Energy and AI 26 (2026) 100887

Contents lists available at ScienceDirect

Energy and AI

journal homepage: www.elsevier.com/locate/egyai

Improving day-ahead probabilistic solar irradiance forecasting via shape-prior flow matching for numerical weather prediction downscaling

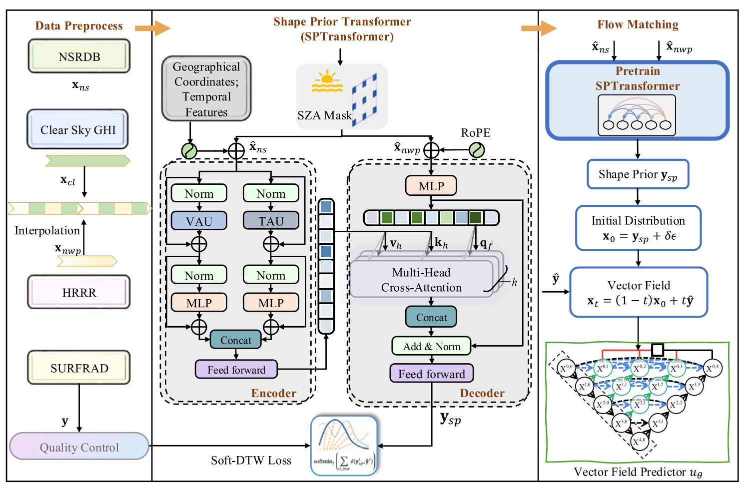
Tao Jing [a], Shanlin Chen [b], Jinqing Peng [c,*], Mengying Li [a,*]

[a] Department of Mechanical Engineering and Research Centre for Data Sciences & Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region
[b] Wind Engineering and Renewable Energy Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne 1015, Switzerland
[c] College of Civil Engineering, Hunan University, Changsha, 410082, Hunan, China

HIGHLIGHTS

- Shape-prior learning converts hourly weather forecasts into 15-min irradiance.
- Soft-Dynamic Time Warping preserves sub-hourly ramps under timing errors.
- Flow matching generates reliable and traceable probabilistic trajectories.
- The method lowers the continuous ranked probability score by up to 10.9%.
- Zero-shot tests show promising transferability at two unseen stations.

GRAPHICAL ABSTRACT



ARTICLE INFO

Dataset link: https://gml.noaa.gov/aftp/data/radiation/surfrad/, https://nsrdb.nrel.gov/data-viewer, https://registry.opendata.aws/noaa-hrrr-pds/

Keywords:
Day-ahead solar forecasting
Numerical weather prediction
Probabilistic temporal downscaling
Conditional flow matching

ABSTRACT

Accurate day-ahead solar irradiance forecasting is essential for photovoltaic integration, reserve scheduling, and energy storage operation. However, operational Numerical Weather Prediction (NWP) products are commonly available at hourly resolution, which limits their ability to represent sub-hourly irradiance ramps and cloud-induced fluctuations relevant to power system operation. To address this limitation, this study proposes Shape-Prior conditional Flow Matching (SPFlow), a probabilistic temporal-downscaling framework that converts hourly NWP forecasts into 15-min irradiance trajectories. SPFlow first employs a deterministic SPTransformer to construct a physically informed high-resolution shape prior from coarse NWP forecasts, satellite-derived meteorological variables, clear-sky irradiance, and site information. Conditioned on this informative prior, a conditional flow-matching model probabilistically refines the deterministic trajectory to represent unresolved sub-hourly variability and forecast uncertainty. Experiments on the SURFRAD dataset show that SPFlow consistently outperforms statistical, deterministic deep-learning, and probabilistic generative baselines. It reduces the continuous ranked probability score by up to 10.9% relative to the strongest probabilistic baseline, improves sub-hourly ramp representation, produces prediction intervals with near-nominal empirical coverage, and demonstrates promising zero-shot transferability to unseen locations. These

* Corresponding authors.
E-mail addresses: jqpeng@hnu.edu.cn (J. Peng), mengying.li@polyu.edu.hk (M. Li).

https://doi.org/10.1016/j.egyai.2026.100887
Received 22 May 2026; Received in revised form 16 August 2026; Accepted 25 August 2026
Available online 4 September 2026

results indicate that combining physically informed shape-prior construction with conditional flow-based probabilistic refinement provides an effective approach for converting coarse NWP forecasts into operationally useful sub-hourly probabilistic solar forecasts.

1. Introduction

Accurate solar irradiance forecasting is regarded as a crucial approach to enhancing the synergistic integration of photovoltaic (PV) systems with the grid and significantly reducing associated integration costs [1,2]. Given the sequential operation of day-ahead, intraday, and real-time electricity markets, forecasts at different lead times must support distinct operational decisions. Particularly as most critical decisions are concentrated in the day-ahead market phase, there is an increased demand for highly accurate day-ahead Global Horizontal Irradiance (GHI) forecasts [3]. Numerical Weather Prediction (NWP) models demonstrate excellent performance in day-ahead forecasting and continue to improve [4,5]. However, the prevalent hourly resolution of NWP products often fails to capture sub-hourly ramp events and high-frequency fluctuations triggered by rapid cloud field evolution [6,7]. This resolution mismatch poses severe risks to future grid stability, a problem exacerbated by the declining share of conventional power plants. Importantly, the value of sub-hourly information at the day-ahead stage lies in anticipatory resource planning before more accurate intraday or real-time updates become available. Resolving the next day's expected intra-hour variability enables operators to identify potentially high-variability periods in advance, pre-position storage and flexible resources, and construct temporally resolved scenarios for day-ahead scheduling and reserve allocation [8].

