}, \qquad \rho_q(e) = \max\left(qe, (q-1)e\right), \tag{12}$$

where $m_{b,h}$ identifies valid targets. Thus, all eleven quantiles are supervised jointly while invalid or padded positions are excluded.

4.5 Satellite-Image Adapter

Pretraining uses only PV and meteorological time series. For subsets with satellite coverage, we attach a trainable residual branch that injects historical cloud imagery into the frozen PVFM backbone. The backbone remains frozen. The adapter sits after the historical encoder and before the future decoder, and it corrects the historical PV representation rather than replacing it. Figure 3 summarizes the information flow.

Let $\mathbf{Z}_{\mathrm{PV}}$ denote the historical PV tokens and $\mathbf{S}$ the satellite representation aligned with the context window. Satellite information is injected through gated cross-attention:

$$\mathbf{Z}'_{\mathrm{PV}} = \mathbf{Z}_{\mathrm{PV}} + g \odot \mathrm{Proj}_0(\mathrm{CrossAttn}(\mathbf{Z}_{\mathrm{PV}}, \mathbf{S}, \mathbf{S})) \,. \quad (13)$$

The gate masks unavailable satellite inputs, while the zero-initialized projection preserves the pretrained representation at initialization. Only the adapter is optimized, with the decoder kept frozen. Satellite images are restricted to the context window to prevent future leakage (Appendix B.2.2).

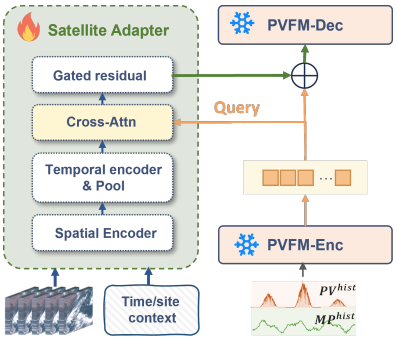

Figure 3: Satellite-image adapter.

5 Experimental Setup

5.1 Datasets

We collect PV observations from 845 real-world sites across 10 countries and regions on five continents, covering centralized, distributed, and rooftop installations. Hourly and 15-minute series are constructed from source data with sampling intervals ranging from 1 minute to 1 hour. Historical weather variables are obtained from ERA5 reanalysis Hersbach et al. (2020), while GFS forecasts National Centers for Environmental Prediction (2026) provide known future weather covariates. Satellite experiments additionally use Himawari-9 AHI imagery. Dataset sources, preprocessing, and geographic distributions are detailed in Appendix A.1.

5.2 Baselines

We compare PVFM against baselines from three categories, with detailed configurations and input interfaces provided in Appendix B.3.

PV-domain forecasting models. Two sub-categories are considered: **(i) Temporal-only:** Cross-Unet Zhang et al. (2026); and **(ii) Multimodal:** FusionSF Ma et al. (2024), Solar-VLM Fan et al. (2026), and MCloudNet Wan et al. (2025).

Time-series forecasting models. Three architectural categories are covered: **(i) Channel-independent:** PatchTST Nie et al. (2023); **(ii) Cross-variate:** Crossformer Zhang & Yan (2023), iTransformer Liu et al. (2024a), TimeMixer Wang et al. (2024a), and LightTS Zhang et al. (2022); and **(iii) Known-future-covariate-aware:** TiDE Das et al. (2023), TimeXer Wang et al. (2024b), DAG Qiu et al. (2026), and GCGNet Li et al. (2026).

Pretrained TSFMs. Six pretrained TSFMs with support for exogenous covariates are evaluated zero-shot using their officially released pretrained weights: Chronos-2 Ansari et al. (2025), Moirai-2 Liu et al. (2025), TiRex-2 Podest et al. (2026), TimesFM-3 Jain & Sen (2026), CITRAS-FM Yamaguchi et al. (2026), and TabPFN-TS3 Grinsztajn et al. (2026); Hoo et al. (2025).

5.3 Implementation Details

We evaluate PVFM and the baseline models across multiple forecasting horizons and temporal resolutions under zero-shot and full-shot settings, as detailed below.

Pretraining Setup. PVFM is pretrained on 746 sites, with 99 unseen sites for hourly zero-shot evaluation and 79 for 15-minute evaluation. Pretraining uses multi-resolution data with variable context and forecast lengths, capped at 672 and 168 steps, respectively. Detailed sampling and optimization settings are provided in Appendix B.2.3.

Evaluation Protocol. For each target station (99 for the hourly tasks, including 79 with 15-minute data), timestamps are partitioned chronologically into approximately 70%/10%/20% training, validation, and test segments. (i) **Zero-shot evaluation:** PVFM and publicly pretrained TSFMs are

Table 1: Zero-shot forecasting performance across five forecasting tasks. Shaded cells indicate the best result within each evaluation protocol (full-shot and zero-shot), while red bold and underlined values denote the best and second-best results overall, respectively. Imp. denotes the relative improvement over the best non-PVFM baseline. For FC/TC-augmented baselines, only the better-performing variant is shown.

Setting	Model	1 h									15 min					
		$C_{24} \to H_6$			$C_{72} \to H_{24}$			$C_{336} \to H_{168}$			$C_{64} \to H_{16}$			$C_{96} \to H_{24}$		
		$R^2 \uparrow$	MAE↓	CRPS↓	$R^2 \uparrow$	MAE↓	CRPS↓	$R^2 \uparrow$	MAE↓	CRPS↓	$R^2 \uparrow$	MAE↓	CRPS↓	$R^2 \uparrow$	MAE↓	CRPS↓
Full-shot	Cross-Unet	0.713	0.080	0.066	0.726	0.082	0.066	0.417	0.117	0.093	0.798	0.053	0.043	0.756	0.062	0.049
	FusionSF	0.713	0.065	0.048	0.728	0.065	0.049	0.589	0.089	0.065	0.802	0.047	0.034	0.736	0.057	0.042
	TiDE	0.557	0.113	0.092	0.643	0.098	0.084	0.546	0.113	0.092	0.535	0.092	0.076	0.734	0.066	0.057
	TimeXer	0.750	0.068	0.061	0.743	0.074	0.060	0.673	0.085	0.064	0.765	0.062	0.050	0.776	0.062	0.048
	DAG	0.727	0.082	0.074	0.675	0.096	0.078	0.517	0.116	0.088	0.768	0.065	0.051	0.780	0.062	0.049
	GCGNet	0.761	0.070	0.064	0.743	0.078	0.068	0.497	0.124	0.096	0.732	0.069	0.060	0.732	0.072	0.060
	PatchTST (FC)	0.755	0.065	0.055	0.724	0.074	0.055	0.665	0.086	0.061	0.739	0.066	0.049	0.764	0.063	0.046
	iTransformer (TC)	0.789	0.056	0.052	0.788	0.060	0.054	0.698	0.079	0.069	0.802	0.050	0.044	0.806	0.051	0.044
	Crossformer (FC)	0.783	0.067	0.056	0.650	0.091	0.073	0.311	0.135	0.103	0.821	0.048	0.039	0.807	0.052	0.043
	TimeMixer (TC)	0.753	0.073	0.065	0.723	0.082	0.068	0.641	0.092	0.072	0.777	0.055	0.048	0.787	0.056	0.047
	LightTS (FC)	0.637	0.099	0.077	0.529	0.120	0.090	0.352	0.136	0.102	0.776	0.063	0.048	0.776	0.064	0.049
Zero-shot	CITRAS-FM	0.709	0.061	0.046	0.667	0.073	0.054	0.754	0.061	0.046	0.482	0.085	0.066	0.549	0.082	0.063
	Chronos-2	0.618	0.081	0.061	0.758	0.060	0.044	0.774	0.058	0.041	0.410	0.083	0.065	0.593	0.074	0.054
	TiRex-2	0.511	0.098	0.078	0.694	0.075	0.060	0.747	0.065	0.050	0.506	0.081	0.063	0.595	0.076	0.058
	Moirai-2	0.016	0.157	0.122	0.494	0.115	0.087	0.585	0.105	0.078	0.435	0.096	0.078	0.252	0.122	0.093
	TimesFM-3	0.432	0.113	0.088	0.715	0.073	0.056	0.735	0.067	0.050	0.270	0.085	0.066	0.588	0.073	0.054
	TabPFN-TS3	0.366	0.114	0.092	0.647	0.079	0.061	0.674	0.074	0.055	0.383	0.098	0.079	0.280	0.111	0.089
	PVFM	**0.830**	**0.045**	**0.033**	**0.829**	**0.045**	**0.033**	**0.824**	**0.045**	**0.032**	**0.823**	**0.042**	**0.030**	**0.825**	**0.044**	**0.032**
Imp. PVFM (%)		**+5.2**	**+19.4**	**+29.5**	**+5.2**	**+24.3**	**+25.8**	**+6.5**	**+21.2**	**+20.5**	**+0.3**	**+10.2**	**+10.0**	**+2.3**	**+13.1**	**+24.8**

evaluated directly on the target stations without target-site parameter updates. (ii) **Full-shot training and evaluation:** Time-series forecasting models and PV-domain models are trained separately on each target station. For models without native support for known future covariates, we use *feature concatenation* (FC) along the variable dimension or *time concatenation* (TC) along the temporal dimension. See Appendix B.3.1 for details.

Forecasting Tasks and Evaluation. We evaluate five tasks at hourly and 15-minute resolutions. The hourly context/horizon settings are $C/H = 24/6$, $72/24$, and $336/168$ h, while the 15-minute settings are $16/4$ and $24/6$ h, corresponding to 64/16 and 96/24 time steps. Point forecasts are evaluated using MAE, RMSE, and R^2, and probabilistic forecasts using AQL and CRPS, with all metrics averaged equally across stations. Ramp events are additionally evaluated using POD, FAR, and CSI, with detailed definitions and criteria provided in Appendix B.1.

6 Experiments

We first evaluate PVFM on the main forecasting tasks, followed by cross-region zero-shot transfer and ablation studies. We further examine PV-specific forecasting scenarios through ramp-event evaluation and satellite-enhanced fine-tuning.

6.1 Main Forecasting Results

Protocol. We evaluate PVFM across five forecasting tasks on 99 unseen target stations for the hourly tasks and 79 of these stations for the 15-minute tasks. We compare against full-shot and zero-shot baselines using MAE, RMSE, and R^2. Additional details are provided in Appendix C.1.

Main results. Table 1 shows that PVFM consistently achieves the best overall performance across all five forecasting tasks. Compared with the best non-PVFM result for each task and metric, PVFM reduces MAE by 10.2–24.3% and CRPS by 10.0–29.5% across different horizons and temporal resolutions. Notably, these improvements are achieved under zero-shot evaluation, with PVFM outperforming target-site full-shot models across all five tasks. Among the full-shot baselines, iTransformer with time concatenation performs best most frequently, followed by FusionSF, although the strongest model varies across tasks and metrics. Pretrained TSFMs become more competitive in the long-context hourly setting, with Chronos-2 surpassing the best full-shot baseline across all three metrics at $C/H = 336/168$, whereas their advantage is less evident on the 15-minute tasks. One possible explanation is that general-purpose TSFMs are pretrained on heterogeneous

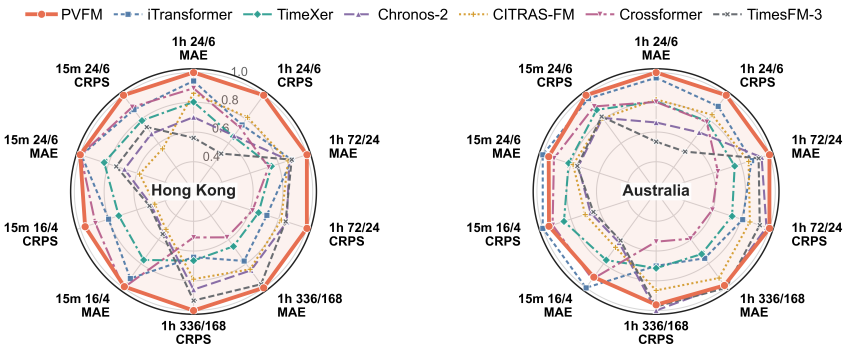

Figure 4: Cross-region performance in Hong Kong and Australia across three hourly and two 15-minute tasks. Radar axes show normalized station-equal MAE and CRPS, with larger values indicating better performance.

time series spanning different temporal resolutions, which may make it more difficult to capture the short-timescale variations prominent in 15-minute PV generation. PVFM addresses this challenge by learning transferable representations from a large-scale PV corpus, with its solar-aware patch embedding preserving shared solar-driven patterns across resolutions. Consistently, removing this embedding increases MAE across all five tasks (Figure 5; Section 6.3).

6.2 Cross-Region Zero-Shot Transfer

Protocol. We further evaluate PVFM under two region-holdout settings, where all stations from either Hong Kong or Australia are held out from pretraining and used exclusively for evaluation. We compare selected zero-shot and full-shot models on the held-out stations using MAE and CRPS. Complete results are provided in Tables 7 and 8.

Main reulsts. As shown in Figure 4, PVFM achieves the lowest MAE on four of five tasks in Hong Kong, with reductions up to 7.5% over the strongest baseline. Across both regions, PVFM consistently outperforms general-purpose TSFMs, except at $C/H = 336/168$ in Australia, where TimesFM-3 and Chronos-2 achieve slightly better MAE and CRPS, respectively. Notably, the advantage is particularly pronounced on the 15-minute tasks. While the best general-purpose TSFMs remain 32.3–77.7% behind the best full-shot models in MAE, PVFM reduces this gap to within 9.7% and even surpasses the best full-shot baseline on the 15-minute Hong Kong $C/H = 24/6$ task. Nevertheless, target-region full-shot models remain stronger on some 15-minute tasks in Australia, indicating that fine-resolution cross-region transfer remains challenging. The remaining gap in Australia may reflect a larger regional distribution shift, as geographically and climatically related PV data are less represented in the pretraining corpus, particularly at finer temporal resolutions.

6.3 Ablation Study

We ablate two key aspects of PVFM: (i) the information flow between historical observations and future weather covariates, and (ii) the embedding strategy for temporal and variable information.

Dual-branch architecture. To assess the separation of historical observations and known future weather covariates, we compare PVFM with encoder-only and encoder–decoder early-fusion variants under the same pretraining and evaluation protocol. The encoder-only variant increases MAE

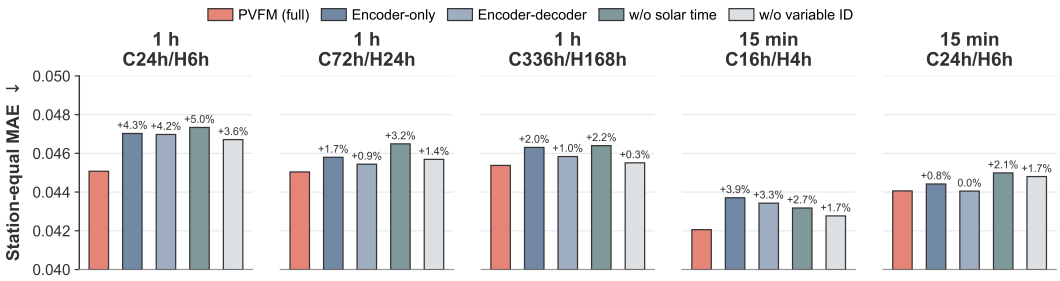

Figure 5: Ablation study of PVFM's information-flow architecture and PV-specific representation design. Lower MAE is better. Percentages indicate the relative MAE increase of each ablation over the full PVFM. All panels share the same truncated y-axis.

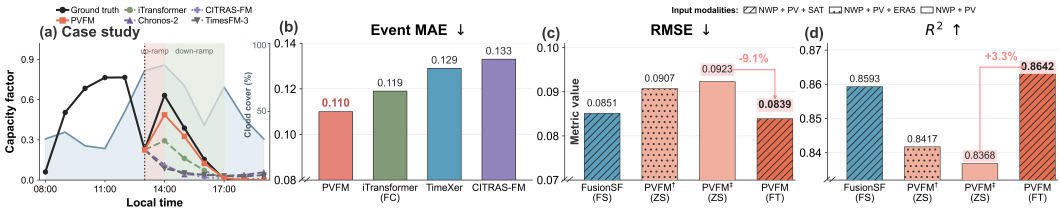

Figure 6: PV-specific forecasting evaluation. (a) Ramp-event forecasting case study. (b) Event-level MAE on ramp events. (c–d) RMSE and R^2 for satellite-enhanced 4-hour forecasting in Anhui under different training and input settings.

across all five tasks, with a maximum degradation of 4.3% relative to the full PVFM. Encoder–decoder early fusion likewise increases MAE on four of the five tasks by up to 4.2%, while remaining effectively tied with the full model on the remaining task. Overall, these results support PVFM's explicit separation of historical representations from known future covariates. We further compare PVFM with representative baseline architectures under the same pretraining protocol in Appendix C.3.2.

PV-specific representations. Replacing the solar-aware time encoding with calendar-based encoding increases MAE across all five tasks, with a maximum degradation of 5.0% relative to the full PVFM. This consistent effect across both hourly and 15-minute tasks supports its role in capturing shared solar-driven temporal patterns across resolutions. Removing the variable-identity embedding also increases MAE across all five tasks, by up to 3.6%, demonstrating the benefit of explicitly distinguishing the input variables.

6.4 PV-Specific Forecasting Scenarios

We further evaluate PVFM in two PV-specific settings beyond the standard forecasting tasks: ramp-event forecasting, which focuses on rapid changes in PV generation, and satellite-enhanced forecasting, which incorporates cloud observations from satellite imagery for short-horizon prediction.

Ramp-Event Evaluation. Ramp events are an important and challenging scenario in PV forecasting, as abrupt power changes are difficult to predict and can directly affect operational control Abuella & Chowdhury (2019); Wen et al. (2021). Figure 6(a) presents a representative case study containing a sharp up-ramp followed by a down-ramp. PVFM follows both transitions more closely than the compared models, which either underestimate the up-ramp magnitude or respond more slowly during the subsequent decline. Figure 6(b) summarizes the overall event-level errors, where PVFM achieves the lowest MAE of 0.110, compared with 0.119 for iTransformer (FC), 0.129 for TimeXer, and 0.133 for CITRAS-FM. Additional results are provided in Appendix C.4.1.

Multimodal Fine-tuning With Satellite Imagery. Considering the latency of ERA5 availability, we further explore satellite imagery as a timely source of recent atmospheric conditions for ultra-short-term forecasting in the Anhui region. Removing ERA5 from PVFM leads to only a slight performance degradation, while satellite-image fine-tuning under this setting improves RMSE and R^2 by 9.1% and 3.3%, respectively. With the proposed multimodal adaptation, PVFM not only recovers the performance of its ERA5-based counterpart but also surpasses the full-shot FusionSF baseline. Additional results are provided in Appendix C.4.2.

7 Conclusion

In this work, we address the challenge of transferable PV forecasting across heterogeneous sites, horizons, and temporal resolutions. To this end, we propose PVFM, a weather-conditioned foundation model pretrained on a large-scale PV corpus with 59.2 million time points. PVFM separately models historical observations and known future weather covariates through a dual-branch architecture, while solar-aware temporal representations encode the solar-driven temporal structure of PV generation. Extensive experiments demonstrate strong zero-shot generalization across unseen stations and regions, consistently outperforming general-purpose TSFMs and full-shot baselines on the main forecasting tasks. Future work will further strengthen the coupling between weather and renewable generation in PVFM, extending its capabilities to broader forecasting scenarios and weather information across spatial and temporal scales.

AI USE STATEMENT

Generative AI tools were used solely to assist with language editing and polishing. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve human subjects or personally identifiable information, and we are not aware of any significant ethical concerns associated with this study.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide an anonymous repository containing the implementation of PVFM, training and evaluation code, preprocessing scripts, and configuration files at `https://anonymous.4open.science/r/PVFM-173C/`. Detailed dataset descriptions, preprocessing procedures, training configurations, baseline settings, and evaluation protocols are provided in the appendix.

REFERENCES

Mohamed Abuella and Badrul Chowdhury. Forecasting of solar power ramp events: A post-processing approach. *Renewable Energy*, 133:1380–1392, 2019. doi: 10.1016/j.renene.2018.09.005.

Razin Ahmed, Victor Sreeram, Yateendra Mishra, and Muammer Din Arif. A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. *Renewable and Sustainable Energy Reviews*, 124:109792, 2020. doi: 10.1016/j.rser.2020.109792. URL `https://doi.org/10.1016/j.rser.2020.109792`.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=gerNCVqqtR`.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025. doi: 10.48550/arXiv.2510.15821. URL `https://arxiv.org/abs/2510.15821`.

J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de Pison, and F. Antonanzas-Torres. Review of photovoltaic power forecasting. *Solar Energy*, 136:78–111, 2016. doi: 10.1016/j.solener.2016.06.069. URL `https://doi.org/10.1016/j.solener.2016.06.069`.

Australian Energy Market Operator. National electricity market generation and load data, 2026. URL `https://www.aemo.com.au/energy-systems/electricity/national-electricity-market-nem/data-nem/market-management-system-mms-data/generation-and-load`. Accessed: 2026-09-15.

Oussama Boussif, Ghait Boukachab, Dan Assouline, Stefano Massaroli, Tianle Yuan, Loubna Benabbou, and Yoshua Bengio. Improving day-ahead solar irradiance time series forecasting by leveraging spatio-temporal context. *arXiv preprint arXiv:2306.01112*, 2023. URL `https://arxiv.org/abs/2306.01112`.

Dayin Chen, Xiaodan Shi, Mingkun Jiang, Shibo Zhu, Haoran Zhang, Dongxiao Zhang, Yuntian Chen, and Jinyue Yan. Selecting effective NWP integration approaches for PV power forecasting with deep learning. *Solar Energy*, 301:113939, 2025. doi: 10.1016/j.solener.2025.113939. URL https://doi.org/10.1016/j.solener.2025.113939.

Mingjian Cui, Jie Zhang, Anthony Florita, Bri-Mathias Hodge, Deping Ke, and Yuanzhang Sun. Solar power ramp events detection using an optimized swinging door algorithm. In *Proceedings of the ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Boston, Massachusetts, USA, 2015. doi: 10.1115/DETC2015-46849. URL https://www.nrel.gov/docs/fy15osti/64093.pdf. NREL/CP-5D00-64093.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with TiDE: Time-series dense encoder. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=pCbC3aQB5W.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167. PMLR, 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Desert Knowledge Australia Solar Centre. Alice springs solar centre data portal, 2026. URL https://www.dkasolarcentre.com.au/download?location=alice-springs. Accessed: 2026-09-15.

Caian Dutra de Jesus, Rhuan Nascimento Ferreira, and Eduardo Pestana de Aguiar. BR-PVGen: The brazilian photovoltaic generation dataset, 2026. URL https://doi.org/10.5281/zenodo.21511487.

Electric Power Research Institute. Solar forecast arbiter: Metrics. https://forecastarbiter.epri.com/metrics/, 2026. Accessed: 2026-09-13.

Rasool Fakoor, Taesup Kim, Jonas Mueller, Alexander J. Smola, and Ryan J. Tibshirani. Flexible model aggregation for quantile regression. *Journal of Machine Learning Research*, 24(162):1–45, 2023. URL https://jmlr.org/papers/v24/22-0799.html.

Hang Fan, Haoran Pei, Runze Liang, Weican Liu, Long Cheng, and Wei Wei. Solar-VLM: Multimodal vision-language models for augmented solar power forecasting. *arXiv preprint arXiv:2604.04145*, 2026. doi: 10.48550/arXiv.2604.04145. URL https://arxiv.org/abs/2604.04145.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152. PMLR, 2024. URL https://proceedings.mlr.press/v235/goswami24a.html.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger, Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Eliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kursat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. TabPFN-3: Technical report. *arXiv preprint arXiv:2605.13986*, 2026. doi: 10.48550/arXiv.2605.13986. URL https://arxiv.org/abs/2605.13986.

Hans Hersbach, Bill Bell, Paul Berrisford, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803. URL https://doi.org/10.1002/qj.3803.

Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. From tables to time: Extending TabPFN-v2 to time series forecasting. *arXiv preprint arXiv:2501.02945*, 2025. doi: 10.48550/arXiv.2501.02945. URL https://arxiv.org/abs/2501.02945. Title of version 4, revised January 2026; first submitted January 2025.

Zehuan Hu, Yuan Gao, Siyu Ji, Masayuki Mae, and Taiji Imaizumi. Improved multistep ahead photovoltaic power prediction model based on lstm and self-attention with weather forecast data. *Applied Energy*, 359:122709, 2024.

International Renewable Energy Agency. Renewable capacity statistics 2026, 2026. URL https://www.irena.org/Publications/2026/Mar/Renewable-capacity-statistics-2026.

Ayush Jain and Rajat Sen. TimesFM-3: A zero-shot foundation model for multivariate forecasting. Google Research Blog, August 2026. URL https://www.research.google/blog/timesfm-3-a-zero-shot-foundation-model-for-multivariate-forecasting/. Official release announcement, August 31, 2026; accessed September 13, 2026.

Elizaveta Kharlova, Daniel C. May, and Petr Musilek. Forecasting photovoltaic power production using a deep learning sequence-to-sequence model with attention. In *2020 International Joint Conference on Neural Networks*, pp. 1–7. IEEE, 2020. doi: 10.1109/IJCNN48605.2020.9207573. URL https://doi.org/10.1109/IJCNN48605.2020.9207573.

Zhengyu Li, Xiangfei Qiu, Yuhan Zhu, Xingjian Wu, Jilin Hu, Chenjuan Guo, and Bin Yang. GCGNet: Graph-consistent generative network for time series forecasting with exogenous variables. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EO5jwQ5NCw.

Zinan Lin, Qi Zhou, Zhe Wang, Ce Wang, Davis Boyd Bookhart, and Marcus Leung-Shea. A high-resolution three-year dataset supporting rooftop photovoltaics (PV) generation analytics, 2024. URL https://doi.org/10.5061/dryad.m37pvmd99.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025. doi: 10.48550/arXiv.2511.11698. URL https://arxiv.org/abs/2511.11698.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=JePfAI8fah.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32369–32399. PMLR, 2024b. URL https://proceedings.mlr.press/v235/liu24cb.html.

Atse Louwen, Alexander Astigarraga, Enrico Dalla Maria, Lukas Koester, and David Moser. Monitoring data TRUST-PV bifacial HJT SAT plant, bolzano, italy, 2025. URL https://doi.org/10.5281/zenodo.14766289.

Ziqing Ma, Wenwei Wang, Tian Zhou, Chao Chen, Bingqing Peng, Liang Sun, and Rong Jin. FusionSF: Fuse heterogeneous modalities in a vector quantized framework for robust solar power forecasting. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 2024. doi: 10.1145/3637528.3671509. URL https://arxiv.org/abs/2402.05823.

Martin János Mayer and Gyula Grőf. Extensive comparison of physical models for photovoltaic power forecasting. *Applied Energy*, 283:116239, 2021. doi: 10.1016/j.apenergy.2020.116239. URL https://doi.org/10.1016/j.apenergy.2020.116239.

National Centers for Environmental Prediction. Global forecast system (GFS). National Oceanic and Atmospheric Administration, 2026. URL https://www.ncei.noaa.gov/products/weather-climate-models/global-forecast. Accessed: 2026-09-14.

National Renewable Energy Laboratory. PVDAQ: Photovoltaic data acquisition, 2026. URL https://github.com/openEDI/documentation/blob/main/pvdaq.md. Accessed: 2026-09-15.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Open Climate Fix. UK PV, 2025. URL https://huggingface.co/datasets/openclimatefix/uk_pv.

Patrick Podest, Marco Pichler, Elias Bürger, Levente Zólyomi, Bernhard Voggenberger, Wilhelm Berghammer, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. TiRex-2: Generalizing TiRex to multivariate data and streaming. *arXiv preprint arXiv:2607.01204*, 2026. doi: 10.48550/arXiv.2607.01204. URL https://arxiv.org/abs/2607.01204.

Daniel Vázquez Pombo. The SOLETE dataset, 2022. URL https://doi.org/10.11583/DTU.17040767.

Xiangfei Qiu, Yuhan Zhu, Zhengyu Li, Xingjian Wu, Bin Yang, and Jilin Hu. DAG: A dual correlation network for time series forecasting with exogenous variables. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2509.14933.

Luis Ramírez Camargo and Johannes Schmidt. Time series of electricity output for large grid connected photovoltaic installations in chile, 2020. URL https://doi.org/10.5281/zenodo.3939047.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, Marin Biloš, Sahil Garg, Anderson Schneider, Nicolas Chapados, Alexandre Drouin, Valentina Zantedeschi, Yuriy Nevmyvaka, and Irina Rish. Lag-llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*, 2023. doi: 10.48550/arXiv.2310.08278. URL https://arxiv.org/abs/2310.08278.

Zhiyuan Si, Ming Yang, Yixiao Yu, and Tingting Ding. Photovoltaic power forecast based on satellite images considering effects of solar position. *Applied Energy*, 302:117514, 2021. doi: 10.1016/j.apenergy.2021.117514. URL https://doi.org/10.1016/j.apenergy.2021.117514.

Meng Wan, Tiantian Liu, Yuxuan Bi, Jue Wang, Hui Cui, Rongqiang Cao, Jiaxiang Wang, Peng Shi, Ningming Nie, and Yangang Wang. MCloudNet: An ultra-short-term photovoltaic power forecasting framework with multi-layer cloud coverage. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 9908–9917. International Joint Conferences on Artificial Intelligence Organization, August 2025. doi: 10.24963/ijcai.2025/1101. URL https://doi.org/10.24963/ijcai.2025/1101. AI and Social Good.