Beyond operational safety, coarse temporal resolution also significantly impacts techno-economic assessments and system planning. Previous research [8] emphasized that sub-hourly resolution is critical for policy analysis and electricity sector planning [9]. Studies indicate that reliance on hourly data leads to an underestimation of total annualized costs by up to 2% [8] and distorts levelized cost of electricity calculations [10]. Crucially, in renewable energy system design, coarse data may yield unfeasible solutions, particularly by under-provisioning dynamic components such as battery energy storage systems [10] and inverters [8]. To bridge the critical resolution gap between meteorological service provision and the stringent requirements of power system operational planning, temporal downscaling techniques for NWP day-ahead solar irradiance forecasting are emerging as a practical, economical, and efficient solution pathway.

1.1. Literature review

The core challenge of temporal downscaling lies in recovering high-frequency details from low-frequency data. Existing approaches can broadly be categorized into Markov [11], deterministic [12], stochastic [13], non-dimensional, and machine learning-based interpolation [14].

In Markov-based downscaling, the clearness index is first calculated and classified into states representing atmospheric cloudiness. A transition matrix is then constructed based on state probabilities, which is used to weight and fit the downscaled clearness index. Foundational work by Ngoko et al. [15] established the core methodology by removing zenith angle dependence and classifying weather conditions into discrete cloudiness states based on the clearness index. These states were used to construct Markov transition matrices for synthesizing high-resolution solar irradiance. Bright et al. [16] subsequently enhanced this framework by incorporating meteorological drivers such as seasonality, pressure, and wind speed, while also extending the approach to spatial dimensions and allowing for an arbitrary number of matrix states. While effective for generating typical meteorological years, these methods suffer in an NWP forecasting context because they rely on static, historical transition probabilities. The methods assume that future volatility will mirror the average historical behavior, often failing to capture the specific, evolving dynamic weather patterns predicted by the NWP for a specific future window.

Deterministic and stochastic downscaling approaches attempt to model the trajectory between coarse data points mathematically. Linear approaches have been widely adapted for various datasets. Researchers have further refined these methods by incorporating time-step dependencies [17] or by restricting their application to clear-sky conditions [13]. In NWP downscaling, where future high-frequency ground truth is unavailable, these minimum-variance estimators produce over-smoothed trajectories that strip away the critical intermittency required for grid stability analysis. To restore this variance, stochastic methods superimpose noise based on Dirichlet distributions [18] or log-additive models [19]. However, these probabilistic additions are often generic and stationary, and they fail to ensure that the synthesized fluctuations are consistent with the specific atmospheric conditions of the forecast.

The non-dimensionalization downscaling approach normalizes irradiance and time parameters to a 0–1 scale, allowing coarse data to be downscaled by matching it with the closest high-resolution daily values in a database via Euclidean distance. Pioneered by Peruchena et al. [20] and generalized globally by Larraneta et al. [21], this method relies on finding the nearest neighbor day in a historical database. The critical limitation for NWP downscaling is its dependency on the existence of a perfect historical analog. If the NWP predicts a unique or extreme weather event not represented in the library, the method cannot generate an accurate profile, effectively limiting the downscaling to a repetition of past events rather than a prediction of future dynamics.

Machine learning paradigms have increasingly been adopted to model the non-linear stochasticity of solar irradiance [22]. Schreck et al. [23] utilized Artificial Neural Networks (ANNs) to bridge satellite-derived cloud parameters with surface irradiance, while Tang et al. [24] pioneered the use of Generative Adversarial Networks (GANs) to reconstruct high-frequency details directly from low temporal resolution solar irradiance. While machine learning approaches surpass traditional methods in capturing non-linear dependencies, they face inherent limitations in temporal downscaling. Deterministic regression models, typically optimized via Mean Squared Error (MSE), suffer from spectral bias, resulting in over-smoothed trajectories that fail to capture ramp events critical for grid stability. Liu and Zhang [25] proposed a domain-invariant feature learning framework for probabilistic day-ahead renewable power forecasting, enabling knowledge transfer across multiple plants while preserving data privacy. Song et al. [26] integrated attention-enhanced neural networks with natural gradient boosting to improve the reliability and sharpness of probabilistic PV power forecasts. More recently, Kosmopoulos et al. [27] developed a similarity-based day-ahead solar forecasting framework that exploits cloud-transition dynamics to improve geographical generalization without site-specific retraining. Recent conditional generative models, including TimeGrad [28], CSDI [29], and TimeDiff [30], demonstrate that diffusion models can start from Gaussian noise while still being guided by historical observations and exogenous covariates. Conditional flow-matching models similarly condition the vector field on auxiliary information while transporting samples from a simple source distribution [31]. More recent methods, such as TSFlow [32], further introduce temporally structured priors to reduce the mismatch between unstructured noise and time-series data. Our earlier work [33] proposed a prior-guided residual diffusion framework for short-term regional solar irradiance forecasting. Nevertheless, these approaches are not specifically designed for day-ahead temporal downscaling of coarse NWP forecasts. Although conditional information guides the generative process, the source distribution itself commonly lacks an explicit representation of the expected high-resolution irradiance trajectory. Consequently, the generative model must recover substantial temporal structure while also modeling unresolved stochastic variability. For NWP temporal downscaling, a more informative formulation is to center the source distribution on a weather-dependent trajectory that preserves the low-frequency solar profile and clear-sky envelope already contained in the NWP inputs, allowing the probabilistic model to focus on unresolved sub-hourly corrections.

1.2. Contributions

To address the above limitations, this study develops Shape-Prior conditional Flow Matching (SPFlow), a day-ahead probabilistic solar irradiance forecasting framework that converts coarse NWP forecasts into probabilistic sub-hourly irradiance trajectories. The central idea is to reformulate NWP-based solar forecasting as a trajectory refinement problem: a deterministic model first extracts the physically plausible irradiance shape from NWP and satellite-derived information, and a conditional flow matching model then learns the stochastic residual correction towards the high-resolution observation distribution.

The main contributions are summarized as follows:

- A day-ahead probabilistic solar irradiance forecasting framework is proposed for improving coarse NWP forecasts at sub-hourly resolution. By treating temporal downscaling as trajectory refinement rather than simple interpolation, SPFlow directly addresses the resolution mismatch between operational NWP products and renewable energy operation requirements.
- A shape-prior Transformer is developed to construct high-fidelity deterministic irradiance trajectories from multi-source inputs. The proposed Variable Attention Unit (VAU) and Temporal Attention Unit (TAU) disentangle cross-variable meteorological interactions and multi-scale temporal dependencies, enabling the model to extract informative forecasting priors from satellite-derived data and coarse NWP forecasts.
- A Soft-Dynamic Time Warping (Soft-DTW) objective is introduced to train the shape-prior backbone under temporal phase errors. By relaxing rigid pointwise alignment, the proposed training strategy mitigates the double-penalty effect, reduces spectral over-smoothing, and improves the preservation of sub-hourly ramp structures in day-ahead solar forecasts.
- An informative-prior conditional flow matching model is designed to generate probabilistic forecasts with a favorable reliability–sharpness trade-off. Anchoring the source distribution around the learned shape prior allows the flow model to focus on stochastic residual correction, improving empirical reliability at the evaluated prediction-interval level, sampling efficiency, and traceability of the forecast refinement trajectory.

The remainder of this work is structured as follows. Section 2 describes the data used and the pre-processing. Section 3 describes the proposed SPFlow framework, elaborating on the construction of the shape prior and the conditional flow matching mechanism. Section 4 outlines the experimental procedures, specifying the training and inference protocols, benchmarks, and evaluation metrics. Section 5 presents the results and discussion. Finally, the key findings of this study are summarized in Section 6.

2. Data description and pre-processing

2.1. SURFRAD data: In-situ measurements

High-fidelity radiometric measurements acquired from the Surface Radiation Budget Network (SURFRAD) [35] are utilized as the primary ground truth for this study. These datasets serve a dual purpose: establishing the supervisory targets during the model training phase and functioning as the true values for quantitative performance evaluation. Renowned for its long-term reliability, the SURFRAD provides critical in-situ observations that support a wide array of applications in climatology, meteorology, and renewable energy research [1]. Geographically dispersed across the continental United States, the network comprises seven distinct observation sites: Bondville (BON), Desert Rock (DRA), Fort Peck (FPK), Goodwin Creek (GWN), Penn State University (PSU), Sioux Falls (SXF), and Table Mountain (TBL). The spatial distribution covers five diverse climatological zones, ensuring a representative sampling of varying atmospheric conditions, as shown in Fig. 1. To evaluate the models' robustness and generalizability, the SURFRAD datasets are utilized for training, with the exception of the DRA and TBL stations, which are reserved exclusively for zero-shot testing. DRA and TBL represent two contrasting zero-shot environments, with differences in climate type, irradiance variability, seasonal solar conditions, and cloud regimes.