Jianxun Wang, Xin Chen, Weicheng Jiang, Li Hua, Junyi Liu, and Haigang Sui. PVNet: A novel semantic segmentation model for extracting high-quality photovoltaic panels in large-scale systems from high-resolution remote sensing imagery. *International Journal of Applied Earth Observation and Geoinformation*, 119:103309, 2023. doi: 10.1016/j.jag.2023.103309. URL https://doi.org/10.1016/j.jag.2023.103309.

Kai Wang, Shuo Shan, Weijing Dou, Haikun Wei, and Kanjian Zhang. A robust photovoltaic power forecasting method based on multimodal learning using satellite images and time series. *IEEE Transactions on Sustainable Energy*, 16(2):970–980, 2025a.

Keqi Wang, Lijie Wang, Qiang Meng, Chao Yang, Yangshu Lin, Junye Zhu, Zhongyang Zhao, Can Zhou, Chenghang Zheng, and Xiang Gao. Accurate photovoltaic power prediction via temperature correction with physics-informed neural networks. *Energy*, 328:136546, 2025b. doi: 10.1016/j.energy.2025.136546. URL https://doi.org/10.1016/j.energy.2025.136546.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=7oLshfEIC2.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. TimeXer: Empowering transformers for time series forecasting with exogenous variables. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0113ef4642264adc2e6924a3cbbdf532-Abstract-Conference.html.

Haoran Wen, Yang Du, Xiaoyang Chen, Enggee Lim, Huiqing Wen, Lin Jiang, and Wei Xiang. Deep learning based multistep solar forecasting for pv ramp-rate control using sky images. *IEEE Transactions on Industrial Informatics*, 17(2):1397–1406, 2021. doi: 10.1109/TII.2020.2987916.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164. PMLR, 2024.

Yosuke Yamaguchi, Issei Suemitsu, Yuki Kajihara, and Wenpeng Wei. CITRAS-FM: Tiny time series foundation model for covariate-informed zero-shot forecasting. *arXiv preprint arXiv:2606.10798*, 2026. doi: 10.48550/arXiv.2606.10798. URL https://arxiv.org/abs/2606.10798. Accepted to EUSIPCO 2026.

T. Yao et al. PVOD v1.0: A photovoltaic power output dataset, 2022. URL https://doi.org/10.11922/sciencedb.01094.

Jinsong Zhang, Rodrigo Verschae, Shohei Nobuhara, and Jean-François Lalonde. Deep photovoltaic nowcasting. *Solar Energy*, 176:267–276, 2018. doi: 10.1016/j.solener.2018.10.024. URL https://doi.org/10.1016/j.solener.2018.10.024.

Tianping Zhang, Yizhuo Zhang, Wei Cao, Jiang Bian, Xiaohan Yi, Shun Zheng, and Jian Li. Less is more: Fast multivariate time series forecasting with light sampling-oriented MLP structures. *arXiv preprint arXiv:2207.01186*, 2022. doi: 10.48550/arXiv.2207.01186. URL https://arxiv.org/abs/2207.01186.

Yujia Zhang, Yuzhou Zhang, Zhixiang Dai, and Rita Zhang. Rethinking the use of deep learning methods for photovoltaic power forecasting. *Nature Communications*, 17:7613, 2026. doi: 10.1038/s41467-026-73817-3. URL https://www.nature.com/articles/s41467-026-73817-3.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vSVLM2j9eie.

APPENDIX

TABLE OF CONTENTS

A Dataset Details **16**
A.1 PV Bench 16
A.2 Data Preprocessing and Alignment 16
B Experimental Details **16**
B.1 Metrics 17
B.1.1 Probabilistic Metrics 17
B.1.2 Deterministic Metrics 17
B.1.3 Ramp-Event Evaluation 18
B.2 PVFM Implementation Details 20
B.2.1 Solar Geometry Computation 20
B.2.2 Satellite Adapter Details 20
B.2.3 Training and Optimization Details 21
B.3 Baseline Implementation Details 21
B.3.1 Full-Shot Forecasting Models 21
B.3.2 Multi-Modal Baselines 23
B.3.3 Pretrained Time-Series Foundation Models 24
C Additional Results **24**
C.1 Full Forecasting Results 24
C.2 Cross-Region Generalization 25
C.3 Architecture Comparisons 25
C.3.1 Model Capacity Scaling 25
C.3.2 Architecture Comparison under Shared Pretraining 25
C.3.3 Pretraining Dataset Scaling 26
C.4 PV-Specific Forecasting Scenarios 27
C.4.1 Ramp-Event Evaluation 27
C.4.2 Satellite-Image Fine-Tuning 29
C.5 Case Study 29
C.6 Seed Variability 31

Table 2: Geographic distribution and original data sources for stations in the pretraining and zero-shot splits.

Continent	Dataset / region	PV deployment type	Pretraining	Zero-shot	Native cadence	Temporal coverage	Original data source
Asia	Anhui, China	Centralized	218	17	5 min	2023–2024	Project industrial PV records
	Guangxi, China	Centralized	40	13	15 min	2023–2025	
	Yunnan, China	Centralized	1	0	1 min	2025	
	Guangdong, China	Centralized	1	0	15 min	2024–2025	
	Hubei, China	Centralized	1	0	15 min	2024–2025	
	Xinjiang, China	Centralized	4	0	15 min	2016–2018	
	Hebei, China	Centralized	10	0	15 min	2018–2019	PVOD (2022)
	Hong Kong, China	Rooftop	25	12	15 min	2021–2023	HKUST (2024)
Europe	United Kingdom	Rooftop	174	17	5/30 min	2010–2025	UK PV (2025)
	Italy	Distributed	1	0	5 min	2024	TRUST-PV (2025)
	Denmark	Distributed	1	0	5 min	2018–2019	SOLETE (2022)
North America	United States	Centralized	4	0	5–30 min	2008–2025	NREL PVDAQ (2026)
	United States	Distributed	181	17	1–15 min	2013–2019	
South America	Chile	Centralized	44	12	1 h	2014–2019	Chile PV (2020)
	Brazil	Distributed	15	0	15 min	2024–2025	BR-PVGen (2026)
	Colombia	Distributed	1	0	5 min	2024	UNC local archive
Oceania	Australia	Centralized	24	11	5 min	2024–2026	AEMO SCADA (2026)
	Australia	Distributed	1	0	5 min	2008–2026	DKASC (2026)
Total	**10 countries/regions**	**400 centralized / 217 distributed / 228 rooftop**	**746**	**99**	**1 min–1 h**	**2008–2026**	–

A Dataset Details

A.1 PV Bench

Table 2 provides detailed information on the geographic coverage, PV deployment type, temporal coverage, native sampling cadence, and original source of each dataset, as well as the station counts used for pretraining and zero-shot evaluation.

A.2 Data Preprocessing and Alignment

To harmonize PV measurements collected at different sampling frequencies, we first convert all timestamps to site-local time and organize each station on a regular temporal grid. Invalid or missing PV observations are retained through explicit masks rather than imputed. We construct an hourly representation for all stations, while preserving native 15-minute measurements when available. PV power is normalized by the installed capacity of each site, yielding capacity-factor targets that are comparable across installations of different scales.

Historical meteorological conditions are obtained from ERA5, while future meteorological covariates are derived from archived GFS forecasts. Both are matched to each PV site using its geographic coordinates and represented by six common variables: temperature, cloud cover, relative humidity, surface pressure, shortwave radiation, and wind speed. All meteorological variables are aligned with the PV sequence according to their valid timestamps. For hourly forecasting, the weather variables are matched directly to the hourly PV grid. For native 15-minute forecasting, each hourly weather value is shared by the four quarter-hour observations within that hour, avoiding the introduction of artificial sub-hourly weather variations. Shortwave radiation is treated according to its original interval-averaged semantics.

B Experimental Details

This section provides additional details of the experimental setup, including the evaluation metrics, PVFM implementation and training settings, and the implementation details of the baselines.

B.1 METRICS

We evaluate forecasting performance from deterministic, probabilistic, and ramp-event perspectives. Deterministic performance is assessed using MAE, RMSE, and R^2, while probabilistic forecasts are evaluated using average quantile loss (AQL) and Continuous Ranked Probability Score (CRPS). We additionally evaluate ramp-event forecasting using event-based detection metrics.

B.1.1 PROBABILISTIC METRICS

Probabilistic Evaluation Protocol. All methods are evaluated on the common quantile grid $\mathcal{Q}_{\text{eval}} = \{0.1, 0.2, \ldots, 0.9\}$, with $K = 9$. For PVFM, these nine quantiles are selected from its native Q11 outputs without modifying the trained model. Let $\mathcal{I}_s$ denote the valid forecast-window–horizon pairs for station s, and let $N_s = |\mathcal{I}_s|$. For each pair i, $y_{s,i}$ and $\hat{q}_{s,i,k}$ denote the observed capacity factor and the predicted τ_k-quantile, respectively.

Average Quantile Loss (AQL). AQL measures predictive accuracy across the evaluated quantile levels. The pinball loss is defined as

$$\Lambda_\tau(q, y) = \big(\tau - \mathbf{1}\{y < q\}\big)(y - q). \tag{14}$$

The station-level AQL is

$$\text{AQL}_s = \frac{1}{N_s K} \sum_{i \in \mathcal{I}_s} \sum_{k=1}^{K} \Lambda_{\tau_k}(\hat{q}_{s,i,k}, y_{s,i}). \tag{15}$$

Continuous Ranked Probability Score (CRPS). CRPS measures the discrepancy between a predictive distribution and the observed outcome. For a predictive CDF F and observation y, it admits the quantile representation Fakoor et al. (2023):

$$\text{CRPS}(F, y) = 2 \int_0^1 \Lambda_\tau\big(F^{-1}(\tau), y\big)\, d\tau. \tag{16}$$

Because forecasts are available on a finite quantile grid, we approximate this integral using the trapezoidal rule. Before CRPS integration, quantile crossings, if present, are corrected using a cumulative maximum: $\tilde{q}_{s,i,k} = \max_{1 \le j \le k} \hat{q}_{s,i,j}$. This operation leaves noncrossing predictions unchanged. We extend the grid with $\tau_0 = 0$ and $\tau_{K+1} = 1$, holding the endpoint predictions constant: $\tilde{q}_{s,i,0} = \tilde{q}_{s,i,1}$ and $\tilde{q}_{s,i,K+1} = \tilde{q}_{s,i,K}$. Writing $\ell_{s,i,k} = \Lambda_{\tau_k}(\tilde{q}_{s,i,k}, y_{s,i})$, the reported station-level approximation is

$$\widehat{\text{CRPS}}_s = \frac{2}{N_s} \sum_{i \in \mathcal{I}_s} \sum_{k=0}^{K} \frac{\tau_{k+1} - \tau_k}{2} \left(\ell_{s,i,k} + \ell_{s,i,k+1}\right). \tag{17}$$

Both metrics are computed in capacity-factor units without additional normalization by the sum of observed power. For each forecasting task, we first calculate each station's score and then report the arithmetic mean across stations, giving every station equal weight. Lower values indicate better probabilistic forecasting performance.

B.1.2 DETERMINISTIC METRICS

Deterministic Evaluation Protocol. We evaluate deterministic forecasting performance using mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2). Let $\hat{y}_{s,i}$ denote the point forecast corresponding to observation $y_{s,i}$. For models with probabilistic outputs, we consistently use the predictive median as the point forecast. This includes the deterministic baselines augmented with quantile prediction, for which $\hat{y}_{s,i} = \hat{q}_{s,i}(0.5)$, as well as pretrained time-series foundation models, for which the corresponding median forecast is used. All three metrics are computed over the valid forecast-window–horizon pairs $\mathcal{I}_s$ defined above.

Mean Absolute Error (MAE). MAE measures the average absolute deviation between predictions and observations. The station-level MAE is

$$\mathrm{MAE}_s = \frac{1}{N_s} \sum_{i \in \mathcal{I}_s} |\hat{y}_{s,i} - y_{s,i}| \,. \tag{18}$$

MAE is expressed in capacity-factor units, with lower values indicating more accurate point forecasts.

Root Mean Squared Error (RMSE). RMSE assigns greater weight to larger prediction errors through the squared-error term:

$$\mathrm{RMSE}_s = \sqrt{\frac{1}{N_s} \sum_{i \in \mathcal{I}_s} (\hat{y}_{s,i} - y_{s,i})^2}. \tag{19}$$

Like MAE, RMSE is expressed in capacity-factor units, and lower values indicate better performance.

Coefficient of Determination (R^2). R^2 measures the reduction in squared error relative to predicting the mean of the evaluated observations. Defining

$$\bar{y}_s = \frac{1}{N_s} \sum_{i \in \mathcal{I}_s} y_{s,i}, \tag{20}$$

the station-level score is

$$R^2_s = 1 - \frac{\sum_{i \in \mathcal{I}_s} (y_{s,i} - \hat{y}_{s,i})^2}{\sum_{i \in \mathcal{I}_s} (y_{s,i} - \bar{y}_s)^2}. \tag{21}$$

Higher values indicate better performance, with $R^2 = 1$ corresponding to perfect predictions. Negative values indicate larger squared error than the constant-mean reference.

Station-Equal Aggregation. For each forecasting task, we compute each metric separately for every station and then take the arithmetic mean:

$$\overline{m} = \frac{1}{|\mathcal{S}_m|} \sum_{s \in \mathcal{S}_m} m_s, \qquad m \in \{\mathrm{MAE}, \mathrm{RMSE}, R^2\}, \tag{22}$$

where $\mathcal{S}_m$ contains stations with a defined, finite value of metric m. This gives each station equal weight regardless of its number of valid forecast pairs.

B.1.3 Ramp-Event Evaluation

Ramp-Event Definition. We evaluate ramp-event detection using the directional thresholds of Cui et al. Cui et al. (2015). Let $\mathcal{J}_s$ denote the set of valid adjacent one-hour intervals for station s. For interval $j \in \mathcal{J}_s$, the observed and predicted capacity-factor changes are defined as

$$\Delta y_{s,j} = y_{s,j} - y_{s,j^-}, \qquad \Delta \hat{y}_{s,j} = \hat{y}_{s,j} - \hat{y}_{s,j^-}, \tag{23}$$

where j^- denotes the preceding hourly point. For the first forecast lead, the final observed context value is used as the preceding value for both the observed and predicted changes. For subsequent leads, the preceding observation and preceding point forecast are used, respectively. The point forecast $\hat{y}_{s,j}$ is the predictive median for probabilistic models.

The observed directional event indicators are

$$z^{\uparrow}_{s,j} = \mathbb{I}(\Delta y_{s,j} > 0.10) \,, \qquad z^{\downarrow}_{s,j} = \mathbb{I}(\Delta y_{s,j} < -0.08) \,, \tag{24}$$

and the corresponding predicted indicators are

$$\hat{z}^{\uparrow}_{s,j} = \mathbb{I}(\Delta \hat{y}_{s,j} > 0.10) \,, \qquad \hat{z}^{\downarrow}_{s,j} = \mathbb{I}(\Delta \hat{y}_{s,j} < -0.08) \,. \tag{25}$$

Changes equal to either threshold are treated as non-ramp intervals. Overall ramp occurrence is defined by

$$z^{\mathrm{all}}_{s,j} = z^{\uparrow}_{s,j} \vee z^{\downarrow}_{s,j}, \qquad \hat{z}^{\mathrm{all}}_{s,j} = \hat{z}^{\uparrow}_{s,j} \vee \hat{z}^{\downarrow}_{s,j}. \tag{26}$$

Thus, the overall category evaluates whether a ramp occurs regardless of direction, whereas the upward and downward categories evaluate the corresponding direction separately.

Event Contingency Counts. For each event category $c \in \{\text{all}, \uparrow, \downarrow\}$, we compare the observed indicator $z^c_{s,j}$ with the predicted indicator $\hat{z}^c_{s,j}$. The station-level numbers of hits, misses, and false alarms are

$$\begin{aligned} N^c_{\mathrm{H},s} &= \sum_{j \in \mathcal{J}_s} z^c_{s,j} \hat{z}^c_{s,j}, \\ N^c_{\mathrm{M},s} &= \sum_{j \in \mathcal{J}_s} z^c_{s,j} \left(1 - \hat{z}^c_{s,j}\right), \\ N^c_{\mathrm{F},s} &= \sum_{j \in \mathcal{J}_s} \left(1 - z^c_{s,j}\right) \hat{z}^c_{s,j}. \end{aligned} \tag{27}$$

A hit denotes a ramp observed and predicted over the same one-hour interval; a miss denotes an observed ramp that is not predicted; and a false alarm denotes a predicted ramp without a corresponding observed ramp. Matching is performed strictly at the same station, forecast origin, and lead interval, without temporal tolerance or episode merging.

Ramp-Detection Metrics. Ramp-event detection is quantified using the probability of detection (POD), false alarm ratio (FAR), and critical success index (CSI):

$$\begin{aligned} \mathrm{POD}^c_s &= \frac{N^c_{\mathrm{H},s}}{N^c_{\mathrm{H},s} + N^c_{\mathrm{M},s}}, \\ \mathrm{FAR}^c_s &= \frac{N^c_{\mathrm{F},s}}{N^c_{\mathrm{H},s} + N^c_{\mathrm{F},s}}, \\ \mathrm{CSI}^c_s &= \frac{N^c_{\mathrm{H},s}}{N^c_{\mathrm{H},s} + N^c_{\mathrm{M},s} + N^c_{\mathrm{F},s}}. \end{aligned} \tag{28}$$

Higher POD and CSI indicate better event detection, whereas lower FAR indicates fewer false alarms among the predicted ramp events. A metric with a zero denominator is treated as undefined rather than assigned a value of zero.

Forecast Errors During Ramp Events. In addition to event detection, we measure point-forecast accuracy during observed ramp intervals. For event category c, let

$$\mathcal{E}^c_s = \left\{ j \in \mathcal{J}_s : z^c_{s,j} = 1 \right\} \tag{29}$$

denote the observed ramp intervals at station s, and let $N^c_{\mathrm{E},s} = |\mathcal{E}^c_s|$. The ramp-event MAE and RMSE are

$$\mathrm{EventMAE}^c_s = \frac{1}{N^c_{\mathrm{E},s}} \sum_{j \in \mathcal{E}^c_s} \left| \hat{y}_{s,j} - y_{s,j} \right|, \tag{30}$$

and

$$\mathrm{EventRMSE}^c_s = \sqrt{\frac{1}{N^c_{\mathrm{E},s}} \sum_{j \in \mathcal{E}^c_s} \left(\hat{y}_{s,j} - y_{s,j} \right)^2}. \tag{31}$$

These metrics evaluate capacity-factor prediction errors at the end points of observed ramp intervals, irrespective of whether the event was successfully detected. They therefore measure forecast accuracy under ramp conditions and should not be interpreted as errors in the predicted ramp magnitude.

Station-Equal Aggregation. All ramp metrics are first computed independently for each station and then averaged equally across eligible stations:

$$\overline{m}^c = \frac{1}{|\mathcal{S}^c_m|} \sum_{s \in \mathcal{S}^c_m} m^c_s, \qquad m \in \{\mathrm{POD}, \mathrm{FAR}, \mathrm{CSI}, \mathrm{EventMAE}, \mathrm{EventRMSE}\}. \tag{32}$$

Here, $\mathcal{S}^c_m$ contains stations with a defined, finite metric and at least five unique observed ramp intervals for category c. Consequently, each eligible station contributes equally, regardless of its number of forecast windows or ramp events. Event counts may be reported for auditing, but the headline metrics are not computed by pooling all intervals across stations.

The headline evaluation applies the thresholds directly to the observed and forecast capacity-factor sequences. No clear-sky or weather-attribution filtering is applied to these primary ramp-event metrics. Following the EPRI Solar Forecast Arbiter Electric Power Research Institute (2026), detection performance is evaluated using probability of detection (POD), false alarm ratio (FAR), and critical success index (CSI). These metrics are reported for all ramp events jointly and separately for up-ramp and down-ramp events. Metrics are first computed for each eligible station and then averaged across stations with equal weight. A station is considered eligible if its test period contains at least five observed ramp intervals.

B.2 PVFM Implementation Details

All timestamps are represented in site-local time. Missing or invalid observations are masked rather than interpolated. PV power is converted to capacity factor using the known installed capacity of each station. For PVFM, meteorological normalization statistics are estimated exclusively from the training partition of the pretraining corpus and kept fixed during zero-shot evaluation on unseen stations. For full-shot baselines, these statistics are estimated separately from each target station's training split and applied unchanged to its validation and test splits.

B.2.1 Solar Geometry Computation

For completeness, we provide the solar-position equations used to construct the solar-aware temporal features. Let h_k and n_k denote the hour and day of year at time step k, respectively. For station s with longitude λ_s and UTC offset z_s, the solar hour is approximated as

$$h_{s,k}^{\mathrm{solar}} = h_k - z_s + \frac{\lambda_s}{15}. \tag{33}$$

The solar declination is computed as

$$\delta_k = 23.45^\circ \sin\left(\frac{2\pi(284 + n_k)}{365}\right), \tag{34}$$

and the corresponding hour angle is

$$\omega_{s,k} = 15^\circ \left(h_{s,k}^{\mathrm{solar}} - 12\right). \tag{35}$$

Given the station latitude ϕ_s, the solar-elevation proxy is

$$a_{s,k} = \max\left(0,\, \sin\phi_s^{\mathrm{rad}} \sin\delta_k^{\mathrm{rad}} + \cos\phi_s^{\mathrm{rad}} \cos\delta_k^{\mathrm{rad}} \cos\omega_{s,k}^{\mathrm{rad}}\right), \tag{36}$$

where the superscript rad denotes conversion from degrees to radians. The clipping at zero suppresses negative values when the Sun is below the horizon.

B.2.2 Satellite Adapter Details

For a context window, let $\mathbf{I}$ denote the historical satellite image sequence, c the reused time and site context, and m the per-frame image-validity mask. A spatial encoder extracts patch features from each image and encodes station-centric geography. A temporal encoder and masked attention pooling aggregate images over the historical patches used by PVFM:

$$\mathbf{S} = \mathrm{Enc}_{\mathrm{sat}}(\mathbf{I}, c, m). \tag{37}$$

The context c reuses PVFM's positional, solar, and site embeddings. The mask excludes missing and padded frames from attention and pooling. Spatial tokens are retained within each aligned window and projected to the PVFM hidden size. Queries and keys additionally include station-centric spatial positions.

The scalar gate g is shared across feature channels. It is computed from the PV tokens, satellite tokens, valid-frame fraction, and daylight fraction of the window. It is set to zero when either PV tokens or satellite frames are missing. The affine residual projection Proj_0 has zero-initialized weights and bias, so the adapter initially leaves the frozen backbone unchanged.

Let θ denote frozen PVFM parameters, ϕ the satellite encoder, cross-attention, residual projection, and gate, and Ω the set of valid forecast targets. Fine-tuning updates only ϕ by minimizing the masked squared error of the median forecast:

$$\min_{\phi} \frac{1}{|\Omega|} \sum_{(b,h)\in\Omega} \big(y_{b,h} - f_\theta(\mathbf{Z}'_{\mathrm{PV}}, \mathbf{M}^{\mathrm{fut}})_h\big)^2. \tag{38}$$

Gradients pass through the frozen decoder into the adapter, while only ϕ is updated.

Each station-aligned image is a $64 \times 64 \times 4$ multispectral patch

B.2.3 Training and Optimization Details

Sampling Strategy. Following the task-distribution formulation used in universal time-series pretraining (Woo et al., 2024), we sample the temporal resolution, context length, and forecast horizon from predefined distributions rather than training on a fixed set of (C, H) pairs.

We first sample the temporal resolution $\Delta \in \{1\,\mathrm{h}, 15\,\mathrm{min}\}$ with probabilities 0.75 and 0.25. Conditional on Δ, we sample the forecast horizon from a mixture of task-aligned and uniformly distributed candidate horizons, followed by a compatible context length. Let $\tau = (\Delta, C, H)$ denote a pretraining task, where Δ is the temporal resolution. We factorize its sampling distribution as

$$p(\tau) = p(\Delta)\, p(H \mid \Delta)\, p(C \mid H, \Delta),$$

with

$$p(\Delta = 1\,\mathrm{h}) = 0.75, \qquad p(\Delta = 15\,\mathrm{min}) = 0.25.$$

To reflect representative forecasting needs across different temporal scales in PV applications, the horizon distribution places higher probability on several task-relevant forecast lengths while retaining coverage of other admissible horizons. Specifically, representative horizons are sampled with probability 0.6, while the remaining samples use horizons randomly drawn from the admissible range. For hourly data, we consider 6-, 24-, and 168-hour horizons, whereas for 15-minute data we consider 1-, 4-, 6-, and 24-hour horizons. These horizons correspond to commonly studied ultra-short-term, intraday, day-ahead, and longer-range PV forecasting settings.

To reduce sampling bias caused by substantial differences in data availability across regions, we further reweight regional sampling according to the estimated number of valid temporal windows. Regional probabilities are adjusted using an inverse-supply exponent of 0.25 and capped at 0.15. Within each selected region, valid sites and temporal windows are then sampled subject to the corresponding context–forecast requirements and data-quality constraints.

Optimization. PVFM is trained for 80,000 optimization steps using Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, zero weight decay, and a peak learning rate of 1×10^{-4}. The learning rate follows a warmup–stable–decay schedule with 5,000 warmup steps, 59,000 stable steps, and a 16,000-step cosine decay to 1×10^{-6}. Training minimizes the mean pinball loss over the 11 predictive quantiles on valid target positions. Training uses FP16 mixed precision and distributed data parallelism on four NVIDIA A100-SXM4 GPUs with 80 GB memory each. The per-GPU batch size is 256, resulting in a global batch size of 1,024.

B.3 Baseline Implementation Details

The compared methods are organized into three baseline groups: time-series forecasting models, PV-domain forecasting models, and pretrained TSFMs, with PVFM shown separately as the proposed model. Table 3 summarizes the input information available to each evaluated model, covering historical PV, historical covariates, and future covariates. We next describe the implementation and training settings for the full-shot forecasting models, followed by the evaluation configurations of the publicly pretrained TSFMs.

B.3.1 Full-Shot Forecasting Models

The time-series forecasting models and PV-domain forecasting models are evaluated under the target-site full-shot setting.

Table 3: Input information and future-NWP fusion used by full-shot baselines and PVFM under the evaluation protocol. Blue ticks denote our FC/TC augmentation, where FC and TC are feature and time concatenation, respectively.

Group	Model	Hist_PV	Hist_Cov	Future_Cov	Enhance
Time-series forecasting models	LightTS Zhang et al. (2022)	✓	✓	–	✓
	PatchTST Nie et al. (2023)	✓	✓	–	✓
	Crossformer Zhang & Yan (2023)	✓	✓	–	✓
	iTransformer Liu et al. (2024a)	✓	✓	–	✓
	TimeMixer Wang et al. (2024a)	✓	✓	–	✓
	TiDE Das et al. (2023)	✓	✓	✓	–
	TimeXer Wang et al. (2024b)	✓	✓	✓	–
	DAG Qiu et al. (2026)	✓	✓	✓	–
	GCGNet Li et al. (2026)	✓	✓	✓	–
PV-domain	Cross-Unet Zhang et al. (2026)	✓	✓	✓	–
	FusionSF Ma et al. (2024)	✓	✓	✓	–
Pretrained TSFMs	Chronos-2 Ansari et al. (2025)	✓	✓	✓	–
	Moirai-2 Liu et al. (2025)	✓	✓	✓	–
	TiRex-2 Podest et al. (2026)	✓	✓	✓	–
	TimesFM-3 Jain & Sen (2026)	✓	✓	✓	–
	CITRAS-FM Yamaguchi et al. (2026)	✓	✓	✓	–
	TabPFN-TS3 Grinsztajn et al. (2026); Hoo et al. (2025)	✓	✓	✓	–
Proposed	PVFM	✓	✓	✓	–

Future-covariate adaptation For time-series models without native support for future covariates, we use FC or TC to incorporate future NWP information. Specifically, let the historical PV sequence, historical weather covariates, and future NWP covariates be

$$\mathbf{p} \in \mathbb{R}^{C\times 1}, \qquad \mathbf{W}^{\text{hist}} \in \mathbb{R}^{C\times d}, \qquad \mathbf{W}^{\text{fut}} \in \mathbb{R}^{H\times d}, \tag{39}$$

where C and H denote the context and forecast lengths, respectively, and $d = 6$ is the number of weather variables. The batch dimension is omitted for clarity.

Feature concatenation (FC). FC first projects the H-step future NWP sequence to the context length C:

$$\widetilde{\mathbf{W}}_{\text{FC}} = \mathbf{A}_{\text{FC}}\mathbf{W}^{\text{fut}} + \mathbf{b}_{\text{FC}}\mathbf{1}_d^\top, \qquad \mathbf{A}_{\text{FC}} \in \mathbb{R}^{C\times H}, \quad \mathbf{b}_{\text{FC}} \in \mathbb{R}^{C\times 1}. \tag{40}$$

The projected future covariates are then concatenated with historical weather and PV along the feature dimension:

$$\mathbf{Z}_{\text{FC}} = \left[\mathbf{W}^{\text{hist}} \,\middle|\, \widetilde{\mathbf{W}}_{\text{FC}} \,\middle|\, \mathbf{p}\right] \in \mathbb{R}^{C\times(2d+1)}. \tag{41}$$

Time concatenation (TC). TC retains a separate future segment of length H. The future NWP sequence is projected as

$$\widetilde{\mathbf{W}}_{\text{TC}} = \mathbf{A}_{\text{TC}}\mathbf{W}^{\text{fut}} + \mathbf{b}_{\text{TC}}\mathbf{1}_d^\top, \qquad \mathbf{A}_{\text{TC}} \in \mathbb{R}^{H\times H}, \quad \mathbf{b}_{\text{TC}} \in \mathbb{R}^{H\times 1}, \tag{42}$$

and concatenated with the historical segment along the temporal dimension:

$$\mathbf{Z}_{\text{TC}} = \begin{bmatrix} \mathbf{W}^{\text{hist}} & \mathbf{0}_{C\times d} & \mathbf{p} \\ \mathbf{0}_{H\times d} & \widetilde{\mathbf{W}}_{\text{TC}} & \mathbf{0}_{H\times 1} \end{bmatrix} \in \mathbb{R}^{(C+H)\times(2d+1)}. \tag{43}$$

Here, $\mathbf{A}$ and $\mathbf{b}$ are learnable temporal-projection parameters shared across weather variables, and $\mathbf{1}_d$ denotes an all-ones vector. The future PV channel in TC is zero padded and therefore contains no ground-truth future PV information.