2.2. NSRDB data: Satellite-derived data

The National Solar Radiation Database (NSRDB), developed and maintained by the National Renewable Energy Laboratory (NREL), represents a state-of-the-art repository of high-resolution solar radiation and meteorological data [36]. By employing a physical radiative transfer model that integrates satellite-derived cloud properties with auxiliary atmospheric datasets, the NSRDB generates spatially continuous and temporally complete solar irradiance data across regions monitored by geostationary satellites. For the United States, the archive provides data at a 30-min, 4-km resolution spanning 1998 to 2024, with enhanced 5-min, 2-km capabilities introduced in 2018 following the deployment of GOES-16 and GOES-17 [37]. Furthermore, the database extends its coverage to Europe, Africa, and Asia via the Meteosat and Himawari series, establishing itself as a critical resource for global solar energy modeling. While satellite-derived products inherently possess higher uncertainty compared to in-situ measurements [1], the NSRDB serves as an indispensable proxy in scenarios where long-term, high-quality ground observations are unavailable [38]. Consequently, these gridded datasets are frequently utilized for bias correction in NWP outputs and the development of robust solar forecasting models [39].

In this study, we retrieved data from the NSRDB corresponding to the locations of SURFRAD stations. The dataset covers the period from 2019 to 2022 with a 15-min temporal resolution and includes GHI, temperature, relative humidity, wind direction, and wind speed, alongside the clear-sky irradiance data derived from the REST2 model [36].

2.3. HRRR data: Numerical weather prediction product

The High-Resolution Rapid Refresh (HRRR) is a convection-allowing, hourly updated NWP model operated by the National Oceanic and Atmospheric Administration (NOAA) [40]. Built upon the Advanced Research Weather Research and Forecasting (WRF-ARW) core, the HRRR features a 3-km spatial resolution and is initialized via boundary conditions and data assimilation from the 13-km Rapid Refresh (RAP) system. Since its operational inception in 2014, the framework has undergone significant refinement; notably, the 2020 upgrade to HRRRv4 introduced enhanced physical parameterizations and extended forecast capabilities compared to earlier iterations. Operationally, the HRRR generates 18-hour-ahead forecasts every hour, while the prediction horizon is extended to 48 h during the synoptic cycle runs (i.e., 00Z, 06Z, 12Z, and 18Z). To align with the operational requirements of day-ahead solar power trading and scheduling, this study exclusively utilizes the forecast data initiated at 00Z. Specifically, the GHI predictions corresponding to lead times of 1 to 24 h are extracted to represent the day-ahead solar resource profile [41]. In this study, day-ahead forecasting refers to fixed-origin downscaling of 00Z HRRR forecasts at lead times of 1–24 h from hourly to 15-min resolution, providing

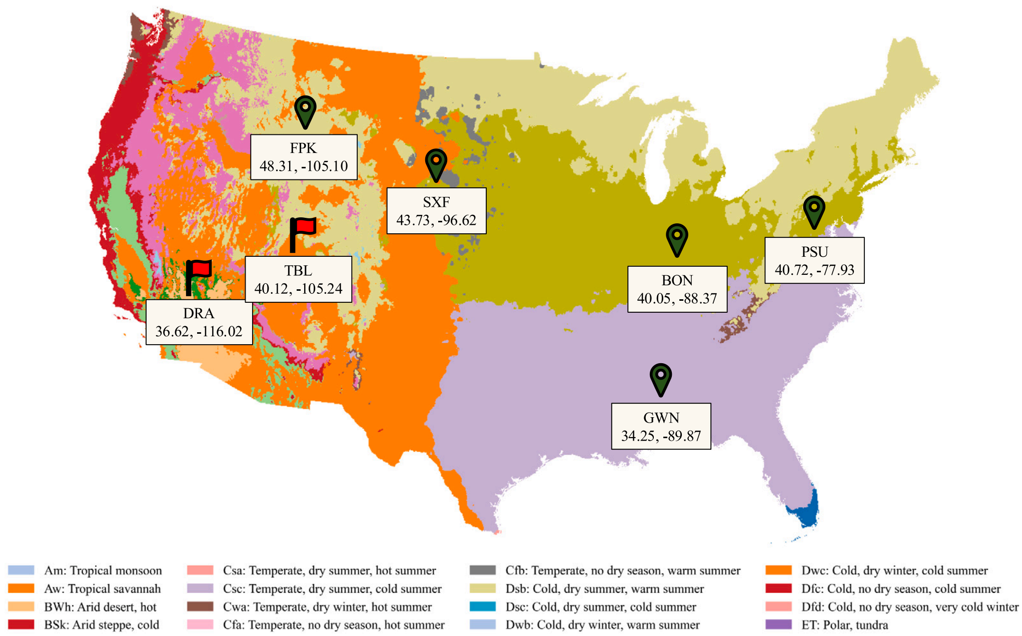


Fig. 1. Köppen-Geiger weather classification of the United States for SURFRAD data [34]. Data from stations marked with droplet symbols are used for training and testing, while data from stations marked with flag symbols are used for zero-shot testing.

sub-hourly probabilistic trajectories for ramp-aware scheduling and energy-storage planning without using post-initialization observations or forecast updates. The operational value of this resolution at the day-ahead origin is to reveal potentially high-variability periods before shorter-horizon forecasts become available. Such information can support advance allocation of ramping and reserve capability, prepositioning of battery state of charge and other flexible resources, and the construction of 15-min scenarios for day-ahead stochastic scheduling and optimization.