Probabilistic prediction head. For probabilistic forecasting, the full-shot baselines use a q9 quantile prediction head with levels $\{0.1, 0.2, \ldots, 0.9\}$. Quantile crossing is prevented through a non-crossing parameterization, and the models are optimized using the pinball loss. The median prediction, $q_{0.5}$, is used as the point forecast when computing deterministic metrics.

Table 4: Baseline architectures and registered parameter counts for 24-hour forecasting (nine output quantiles). Dagger marks explicitly reduced-capacity variants.

Family	Model	d_{model}	d_{ff}	Heads	Layers (E/D)	Params (M)	Other settings
Native models	Cross-Unet†	192	384	4	3/4	7.7861	segment=24
	FusionSF*	64/128	256/512	4/6	4/4	3.3483	VQ off; 9 heads
	TiDE	256	256	–	1/2	0.4995	residual MLP
	TimeXer	256	512	8	1/–	1.0972	patch=16; factor=3
	DAG	64	256	4	1/–	0.3515	patch/stride=48/48
	GCGNet	512	256	4	1/–	3.6626	patch=24; rank=4
FC/TC-augmented	PatchTST(FC)	512	2048	8	2/–	7.3113	patch=16, stride=8
	PatchTST(TC)	512	2048	8	2/–	7.6419	patch=16, stride=8
	iTransformer(FC)	512	2048	8	2/–	6.4558	–
	iTransformer(TC)	512	2048	8	2/–	6.4669	–
	Crossformer(FC)†	192	768	8	2/3	6.0236	segment=12
	Crossformer(TC)†	192	768	8	2/3	6.0280	segment=12
	TimeMixer(FC)	512	2048	–	2/–	8.5093	CI=0; downsample=2
	TimeMixer(TC)	512	2048	–	2/–	8.5470	CI=0; downsample=2
	LightTS(FC)	512	–	–	–	0.1372	chunk=24; 3 IE blocks
	LightTS(TC)	512	–	–	–	0.1412	chunk=24; 3 IE blocks

Table 5: Architecture configurations and parameter counts of the multi-modal forecasting models. Slash-separated entries denote encoder/decoder settings. Parameter counts are reported in millions. S/T/F denotes spatial/temporal/fusion blocks.

Model	Branch	d_{model}	d_{ff}	Heads	Layers	Params (M)	Other settings
FusionSF-SAT	Enc./Dec.	64/128	256/512	8/6	12/4	11.154	8×8 patches; PV/sat 8-level ResidualVQ; codebook 1024; future NWP
CrossViViT	Enc./Dec.	256/128	256/128	8/6	12/4	36.987	8×8 patches; time-aligned NWP; RoPE; ReLU output
MCloudNet	Enc.	512	2048	8	2	42.940	Three ResNet-18 branches; TV-L1 flow; db4 wavelets; NWP/cloud fusion
Solar-VLM	Main	128	512	4	3×2	2244.430	Frozen Qwen; cross-site attention; memory=100; top-k=5; single-site input
	Vision	2048	8192	–	–		
	GAT	–	–	1	3		
PVFM-Sat	Main E/D	256/256	1024/1024	4/4	6/6	16.214	8×8 patches; gated residual fusion;
	Sat.	128	512	4	1 each (S/T/F)		

Table 4 reports the architecture-specific configurations and registered parameter counts of the full-shot baselines. Because the size of the prediction head depends on the forecast horizon, the reported parameter counts correspond to the 24-hour forecasting task. Counts include input adapters and quantile prediction heads but exclude buffers and optimizer states. Dagger-marked Crossformer and Cross-Unet configurations use reduced model widths, with their dimensions reported explicitly in the table.

Training and Optimization Details. For optimization, the station-wise full-shot models are trained for up to 50 epochs with a batch size of 128, an initial learning rate of 10^{-4}, early-stopping patience of 3, the type1 learning-rate schedule, and 16 data-loader workers. The same optimization settings are used across the applicable time-series forecasting models and their FC/TC configurations, with architecture-specific hyperparameters set separately for each model.

B.3.2 Multi-Modal Baselines

We consider FusionSF-SAT, CrossViViT, MCloudNet, and Solar-VLM as multi-modal baselines for satellite-enhanced PV forecasting. Their main architectural configurations and parameter counts are summarized in Table 5, together with the corresponding configuration of PVFM-Sat.

B.3.3 Pretrained Time-Series Foundation Models

The following publicly available TSFMs are evaluated without parameter updates on the unseen PV sites:

- **Chronos-2 (119.48M)** Ansari et al. (2025): an encoder-only Transformer with group attention for zero-shot univariate, multivariate, and covariate-informed forecasting.
- **Moirai-2 (11.39M)** Liu et al. (2025): a compact decoder-only Transformer for zero-shot forecasting with multivariate historical inputs and quantile output.
- **TiRex-2 (82.50M)** Podest et al. (2026): a recurrent xLSTM foundation model supporting multivariate forecasting with past and known future covariates, and producing nine forecast quantiles.
- **TimesFM-3 (330.71M)** Jain & Sen (2026): a patch-based decoder-only Transformer with alternating temporal and variate attention, supporting known future covariates and quantile forecasts.
- **CITRAS-FM (7.22M)** Yamaguchi et al. (2026): a patch-based decoder-only Transformer with shifted cross-variate attention for covariate-informed probabilistic forecasting.
- **TabPFN-TS3 (∼58M)** Grinsztajn et al. (2026); Hoo et al. (2025): a time-series-adapted prior-data fitted network that forms a predictor from the supplied context and supports known future covariates.

C Additional Results

We provide additional evaluations of PVFM from multiple perspectives, covering forecasting performance, cross-region generalization, architecture and scaling behavior, PV-specific forecasting scenarios, qualitative case studies, and robustness across random seeds.

C.1 Full Forecasting Results

Table 6: Complementary forecasting performance across the five tasks in Table 1, reported using RMSE and average quantile loss (AQL). Shaded cells indicate the best result within each evaluation protocol (full-shot and zero-shot), while red bold and underlined values denote the best and second-best results overall, respectively. Imp. denotes the relative improvement over the best non-PVFM baseline. For FC/TC-augmented baselines, only the better-performing variant is shown.

Setting	Model	1 h						15 min			
		$C_{24} \to H_6$		$C_{72} \to H_{24}$		$C_{336} \to H_{168}$		$C_{64} \to H_{16}$		$C_{96} \to H_{24}$	
		RMSE↓	AQL↓	RMSE↓	AQL↓	RMSE↓	AQL↓	RMSE↓	AQL↓	RMSE↓	AQL↓
Full-shot	Cross-Unet	0.1290	0.0361	0.1257	0.0360	0.1687	0.0504	0.0997	0.0229	0.1131	0.0265
	FusionSF	0.1262	0.0262	0.1230	0.0265	0.1525	0.0357	0.0975	0.0184	0.1152	0.0228
	TiDE	0.1654	0.0506	0.1461	0.0459	0.1592	0.0504	0.1543	0.0410	0.1201	0.0312
	TimeXer	0.1194	0.0330	0.1211	0.0325	0.1344	0.0348	0.1093	0.0271	0.1103	0.0264
	DAG	0.1277	0.0405	0.1398	0.0425	0.1652	0.0483	0.1084	0.0275	0.1092	0.0267
	GCGNet	0.1180	0.0349	0.1229	0.0374	0.1721	0.0528	0.1167	0.0327	0.1204	0.0327
	PatchTST (FC)	0.1179	0.0299	0.1246	0.0302	0.1360	0.0331	0.1152	0.0269	0.1129	0.0252
	iTransformer (TC)	0.1092	0.0280	0.1093	0.0293	0.1296	0.0379	0.1000	0.0239	0.1022	0.0238
	Crossformer (FC)	0.1117	0.0306	0.1401	0.0395	0.1909	0.0560	0.0935	0.0211	0.1009	0.0230
	TimeMixer (TC)	0.1195	0.0353	0.1278	0.0369	0.1418	0.0388	0.1057	0.0257	0.1071	0.0250
	LightTS (FC)	0.1461	0.0415	0.1686	0.0485	0.1911	0.0553	0.1049	0.0261	0.1095	0.0265
Zero-shot	CITRAS-FM	0.1278	0.0249	0.1447	0.0294	0.1191	0.0243	0.1638	0.0350	0.1570	0.0335
	Chronos-2	0.1514	0.0331	0.1175	0.0238	0.1128	0.0224	0.1730	0.0347	0.1488	0.0294
	TiRex-2	0.1790	0.0409	0.1410	0.0316	0.1240	0.0265	0.1599	0.0335	0.1495	0.0310
	Moirai-2	0.2487	0.0650	0.1732	0.0475	0.1527	0.0427	0.1714	0.0418	0.2044	0.0504
	TimesFM-3	0.1890	0.0470	0.1333	0.0301	0.1259	0.0268	0.1864	0.0361	0.1487	0.0292
	TabPFN-TS3	0.2020	0.0484	0.1517	0.0324	0.1434	0.0296	0.1793	0.0412	0.2006	0.0469
	PVFM	**0.0970**	**0.0177**	**0.0975**	**0.0176**	**0.0977**	**0.0177**	**0.0911**	**0.0166**	**0.0958**	**0.0173**
Imp. PVFM (%)		**+11.2**	**+28.9**	**+10.8**	**+26.1**	**+13.4**	**+21.0**	**+2.6**	**+9.8**	**+5.1**	**+24.1**

C.2 Cross-Region Generalization

Table 7: Hong Kong (HK) and Australia regional generalization across three hourly and two 15-minute tasks. Entries are station-equal MAE (↓); lower is better. Red bold marks the best unrounded result in each column, and underlining marks the second-best unrounded result. Blue shading marks the best full-shot result, and purple shading marks the best result within the joint zero-shot group.

Setting	Model	1 h						15 min			
		$C_{24} \to H_6$		$C_{72} \to H_{24}$		$C_{336} \to H_{168}$		$C_{64} \to H_{16}$		$C_{96} \to H_{24}$	
		HK	Australia	HK	Australia	HK	Australia	HK	Australia	HK	Australia
Full-shot	Cross-Unet	0.0624	0.1340	0.0667	0.1325	0.0831	0.1415	0.0512	0.1022	0.0533	0.1135
	FusionSF	0.0572	0.1288	0.0571	0.1241	0.0627	0.1296	0.0492	0.1059	0.0522	0.1164
	TiDE	0.0839	0.1794	0.0791	0.1368	0.0957	0.1445	0.0901	0.1535	0.0596	0.1092
	TimeXer	0.0646	0.1153	0.0703	0.1194	0.0829	0.1208	0.0634	0.1169	0.0579	0.1070
	DAG	0.0704	0.1233	0.0811	0.1389	0.0908	0.1405	0.0639	0.1148	0.0588	0.1044
	GCGNet	0.0604	0.1180	0.0656	0.1278	0.0988	0.1555	0.0710	0.1262	0.0665	0.1210
	PatchTST(TC)	0.0610	0.1114	0.0702	0.1186	0.0814	0.1250	0.0638	0.1263	0.0593	0.1041
	PatchTST(FC)	0.0623	0.1111	0.0716	0.1193	0.0830	0.1262	0.0662	0.1264	0.0605	0.1044
	iTransformer(TC)	0.0549	0.0963	0.0591	0.1031	0.0700	0.1152	0.0528	**0.0900**	0.0486	**0.0874**
	iTransformer(FC)	0.0562	0.0976	0.0600	0.1036	0.0701	0.1157	0.0539	0.0928	0.0499	0.0884
	Crossformer(TC)	0.0617	0.1192	0.0722	0.1427	0.0941	0.1607	**0.0488**	0.1030	0.0492	0.0949
	Crossformer(FC)	0.0579	0.1159	0.0735	0.1417	0.0936	0.1468	0.0489	0.0969	0.0483	0.0955
	TimeMixer(TC)	0.0655	0.1176	0.0711	0.1310	0.0734	0.1282	0.0544	0.0988	0.0519	0.0942
	TimeMixer(FC)	0.0663	0.1210	0.0729	0.1327	0.0766	0.1352	0.0579	0.1047	0.0543	0.0979
	LightTS(TC)	0.0811	0.1599	0.0884	0.1618	0.0990	0.1548	0.0580	0.1207	0.0572	0.1077
	LightTS(FC)	0.0801	0.1618	0.0935	0.1742	0.1052	0.1634	0.0592	0.1251	0.0579	0.1112
Zero-shot	CITRAS-FM	0.0602	0.1134	0.0608	0.1057	0.0642	0.0946	0.0948	0.1355	0.0829	0.1112
	Chronos-2	0.0740	0.1399	0.0602	0.0954	0.0636	0.0888	0.0906	0.1442	0.0700	0.1167
	Moirai-2	0.1510	0.2146	0.1131	0.1695	0.1077	0.1595	0.1018	0.1472	0.1228	0.1781
	TiRex-2	0.0768	0.1403	0.0585	0.0944	0.0578	0.0889	0.0867	0.1340	0.0700	0.1194
	TimesFM-3	0.0923	0.1737	0.0591	0.0976	0.0560	**0.0870**	0.0882	0.1483	0.0645	0.1156
	TabPFN-TS3	0.0913	0.1831	0.0570	0.0969	0.0606	0.0909	0.1072	0.1386	0.1113	0.1630
	PVFM (ours)	**0.0518**	**0.0929**	**0.0527**	**0.0901**	**0.0544**	0.0887	0.0492	0.0987	**0.0481**	0.0912

C.3 Architecture Comparisons

We further examine the scaling and architectural properties of PVFM from three complementary perspectives: (i) **model capacity**, (ii) **backbone architecture**, and (iii) **pretraining data scale**. These experiments investigate whether the observed performance gains arise from increased model capacity, the proposed architecture, or broader pretraining data.

C.3.1 Model Capacity Scaling

We first study how PVFM benefits from increased model capacity while keeping the overall architecture and training protocol unchanged. We progressively scale PVFM from 3.9M to 15.3M and 45.0M parameters by increasing the model width and, for the largest model, the network depth. The corresponding model configurations are summarized in Table 9. Figure 7 further shows the training behavior and forecasting performance as model capacity increases.

Table 9: PVFM model configurations used for the capacity scaling study. The model size is progressively increased from 3.9M to 45.0M parameters through changes in model width and depth, with the corresponding architectural hyperparameters reported for each scale.