2.4. Data preprocessing and quality control

To ensure robustness, rigorous quality control (QC) and preprocessing are applied to the raw 1-min SURFRAD measurements. Following the methodology outlined by Yang et al. [42], we implemented a two-stage filtering process comprising physically based limit checks and a closure equation test to identify and exclude erroneous data points. The first stage enforces extremely rare limits on GHI, Diffuse Horizontal Irradiance (DHI), and Direct Normal Irradiance (DNI) based on the solar zenith angle and the extraterrestrial irradiance:

$$\begin{bmatrix} -2 < \mathrm{GHI} < 1.2 E_{0n} \cos^{1.2}(Z) + 50 \\ -2 < \mathrm{DHI} < 0.75 E_{0n} \cos^{1.2}(Z) + 30 \\ -2 < \mathrm{DNI} < 0.95 E_{0n} \cos^{0.2}(Z) + 10 \end{bmatrix}, \tag{1}$$

where Z denotes the solar zenith angle, and E_{0n} denotes the extraterrestrial irradiance incident on a surface normal to the solar rays. Subsequently, a three-component closure test is applied to verify the consistency between the measured GHI and the sum of its components. Defining the closure error as: $|closr| = |\mathrm{GHI} - (\mathrm{DNI}\cos Z + \mathrm{DHI})|$, the data must satisfy:

$$\begin{bmatrix} |closr| < 8\% & \text{for } Z < 75^\circ \text{ and GHI} > 50 \\ |closr| < 15\% & \text{for } 75^\circ < Z < 93^\circ \text{ and GHI} > 50 \end{bmatrix}. \tag{2}$$

Post-QC, the valid high-frequency data are aggregated into 15-min intervals using the center indexing scheme [43], and aligned with Coordinated Universal Time (UTC) to ensure temporal consistency with the NSRDB inputs.

To ensure temporal alignment across heterogeneous data sources, we upsample the hourly HRRR forecast values to a 15-min resolution. Direct interpolation of NWP GHI may fail to account for the solar diurnal cycle and can therefore introduce interpolation errors. Therefore, we adopt a clear-sky index-based interpolation scheme, which uses clear-sky GHI values derived from the REST2 model [13]. The clear-sky index k_c is obtained by normalizing the NWP GHI forecasts against the corresponding clear-sky GHI:

$$k_c(t_h) = \frac{\mathrm{GHI}_{\mathrm{NWP}}(t_h)}{\mathrm{GHI}_{\mathrm{cl}}(t_h)}, \tag{3}$$

where t_h represents the hourly timestamps. The resulting hourly k_c is then linearly interpolated to the target 15-min intervals. The 15-min NWP GHI is subsequently reconstructed by multiplying the interpolated clear-sky index by the 15-min REST2 clear-sky GHI. Two boundary cases require special treatment in the clear-sky index interpolation. First, when the clear-sky GHI approaches zero near sunrise and sunset, the division becomes numerically unstable. In such intervals, we set $k_c = 0$ and directly assign $\mathrm{GHI}_{\mathrm{NWP}} = 0$ W/m^2, which is consistent with the negligible irradiance at high zenith angles. Second, cloud-enhancement events can cause the NWP forecast to exceed the clear-sky reference, yielding $k_c > 1$. We apply a post-interpolation clipping at $1.2 \times \mathrm{GHI}_{\mathrm{cl}}$ following the extremely rare limit defined in Eq. (1) to prevent unphysical outliers. Finally, all input variables are standardized via Z-score normalization to ensure stable model convergence:

$$x' = \frac{x - \mu_{\mathrm{train}}}{\sigma_{\mathrm{train}}}, \tag{4}$$

where x is the original feature vector, μ_{train} and σ_{train} denote the mean and standard deviation calculated exclusively from the training dataset. For zero-shot stations (DRA and TBL), the same training-set statistics are applied without recalculation, ensuring that no target-site information leaks into the normalization.

3. Forecasting methodology

The primary objective of SPFlow is to generate high-resolution, high-quality GHI forecasts by downscaling coarse-grained NWP outputs with multi-source spatiotemporal data. The input feature space comprises NSRDB data $\mathbf{x}_{ns} \in \mathbb{R}^{L_{in} \times C}$, clear-sky GHI $\mathbf{x}_{cl} \in \mathbb{R}^{L_{in} \times 1}$, NWP GHI forecasts $\mathbf{x}_{nwp} \in \mathbb{R}^{L_{out} \times 1}$, and auxiliary features $\mathbf{c}$ (latitude, longitude, altitude). The target is in-situ SURFRAD measurements $\mathbf{y} \in \mathbb{R}^{L_{out} \times 1}$. The

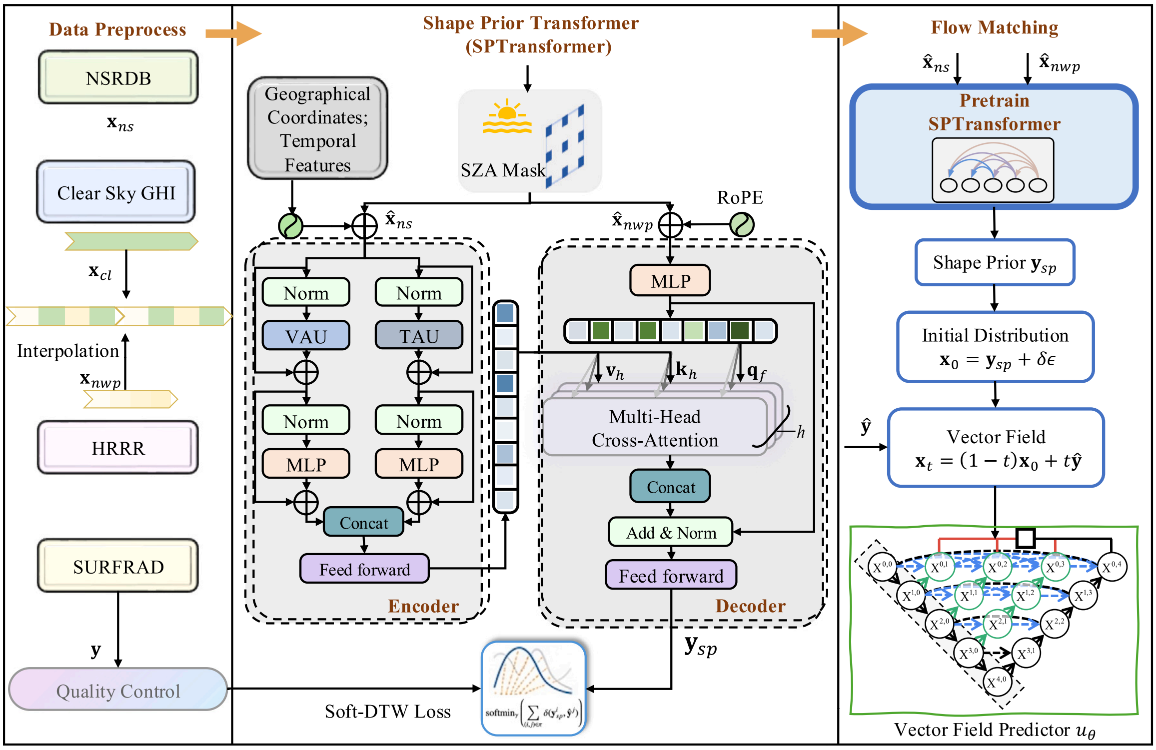


Fig. 2. The proposed SPFlow framework.

goal is to learn $p_\theta(\mathbf{y} \mid \mathbf{x}_{ns}, \mathbf{x}_{nwp}, \mathbf{c})$ that approximates the conditional distribution of $\mathbf{y}$ given the inputs.

As illustrated in Fig. 2, SPFlow adopts a two-stage pipeline to bridge the resolution gap between day-ahead NWP forecasts and the high-frequency demands of grid dispatch. In the first stage, a baseline is constructed by temporally interpolating $\mathbf{x}_{nwp}$ via $\mathbf{x}_{cl}$. A deterministic SPTransformer then corrects this baseline, the novel TAU and VAU disentangle multi-scale dependencies by leveraging satellite-derived contexts from $\mathbf{x}_{ns}$, while a Soft-DTW loss prioritizes morphological fidelity over rigid point-to-point alignment to mitigate phase-shift-induced over-smoothing. In the second stage, the shape-aware deterministic output serves as the informative prior for a flow matching network that models $p_\theta(\mathbf{y} \mid \mathbf{x}_{ns}, \mathbf{x}_{nwp}, \mathbf{c})$ through stable ODE dynamics, efficiently yielding fine-grained probabilistic forecasts whose continuous generation trajectories remain traceable.

3.1. SPTransformer

The structure of SPTransformer is presented in the middle part of Fig. 2. In the task of preliminarily correcting clear sky irradiance interpolated NWP forecasts, the target variable is governed by complex, coupled dependencies. Specifically, the irradiance dynamics are not only auto-correlated with their own historical values but are also significantly influenced by the cross-correlations among various satellite-retrieved meteorological variables within adjacent time steps. To address this, the SPTransformer deploys a modified Transformer [44] designed to explicitly disentangle these two distinct types of dependencies.

3.1.1. Encoder

GHI data inherently contain a deterministic diurnal pattern, with values being zero during the night. Therefore, including nighttime data can introduce spurious correlations and gradient noise, as other meteorological data will be non-zero during nighttime. To mitigate this, a binary masking mechanism based on solar zenith angle (SZA) is employed to handle nighttime values [45]. Specifically, a temporal mask $\mathbf{M} = [M_1, M_2, \ldots, M_t]$ is applied to the input sequence, defined as:

$$M_t = \begin{cases} 1, & \text{if } Z_t \le 85^\circ \\ 0, & \text{if } Z_t > 85^\circ, \end{cases} \tag{5}$$

where Z_t is the SZA at time step t. The input meteorological variables $\mathbf{x}_{ns}$, NWP $\mathbf{x}_{nwp}$, and target $\mathbf{y}$ are filtered element-wise to obtain the valid input:

$$\{\hat{\mathbf{x}}_{ns}, \hat{\mathbf{x}}_{nwp}, \hat{\mathbf{y}}\} = \{\mathbf{x}_{ns}, \mathbf{x}_{nwp}, \mathbf{y}\} \odot \mathbf{M} \tag{6}$$

where $\odot$ denotes the Hadamard product with broadcasting along the feature dimension. This ensures that the subsequent attention mechanisms only attend to effective daylight intervals. Then, the Rotary Positional Embeddings (RoPE) [46] is utilized to embed geographical coordinates and temporal features. RoPE encodes relative positions by rotating the query and key vectors in the complex vector space, which is crucial for capturing the temporal features.