Model	d_{model}	Heads	Enc.	Dec.	d_{ff}	Params (M)	Exact params
PVFM-3.9M	128	8	6	6	512	3.882	3,882,336
PVFM-15.3M	256	4	6	6	1,024	15.331	15,330,912
PVFM-45.0M	384	6	8	8	1,536	45.001	45,001,056

C.3.2 Architecture Comparison under Shared Pretraining

To isolate the effect of model architecture, we compare PVFM with alternative forecasting architectures under a shared pretraining setting. All models use the same pretraining data and evaluation

Table 8: Hong Kong and Australia regional generalization across three hourly and two 15-minute tasks. Entries are station-equal CRPS (↓); lower is better. All models are evaluated using the common Q9 grid. Red bold marks the best unrounded result in each column, and underlining marks the second-best unrounded result. Blue shading marks the best full-shot result, and purple shading marks the best result within the joint zero-shot group, including PVFM.

Setting	Model	1 h						15 min			
		$C_{24} \to H_6$		$C_{72} \to H_{24}$		$C_{336} \to H_{168}$		$C_{64} \to H_{16}$		$C_{96} \to H_{24}$	
		HK	Australia	HK	Australia	HK	Australia	HK	Australia	HK	Australia
Full-shot	Cross-Unet	0.0553	0.1063	0.0556	0.1039	0.0678	0.1088	0.0408	0.0806	0.0422	0.0886
	FusionSF	0.0409	0.0984	0.0410	0.0950	0.0447	0.0985	**0.0347**	0.0836	0.0367	0.0924
	TiDE	0.0764	0.1277	0.0726	0.1057	0.0762	0.1067	0.0723	0.1146	0.0507	0.0860
	TimeXer	0.0566	0.0916	0.0564	0.0909	0.0570	0.0866	0.0477	0.0875	0.0438	0.0786
	DAG	0.0674	0.1009	0.0681	0.1049	0.0685	0.1014	0.0490	0.0883	0.0448	0.0795
	GCGNet	0.0572	0.0916	0.0608	0.0993	0.0794	0.1127	0.0598	0.0992	0.0545	0.0927
	PatchTST(TC)	0.0498	0.0860	0.0502	0.0850	0.0537	0.0843	0.0474	0.0887	0.0422	0.0736
	PatchTST(FC)	0.0504	0.0861	0.0514	0.0858	0.0550	0.0852	0.0485	0.0896	0.0430	0.0742
	iTransformer(TC)	0.0493	0.0791	0.0523	0.0825	0.0601	0.0891	0.0434	**0.0744**	0.0393	0.0710
	iTransformer(FC)	0.0496	0.0793	0.0518	0.0828	0.0590	0.0885	0.0441	0.0768	0.0401	0.0716
	Crossformer(TC)	0.0532	0.0966	0.0603	0.1127	0.0749	0.1231	0.0390	0.0852	0.0389	0.0762
	Crossformer(FC)	0.0508	0.0923	0.0608	0.1121	0.0744	0.1153	0.0389	0.0799	0.0384	0.0774
	TimeMixer(TC)	0.0591	0.0940	0.0598	0.0998	0.0546	0.0937	0.0446	0.0812	0.0412	0.0755
	TimeMixer(FC)	0.0600	0.0957	0.0628	0.1050	0.0585	0.0989	0.0472	0.0855	0.0425	0.0781
	LightTS(TC)	0.0654	0.1215	0.0692	0.1198	0.0725	0.1105	0.0444	0.0964	0.0432	0.0825
	LightTS(FC)	0.0638	0.1230	0.0719	0.1275	0.0765	0.1164	0.0451	0.0987	0.0436	0.0843
Zero-shot	CITRAS-FM	0.0454	0.0856	0.0453	0.0776	0.0480	0.0716	0.0732	0.1061	0.0624	0.0855
	Chronos-2	0.0559	0.1038	0.0433	0.0684	0.0440	**0.0619**	0.0697	0.1122	0.0509	0.0851
	Moirai-2	0.1148	0.1672	0.0875	0.1216	0.0791	0.1116	0.0845	0.1162	0.0954	0.1326
	TiRex-2	0.0592	0.1062	0.0450	0.0721	0.0430	0.0670	0.0666	0.1040	0.0525	0.0888
	TimesFM-3	0.0721	0.1352	0.0437	0.0718	0.0406	0.0636	0.0679	0.1157	0.0471	0.0840
	TabPFN-TS3	0.0687	0.1456	0.0413	0.0694	0.0423	0.0636	0.0855	0.1134	0.0883	0.1326
	PVFM (ours)	**0.0376**	**0.0729**	**0.0376**	**0.0679**	**0.0383**	0.0654	0.0360	0.0778	**0.0346**	**0.0691**

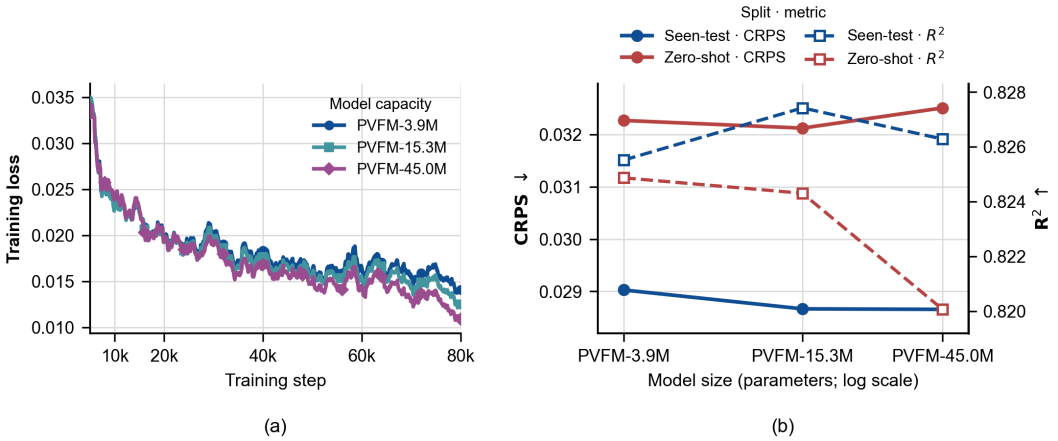

Figure 7: PVFM performance with increasing model capacity. (a) Training loss for different model sizes. (b) Five-task average CRPS and R^2 on seen and zero-shot stations. All models are trained with the same number of training steps and sample exposure.

protocol, reducing the influence of differences in data exposure and allowing a more controlled comparison of architectural design.

C.3.3 Pretraining Dataset Scaling

We further investigate the effect of pretraining data scale by training PVFM with progressively larger subsets of the pretraining corpus while keeping the model configuration fixed. This experiment examines how forecasting performance changes with increased data coverage and whether PVFM continues to benefit from broader pretraining.

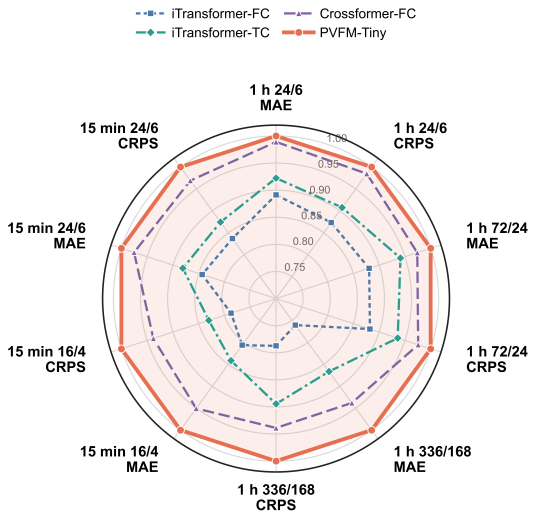

Figure 8: Architecture comparison under the shared the same pretraining protocol. The radar chart reports five-task mean station-equal zero-shot MAE and CRPS for iTransformer(FC), iTransformer(TC), Crossformer(FC), and PVFM.

Figure 9 examines data scaling by varying the number of pretraining stations from 187 to 746 under a fixed 80k-step training budget. Increasing station coverage consistently improves both deterministic and probabilistic forecasting performance. Specifically, MAE decreases from 0.04856 to 0.04507, corresponding to a 7.18% reduction, while CRPS decreases from 0.03501 to 0.03227, yielding a 7.83% reduction. The fitted scaling exponents are $\alpha = 0.054$ for MAE and $\alpha = 0.059$ for CRPS, indicating modest but consistent gains as the diversity of pretraining data increases. Because only three data scales are evaluated, these power-law fits should be interpreted as descriptive trends rather than evidence for reliable extrapolation.

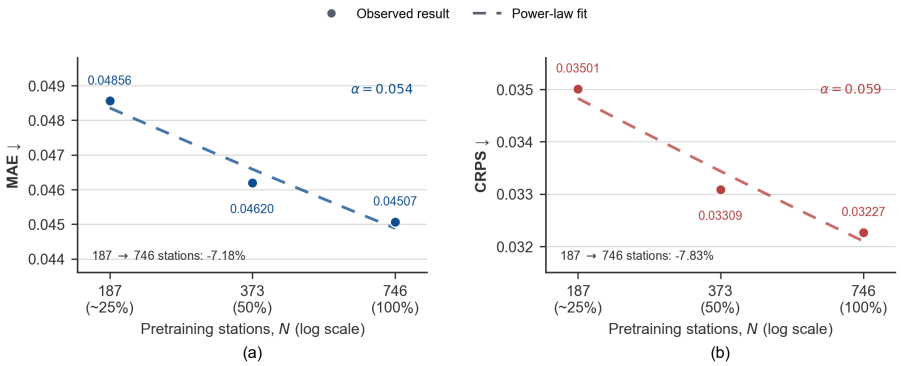

Figure 9: Pretraining Data Scaling with Increasing Station Coverage.

C.4 PV-Specific Forecasting Scenarios

Beyond the standard forecasting evaluation, we further examine PVFM in two PV-specific forecasting scenarios: ramp-event forecasting under rapid generation changes and multimodal fine-tuning with satellite imagery for short-term cloud-field information. We present the corresponding results below.

C.4.1 Ramp-Event Evaluation

We further analyze PVFM under ramp events, where rapid changes in PV generation pose additional challenges beyond overall forecasting accuracy. Figure 10 presents representative examples across

Table 10: Overall ramp-event benchmark on the $C_{24} \rightarrow H_6$ task. Higher POD/CSI and lower FAR/RMSE/MAE are better. Red bold marks the best unrounded result in each column. Blue shading marks the best full-shot result, and purple shading marks the best result within the joint zero-shot group, including PVFM.

Setting	Model	POD↑	FAR↓	CSI↑	Event RMSE↓	Event MAE↓
Full-shot	TiDE	0.555	0.390	0.408	0.209	0.164
	TimeXer	0.622	0.302	0.490	0.172	0.129
	DAG	0.592	0.314	0.461	0.183	0.142
	GCGNet	0.615	0.295	0.487	0.173	0.132
	PatchTST (FC)	0.574	0.244	0.484	0.172	0.128
	PatchTST (TC)	0.575	0.243	0.486	0.173	0.130
	iTransformer (FC)	0.613	0.239	0.516	0.167	0.119
	iTransformer (TC)	0.601	0.239	0.507	0.168	0.120
	Crossformer (FC)	0.633	0.315	0.490	0.163	0.124
	Crossformer (TC)	0.636	0.329	0.483	0.173	0.133
	TimeMixer (FC)	0.635	0.351	0.472	0.173	0.129
	TimeMixer (TC)	0.639	0.345	0.479	0.175	0.130
	LightTS (FC)	0.618	0.374	0.449	0.189	0.148
	LightTS (TC)	**0.667**	0.408	0.452	0.193	0.148
Zero-shot	CITRAS-FM	0.537	**0.212**	0.470	0.188	0.133
	Chronos-2	0.497	0.291	0.414	0.227	0.174
	TiRex-2	0.432	0.351	0.355	0.269	0.211
	Moirai-2	0.070	0.303	0.066	0.347	0.290
	TimesFM-3	0.426	0.372	0.342	0.274	0.217
	TabPFN-TS3	0.315	0.368	0.265	0.299	0.246
	PVFM	0.638	0.227	**0.538**	**0.157**	**0.110**

four PV sites, together with cloud cover and the identified up- and down-ramp periods. These cases illustrate the forecasts of PVFM and representative baselines around abrupt generation changes.

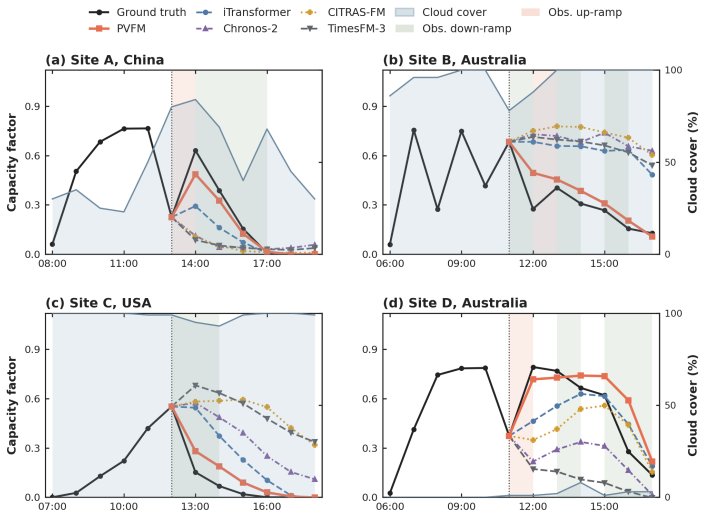

Figure 10: Case studies of ramp-event forecasting across four PV sites. Ground-truth PV generation and forecasts from PVFM and representative baselines are shown together with cloud cover and observed up- and down-ramp periods. The vertical dotted line marks the forecast start time.

Table 10 provides the corresponding quantitative comparison on the $C_{24} \rightarrow H_6$ task. PVFM achieves the highest overall CSI (0.538) and the lowest event RMSE (0.157) and MAE (0.110), while maintaining a POD of 0.638 and a FAR of 0.227. These results indicate that the zero-shot forecasting capability of PVFM extends beyond average prediction accuracy to rapid PV generation changes.

C.4.2 SATELLITE-IMAGE FINE-TUNING

We further evaluate whether PVFM can incorporate satellite imagery through multimodal fine-tuning. We compare satellite-based full-shot baselines, zero-shot PVFM without satellite inputs, and PVFM fine-tuned with historical satellite imagery on the $16 \rightarrow 4$ forecasting task. Table 11 reports the resulting performance. MAE, RMSE, and R^2 are first computed separately for each site and then averaged equally across sites

Table 11: Satellite-image fine-tuning results on the $16 \rightarrow 4$ forecasting task. Check marks indicate the input modalities used by each model, and bold denotes the best result in each column.