This encoder further integrates two core modules: the VAU, dedicated to capturing time-variant inter-variable correlations, and the TAU, specialized in modeling intra-variable dependencies across temporal sequences, as shown in Fig. 3. The operations of the two branches can be summarized as follows:

$$\mathbf{z}_v = \text{MLP}(\text{LN}(\text{VAU}(\text{LN}(\hat{\mathbf{x}}_{ns})))), \tag{7}$$

$$\mathbf{z}_t = \text{MLP}(\text{LN}(\text{TAU}(\text{LN}(\hat{\mathbf{x}}_{ns})))), \tag{8}$$

where LN denotes the layer normalization operation. The outputs of the two streams are fused via a gated mechanism to form a comprehensive representation:

$$\mathbf{z}_{enc} = \text{FFN}\left(\text{Concat}(\mathbf{z}_t, \mathbf{z}_v)\right) + \hat{\mathbf{x}}_{ns} \tag{9}$$

This fused representation $\mathbf{z}_{enc}$ encapsulates both the physical state variables and their temporal dynamics, serving as a robust context for the subsequent decoder.

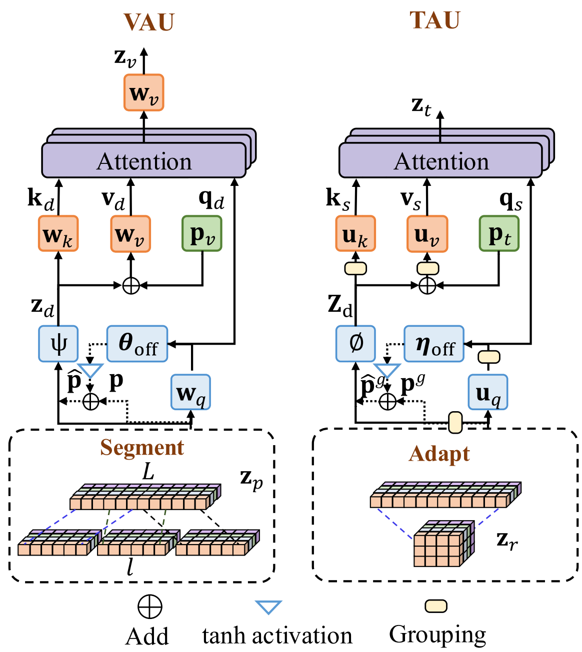


Fig. 3. The structures of TAU and VAU.

A. Variable attention unit. Unlike standard self-attention, which computes global dependencies across fixed positions, VAU incorporates a deformable mechanism [47,48]. This allows the model to adaptively sample relevant features from the spatiotemporal manifold, focusing on informative regions across different meteorological variables and time steps.

Given the input embedding sequence $\mathbf{z}_{in} = \mathrm{LN}(\hat{\mathbf{x}}_{ns}) \in \mathbb{R}^{L\times d}$, computing attention over the entire sequence length L is computationally expensive and may introduce noise from distant time steps. We first segment $\mathbf{z}_{in}$ into overlapping patches along the temporal dimension. Let ℓ be the segment length and s be the stride. The sequence is unfolded into N patches, denoted as $\mathbf{z}_p \in \mathbb{R}^{N\times\ell\times d}$, where $N = \lfloor (L-\ell)/s \rfloor + 1$. This segmentation imposes a locality inductive bias, encouraging the model to capture dependencies within proximal time steps.

The core innovation of VAU lies in its ability to attend to flexible positions. For each patch $\mathbf{z}_p$, we treat the data points as a grid indexed by coordinate $\mathbf{p}$. We aim to learn a position offset $\hat{\mathbf{p}} \in \mathbb{R}^2$ to adjust the sampling locations for Keys and Values, thereby aligning the receptive field with the dynamic movement of meteorological features. First, we project the patch to obtain the query embedding $\mathbf{q}_p = \mathbf{z}_p \mathbf{w}_q$. The offsets are then predicted via a lightweight sub-network θ_{off}, which consists of a $k \times k$ convolutional layer to capture local context and a 1×1 convolutional layer for projection:

$$\hat{\mathbf{p}} = \alpha \cdot \tanh\left(\theta_{\mathrm{off}}(\mathbf{q}_p)\right) \tag{10}$$

where α is a learnable scalar controlling the magnitude of the offset, and tanh denotes the activation function that constrains the range to stabilize training. Since the calculated position $\mathbf{p} + \hat{\mathbf{p}}$ is typically fractional, we employ a differentiable bilinear interpolation kernel $\psi(\cdot)$ to sample the deformed features $\mathbf{z}_d$ from the input patch [49]:

$$\mathbf{z}_d = \psi(\mathbf{z}_p; \mathbf{p} + \hat{\mathbf{p}}) \tag{11}$$

This operation effectively warps the feature map, allowing the attention mechanism to focus on relevant locations that may not align with the rigid grid.