Setting	Model	NWP	Satellite	ERA5	MAE↓	RMSE↓	R^2 ↑
Full-shot	FusionSF	✓	✓	–	**0.0377**	0.0851	0.8593
	Solar-VLM	✓	✓	–	0.0410	0.0856	0.8602
	CrossViViT	✓	✓	–	0.0383	0.0866	0.8554
	MCloudNet	✓	✓	–	0.0433	0.0868	0.8563
Zero-shot	PVFM†	✓	–	✓	0.0427	0.0907	0.8417
	PVFM‡	✓	–	–	0.0440	0.0923	0.8368
Fine-tuned	PVFM	✓	✓	–	0.0390	**0.0842**	**0.8628**

C.5 CASE STUDY

To qualitatively examine PVFM across different regions, temporal resolutions, and forecast horizons, we present representative forecasting examples from seven geographically diverse regions. Each case includes the five evaluation tasks considered in this work, spanning hourly and 15-minute resolutions with forecast horizons from 4 to 168 hours. Figures 11–17 visualize the historical PV context, future ground truth, PVFM median prediction, and the P10–P90 prediction interval. Across these examples, PVFM captures the diurnal generation profile and its variations over different forecast horizons, while the prediction intervals reflect increasing uncertainty when future PV generation becomes more variable.

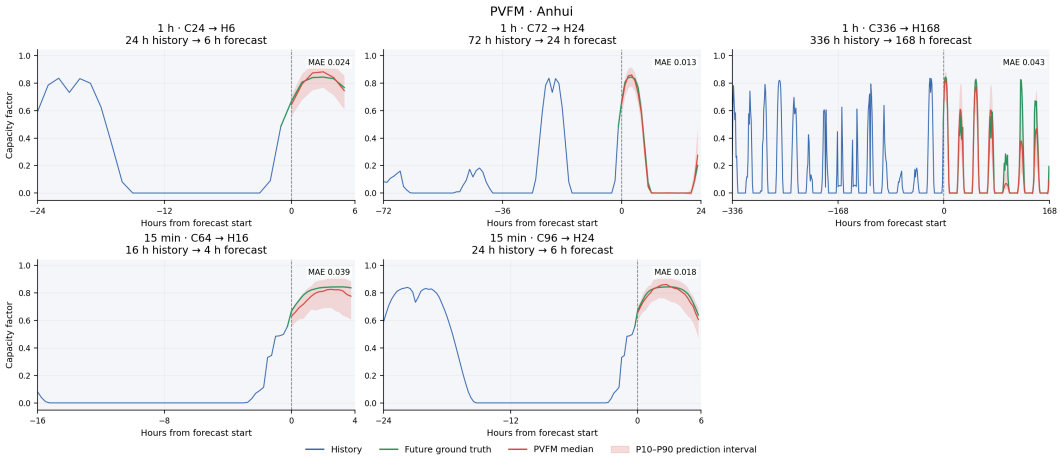

Figure 11: PVFM forecasts for a representative PV site in Anhui across the five evaluation tasks. Shaded areas denote the P10–P90 prediction intervals.

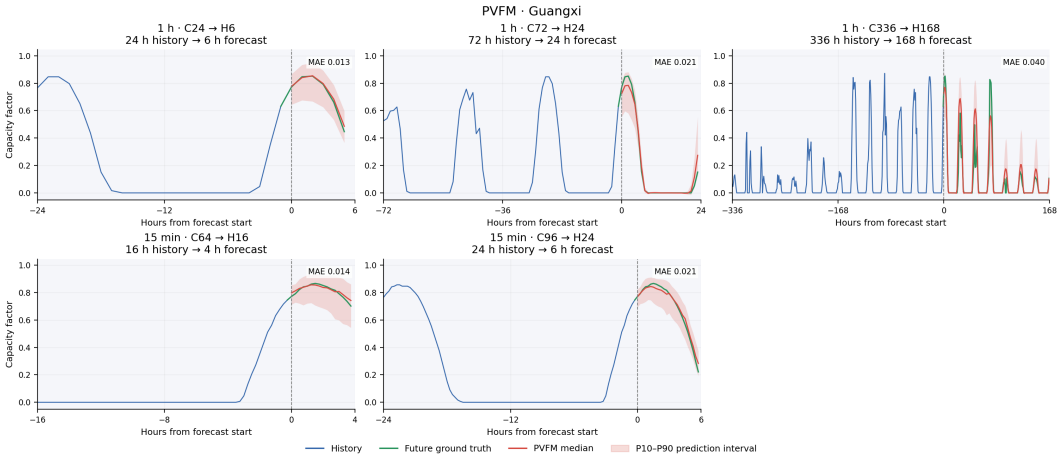

Figure 12: PVFM forecasts for a representative PV site in Guangxi across the five evaluation tasks. Shaded areas denote the P10–P90 prediction intervals.

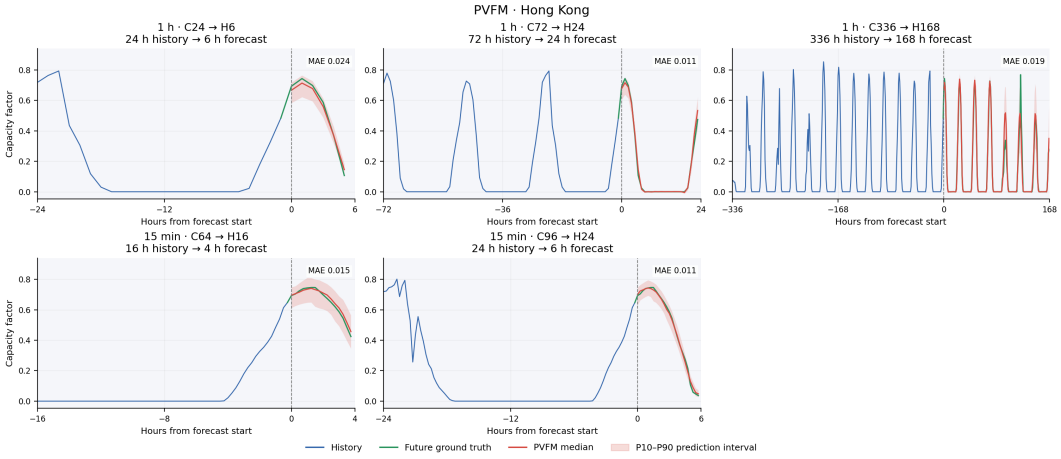

Figure 13: PVFM forecasts for a representative PV site in Hong Kong across the five evaluation tasks. Shaded areas denote the P10–P90 prediction intervals.

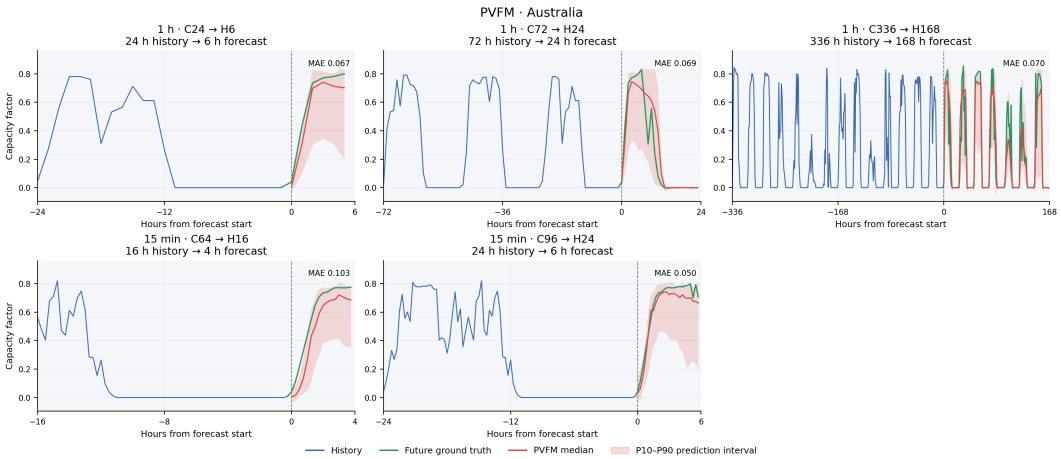

Figure 14: PVFM forecasts for a representative PV site in Australia across the five evaluation tasks. Shaded areas denote the P10–P90 prediction intervals.

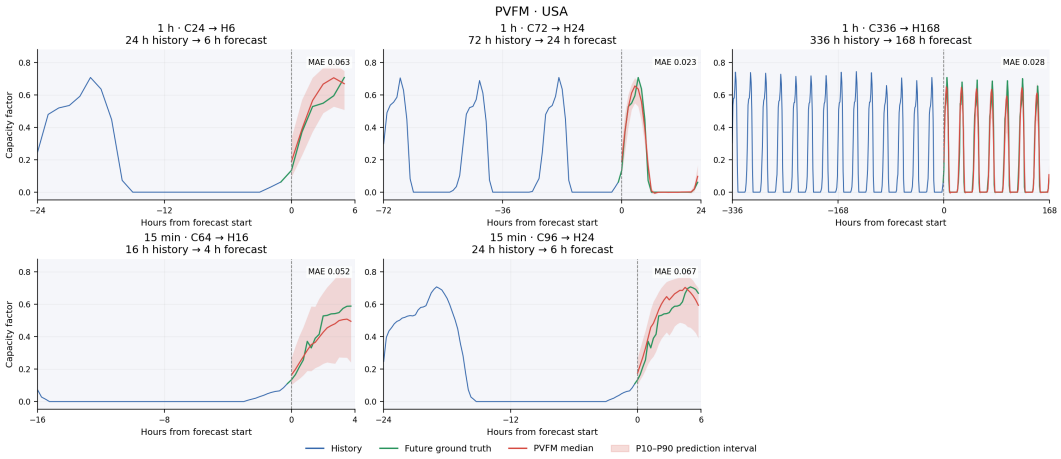

Figure 15: PVFM forecasts for a representative PV site in the United States across the five evaluation tasks. Shaded areas denote the P10–P90 prediction intervals.

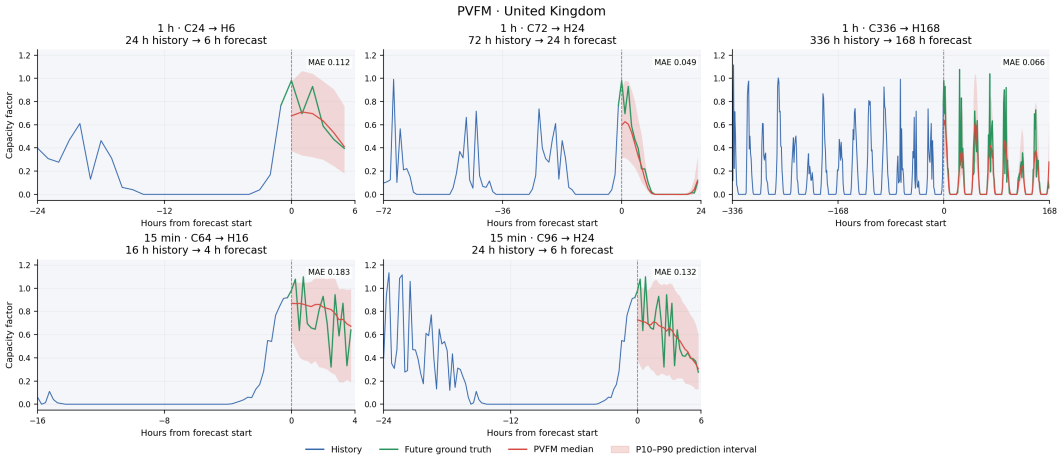

Figure 16: PVFM forecasts for a representative PV site in the United Kingdom across the five evaluation tasks. Shaded areas denote the P10–P90 prediction intervals.

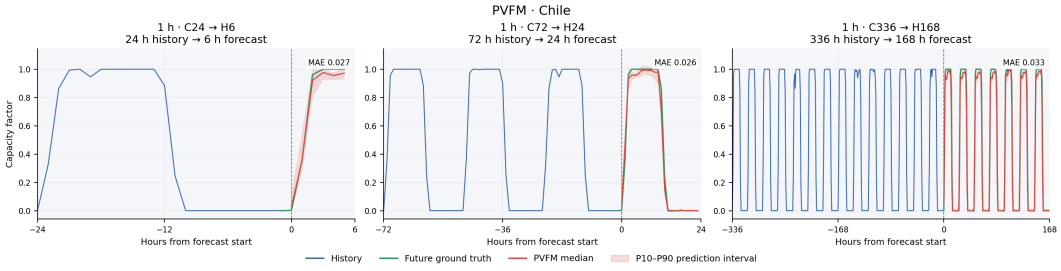

Figure 17: PVFM forecasts for a representative PV site in Chile across the five evaluation tasks. Shaded areas denote the P10–P90 prediction intervals.

C.6 Seed Variability

For the hourly full-shot results, all models are evaluated with three random seeds (2021, 2022, and 2023) on the fixed 99-station cohort. Metrics are first averaged equally across stations within each seed and then averaged across the three seeds. All full-shot runs use native q9 predictions, q50 point forecasts, a maximum of 50 epochs, and early stopping with patience 3. Bold and underlined

values denote the best and second-best results, respectively. Detailed multi-seed results are reported in Tables 12–14 for the three hourly tasks and Tables 15 and 16 for the two 15-minute tasks.

Table 12: Combined seed variability for 6-hour forecasting across all full-shot baselines. Entries report mean ± sample standard deviation across seed-level station-equal metrics.

Setting	Model	R^2 ↑	MAE↓	RMSE↓	CRPS↓	AQL↓
Full-shot	Cross-Unet	0.7133 ± 0.0016	0.0798 ± 0.0008	0.1290 ± 0.0003	0.0665 ± 0.0010	0.0361 ± 0.0006
	FusionSF	0.7132 ± 0.0028	0.0655 ± 0.0003	0.1262 ± 0.0007	$\mathbf{0.0485} \pm 0.0001$	$\mathbf{0.0262} \pm 0.0000$
	TiDE	0.5571 ± 0.1257	0.1133 ± 0.0190	0.1654 ± 0.0249	0.0924 ± 0.0119	0.0506 ± 0.0064
	TimeXer	0.7505 ± 0.0018	0.0682 ± 0.0004	0.1194 ± 0.0004	0.0605 ± 0.0005	0.0330 ± 0.0003
	DAG	0.7274 ± 0.0116	0.0816 ± 0.0022	0.1277 ± 0.0025	0.0738 ± 0.0040	0.0405 ± 0.0022
	GCGNet	0.7615 ± 0.0110	0.0698 ± 0.0028	0.1180 ± 0.0027	0.0637 ± 0.0018	0.0349 ± 0.0010
	PatchTST(FC)	0.7548 ± 0.0005	0.0652 ± 0.0003	0.1179 ± 0.0001	0.0550 ± 0.0001	0.0299 ± 0.0000
	PatchTST(TC)	0.7511 ± 0.0003	0.0645 ± 0.0005	0.1189 ± 0.0000	0.0540 ± 0.0000	0.0294 ± 0.0000
	iTransformer(FC)	$\underline{0.7883} \pm 0.0006$	$\underline{0.0568} \pm 0.0001$	$\underline{0.1095} \pm 0.0002$	$\underline{0.0518} \pm 0.0001$	$\underline{0.0281} \pm 0.0001$
	iTransformer(TC)	$\mathbf{0.7890} \pm 0.0045$	$\mathbf{0.0559} \pm 0.0004$	$\mathbf{0.1092} \pm 0.0010$	$\mathbf{0.0517} \pm 0.0004$	$\mathbf{0.0280} \pm 0.0002$
	Crossformer(FC)	0.7833 ± 0.0046	0.0667 ± 0.0008	0.1117 ± 0.0012	0.0565 ± 0.0008	0.0306 ± 0.0005
	Crossformer(TC)	0.7652 ± 0.0088	0.0699 ± 0.0018	0.1160 ± 0.0022	0.0588 ± 0.0013	0.0319 ± 0.0007
	TimeMixer(FC)	0.7470 ± 0.0101	0.0745 ± 0.0024	0.1205 ± 0.0022	0.0662 ± 0.0023	0.0360 ± 0.0013
	TimeMixer(TC)	0.7533 ± 0.0051	0.0728 ± 0.0023	0.1195 ± 0.0018	0.0651 ± 0.0023	0.0353 ± 0.0013
	LightTS(FC)	0.6367 ± 0.0400	0.0991 ± 0.0079	0.1461 ± 0.0081	0.0769 ± 0.0059	0.0415 ± 0.0033
	LightTS(TC)	0.6155 ± 0.0315	0.1008 ± 0.0064	0.1509 ± 0.0065	0.0785 ± 0.0048	0.0424 ± 0.0027

Table 13: Combined seed variability for 24-hour forecasting across all full-shot baselines. Entries report mean ± sample standard deviation across seed-level station-equal metrics.

Setting	Model	R^2 ↑	MAE↓	RMSE↓	CRPS↓	AQL↓
Full-shot	Cross-Unet	0.7264 ± 0.0041	0.0815 ± 0.0013	0.1257 ± 0.0010	0.0662 ± 0.0007	0.0360 ± 0.0004
	FusionSF	0.7280 ± 0.0028	0.0651 ± 0.0003	0.1230 ± 0.0005	$\mathbf{0.0486} \pm 0.0002$	$\mathbf{0.0265} \pm 0.0001$
	TiDE	0.6435 ± 0.0127	0.0981 ± 0.0039	0.1461 ± 0.0032	0.0836 ± 0.0021	0.0459 ± 0.0011
	TimeXer	0.7435 ± 0.0018	0.0736 ± 0.0002	0.1211 ± 0.0004	0.0597 ± 0.0001	0.0325 ± 0.0001
	DAG	0.6750 ± 0.0098	0.0963 ± 0.0009	0.1398 ± 0.0022	0.0775 ± 0.0012	0.0425 ± 0.0007
	GCGNet	0.7433 ± 0.0078	0.0780 ± 0.0034	0.1229 ± 0.0023	0.0683 ± 0.0011	0.0374 ± 0.0006
	PatchTST(FC)	0.7241 ± 0.0003	0.0744 ± 0.0001	0.1246 ± 0.0001	0.0553 ± 0.0001	0.0302 ± 0.0000
	PatchTST(TC)	0.7156 ± 0.0008	0.0732 ± 0.0001	0.1264 ± 0.0002	$\mathbf{0.0539} \pm 0.0001$	$\underline{0.0294} \pm 0.0001$
	iTransformer(FC)	$\underline{0.7851} \pm 0.0002$	$\underline{0.0605} \pm 0.0002$	$\underline{0.1100} \pm 0.0001$	0.0544 ± 0.0001	0.0295 ± 0.0001
	iTransformer(TC)	$\mathbf{0.7881} \pm 0.0014$	$\mathbf{0.0595} \pm 0.0001$	$\mathbf{0.1093} \pm 0.0003$	$\underline{0.0541} \pm 0.0001$	$\mathbf{0.0293} \pm 0.0001$
	Crossformer(FC)	0.6495 ± 0.0128	0.0912 ± 0.0028	0.1401 ± 0.0029	0.0728 ± 0.0016	0.0395 ± 0.0009
	Crossformer(TC)	0.6495 ± 0.0042	0.0904 ± 0.0002	0.1396 ± 0.0011	0.0724 ± 0.0004	0.0392 ± 0.0002
	TimeMixer(FC)	0.7120 ± 0.0035	0.0861 ± 0.0005	0.1302 ± 0.0004	0.0732 ± 0.0009	0.0400 ± 0.0005
	TimeMixer(TC)	0.7227 ± 0.0015	0.0824 ± 0.0005	0.1278 ± 0.0002	0.0680 ± 0.0002	0.0369 ± 0.0002
	LightTS(FC)	0.5290 ± 0.0461	0.1196 ± 0.0070	0.1686 ± 0.0086	0.0895 ± 0.0044	0.0485 ± 0.0024
	LightTS(TC)	0.5699 ± 0.0385	0.1116 ± 0.0072	0.1606 ± 0.0080	0.0845 ± 0.0049	0.0457 ± 0.0026

Table 14: Combined seed variability for 168-hour forecasting across all full-shot baselines. Entries report mean ± sample standard deviation across seed-level station-equal metrics.

Setting	Model	R^2 ↑	MAE↓	RMSE↓	CRPS↓	AQL↓
Full-shot	Cross-Unet	0.4166 ± 0.0187	0.1172 ± 0.0019	0.1687 ± 0.0022	0.0928 ± 0.0017	0.0504 ± 0.0009
	FusionSF	0.5892 ± 0.0044	0.0892 ± 0.0008	0.1525 ± 0.0009	0.0650 ± 0.0014	0.0357 ± 0.0007
	TiDE	0.5463 ± 0.0443	0.1133 ± 0.0083	0.1592 ± 0.0078	0.0916 ± 0.0047	0.0504 ± 0.0026
	TimeXer	0.6735 ± 0.0013	0.0851 ± 0.0002	0.1344 ± 0.0003	0.0637 ± 0.0001	0.0348 ± 0.0000
	DAG	0.5168 ± 0.0221	0.1159 ± 0.0023	0.1652 ± 0.0028	0.0884 ± 0.0012	0.0483 ± 0.0006
	GCGNet	0.4972 ± 0.0890	0.1244 ± 0.0160	0.1721 ± 0.0166	0.0964 ± 0.0088	0.0528 ± 0.0048
	PatchTST(FC)	0.6652 ± 0.0009	0.0858 ± 0.0001	0.1360 ± 0.0002	$\underline{0.0606} \pm 0.0001$	$\underline{0.0331} \pm 0.0001$
	PatchTST(TC)	0.6608 ± 0.0005	0.0832 ± 0.0001	0.1369 ± 0.0001	$\mathbf{0.0590} \pm 0.0001$	$\mathbf{0.0322} \pm 0.0000$
	iTransformer(FC)	$\underline{0.6927} \pm 0.0024$	$\underline{0.0808} \pm 0.0006$	$\underline{0.1303} \pm 0.0006$	0.0700 ± 0.0003	0.0383 ± 0.0002
	iTransformer(TC)	$\mathbf{0.6977} \pm 0.0016$	$\mathbf{0.0795} \pm 0.0001$	$\mathbf{0.1296} \pm 0.0004$	0.0694 ± 0.0001	0.0379 ± 0.0001
	Crossformer(FC)	0.3113 ± 0.0132	0.1348 ± 0.0010	0.1909 ± 0.0012	0.1033 ± 0.0004	0.0560 ± 0.0002
	Crossformer(TC)	0.2731 ± 0.0249	0.1372 ± 0.0031	0.1964 ± 0.0035	0.1050 ± 0.0019	0.0568 ± 0.0010
	TimeMixer(FC)	0.6211 ± 0.0041	0.1000 ± 0.0004	0.1464 ± 0.0006	0.0786 ± 0.0010	0.0429 ± 0.0006
	TimeMixer(TC)	0.6415 ± 0.0033	0.0923 ± 0.0009	0.1418 ± 0.0007	0.0716 ± 0.0005	0.0388 ± 0.0003
	LightTS(FC)	0.3520 ± 0.0248	0.1360 ± 0.0026	0.1911 ± 0.0034	0.1022 ± 0.0018	0.0553 ± 0.0009
	LightTS(TC)	0.4121 ± 0.0214	0.1272 ± 0.0025	0.1824 ± 0.0029	0.0962 ± 0.0018	0.0521 ± 0.0009

Table 15: Full-shot 15-minute 4-hour forecasting: $C_{64} \rightarrow H_{16}$. Entries report mean ± sample SD across three complete seeds. Best results are shown in bold and second-best results are underlined.

Setting	Model	R^2 ↑	MAE↓	RMSE↓	CRPS↓	AQL↓
Full-shot	Cross-Unet	0.7977 ± 0.0023	0.0527 ± 0.0006	0.0997 ± 0.0005	0.0426 ± 0.0004	0.0229 ± 0.0002
	FusionSF	$\underline{0.8015 \pm 0.0031}$	$\mathbf{0.0468 \pm 0.0005}$	0.0975 ± 0.0007	$\mathbf{0.0339 \pm 0.0002}$	$\mathbf{0.0184 \pm 0.0001}$
	TiDE	0.5354 ± 0.0377	0.0920 ± 0.0041	0.1543 ± 0.0061	0.0756 ± 0.0035	0.0410 ± 0.0018
	TimeXer	0.7652 ± 0.0000	0.0621 ± 0.0002	0.1093 ± 0.0000	0.0495 ± 0.0000	0.0271 ± 0.0000
	DAG	0.7680 ± 0.0039	0.0652 ± 0.0016	0.1084 ± 0.0010	0.0506 ± 0.0009	0.0275 ± 0.0005
	GCGNet	0.7317 ± 0.0206	0.0688 ± 0.0047	0.1167 ± 0.0047	0.0600 ± 0.0025	0.0327 ± 0.0014
	PatchTST(FC)	0.7386 ± 0.0003	0.0658 ± 0.0002	0.1152 ± 0.0000	0.0492 ± 0.0000	0.0269 ± 0.0000
	PatchTST(TC)	0.7352 ± 0.0004	0.0641 ± 0.0002	0.1159 ± 0.0001	0.0482 ± 0.0001	0.0263 ± 0.0001
	iTransformer(FC)	0.7999 ± 0.0005	0.0508 ± 0.0002	0.1004 ± 0.0001	0.0447 ± 0.0001	0.0242 ± 0.0001
	iTransformer(TC)	0.8016 ± 0.0009	0.0499 ± 0.0002	0.1000 ± 0.0003	0.0441 ± 0.0001	0.0239 ± 0.0001
	Crossformer(FC)	$\mathbf{0.8206 \pm 0.0019}$	$\underline{0.0479 \pm 0.0004}$	$\mathbf{0.0935 \pm 0.0004}$	$\underline{0.0393 \pm 0.0003}$	$\underline{0.0211 \pm 0.0001}$
	Crossformer(TC)	0.8153 ± 0.0022	0.0491 ± 0.0003	$\underline{0.0947 \pm 0.0005}$	0.0401 ± 0.0002	0.0216 ± 0.0001
	TimeMixer(FC)	0.7759 ± 0.0053	0.0588 ± 0.0019	0.1061 ± 0.0014	0.0503 ± 0.0014	0.0271 ± 0.0007
	TimeMixer(TC)	0.7767 ± 0.0029	0.0554 ± 0.0008	0.1057 ± 0.0006	0.0478 ± 0.0005	0.0257 ± 0.0002
	LightTS(FC)	0.7758 ± 0.0079	0.0633 ± 0.0014	0.1049 ± 0.0018	0.0484 ± 0.0009	0.0261 ± 0.0005
	LightTS(TC)	0.7694 ± 0.0077	0.0636 ± 0.0015	0.1059 ± 0.0014	0.0486 ± 0.0009	0.0262 ± 0.0005

Table 16: Full-shot 15-minute 6-hour forecasting: $C_{96} \rightarrow H_{24}$. Entries report mean ± sample SD across three complete seeds. Best results are shown in bold and second-best results are underlined.

Setting	Model	R^2 ↑	MAE↓	RMSE↓	CRPS↓	AQL↓
Full-shot	Cross-Unet	0.7555 ± 0.0033	0.0618 ± 0.0006	0.1131 ± 0.0007	0.0493 ± 0.0004	0.0265 ± 0.0002
	FusionSF	0.7361 ± 0.0070	0.0566 ± 0.0007	0.1152 ± 0.0015	$\mathbf{0.0423 \pm 0.0006}$	$\mathbf{0.0228 \pm 0.0003}$
	TiDE	0.7340 ± 0.0098	0.0657 ± 0.0019	0.1201 ± 0.0023	0.0575 ± 0.0016	0.0312 ± 0.0009
	TimeXer	0.7761 ± 0.0007	0.0617 ± 0.0002	0.1103 ± 0.0002	0.0485 ± 0.0001	0.0264 ± 0.0001
	DAG	0.7795 ± 0.0063	0.0624 ± 0.0013	0.1092 ± 0.0015	0.0492 ± 0.0009	0.0267 ± 0.0005
	GCGNet	0.7318 ± 0.0315	0.0717 ± 0.0072	0.1204 ± 0.0068	0.0602 ± 0.0046	0.0327 ± 0.0025
	PatchTST(FC)	0.7643 ± 0.0004	0.0631 ± 0.0001	0.1129 ± 0.0001	0.0462 ± 0.0001	0.0252 ± 0.0001
	PatchTST(TC)	0.7607 ± 0.0007	0.0619 ± 0.0002	0.1138 ± 0.0002	0.0454 ± 0.0001	0.0247 ± 0.0001
	iTransformer(FC)	0.8036 ± 0.0001	$\underline{0.0516 \pm 0.0001}$	0.1029 ± 0.0000	0.0445 ± 0.0001	0.0240 ± 0.0000
	iTransformer(TC)	$\underline{0.8062 \pm 0.0006}$	$\mathbf{0.0507 \pm 0.0001}$	0.1022 ± 0.0002	0.0440 ± 0.0001	0.0238 ± 0.0000
	Crossformer(FC)	$\mathbf{0.8068 \pm 0.0014}$	0.0522 ± 0.0002	$\mathbf{0.1009 \pm 0.0003}$	$\underline{0.0430 \pm 0.0001}$	$\underline{0.0230 \pm 0.0000}$
	Crossformer(TC)	0.8061 ± 0.0019	0.0526 ± 0.0003	$\underline{0.1012 \pm 0.0004}$	$\underline{0.0430 \pm 0.0004}$	0.0231 ± 0.0002
	TimeMixer(FC)	0.7804 ± 0.0011	0.0585 ± 0.0004	0.1089 ± 0.0003	0.0488 ± 0.0003	0.0263 ± 0.0001
	TimeMixer(TC)	0.7871 ± 0.0024	0.0562 ± 0.0006	0.1071 ± 0.0006	0.0466 ± 0.0007	0.0250 ± 0.0004
	LightTS(FC)	0.7755 ± 0.0041	0.0643 ± 0.0004	0.1095 ± 0.0010	0.0493 ± 0.0004	0.0265 ± 0.0002
	LightTS(TC)	0.7796 ± 0.0008	0.0627 ± 0.0004	0.1085 ± 0.0001	0.0484 ± 0.0003	0.0260 ± 0.0001