With the deformed features, we compute the deformable keys $\mathbf{k}_d$ and values $\mathbf{v}_d$. To preserve relative positional information that might be distorted during deformation, a relative positional bias $\mathbf{p}_v$ is explicitly added to the value projection [49]:

$$\mathbf{k}_d = \mathbf{z}_d \mathbf{w}_k, \tag{12}$$

$$\mathbf{v}_d = \mathbf{z}_d \mathbf{w}_v + \mathbf{p}_v. \tag{13}$$

The attention score is then computed between the original query $\mathbf{q}_p$ and the deformed key $\mathbf{k}_d$, aggregating information from the deformed values $\mathbf{v}_d$:

$$\mathbf{Atten}_i^v = \mathrm{Softmax}\left(\frac{\mathbf{q}_p \mathbf{k}_d^\top}{\sqrt{d}}\right)\mathbf{v}_d \tag{14}$$

where $\mathbf{Atten}_i^v$ represents the attention output for the ith patch. Finally, the outputs from all patches are concatenated and projected via a linear layer $\mathbf{w}_v$ to reconstruct the sequence $\mathbf{z}_v \in \mathbb{R}^{L\times d}$. This process enables VAU to dynamically capture interactions between neighboring variables across time, making it robust to temporal shifts and warping.

B. Temporal attention unit (TAU). While VAU focuses on inter-variable dependencies, TAU is designed to capture intra-variable dependencies across different temporal granularities. Meteorological data often exhibit strong periodicity and temporal shifts. To address this, TAU employs a grouped deformable mechanism that learns adaptive temporal offsets for different feature subspaces. To capture dependencies across specific time windows, we first pass $\mathbf{z}_{in}$ through an *Adapt* module. This module reshapes the sequence into a structured tensor $\mathbf{z}_r \in \mathbb{R}^{r\times\kappa\times d}$, where r denotes the number of temporal segments and κ represents the segment length. This transformation allows the module to explicitly model relationships between non-adjacent time steps that share semantic similarity.

Unlike the spatial deformation in VAU where each point has a unique offset, temporal dependencies are often consistent across subsets of latent channels. To reduce computational redundancy and impose structural regularization, we employ a channel grouping strategy. We project the input to obtain a query $\mathbf{q}_s = \mathbf{z}_{in}\mathbf{u}_q$, where $\mathbf{u}_q \in \mathbb{R}^{d\times d}$. The query is then split along the feature dimension into G subgroups, denoted as $\mathbf{q}_g \in \mathbb{R}^{r\times\kappa\times(d/G)}$ for $g = \{1,\ldots,G\}$.

For each subgroup, we employ a lightweight 1D Convolutional Neural Network (1DCNN), denoted as η_{off}, to predict a shared temporal offset $\Delta p^{(g)}$. This network comprises a $k \times 1$ convolution to aggregate local temporal context, followed by a 1×1 convolution for projection. The offset is formulated as:

$$\hat{\mathbf{p}}^g = \alpha \cdot \tanh\left(\eta_{\mathrm{off}}(\mathbf{q}_g)\right), \tag{15}$$

where, $\hat{\mathbf{p}}^g \in \mathbb{R}$ represents a scalar shift along the temporal axis for the gth group. Since time is continuous in this latent representation, we use a linear interpolation function $\phi(\cdot)$ to sample the deformed features $\mathbf{z}_d$ from the discrete grid indices $\mathbf{p}^g$:

$$\mathbf{z}_d^g = \phi\left(\mathbf{z}_r^{(g)}; \mathbf{p}^g + \hat{\mathbf{p}}^g\right), \tag{16}$$

where $\mathbf{z}_d$ is the concatenation of all deformed subgroups. The deformed features $\mathbf{z}_d$ are then projected to generate keys and values using weight matrices $\mathbf{u}_k$ and $\mathbf{u}_v$. To preserve the temporal order information that might be disturbed by the deformation, a learnable relative temporal bias $\mathbf{p}_t \in \mathbb{R}^{\kappa\times d}$ is injected into the value representation:

$$\mathbf{k}_s = \mathbf{z}_d \mathbf{u}_K, \tag{17}$$

$$\mathbf{v}_s = \mathbf{z}_d \mathbf{u}_v + \mathbf{p}_t. \tag{18}$$

The multi-head attention is performed over the temporal dimension. The attention mechanism aggregates information from the adaptively sampled time steps:

$$\mathbf{Atten}_t = \mathrm{Softmax}\left(\frac{\mathbf{q}_s \mathbf{k}_s^\top}{\sqrt{d}}\right)\mathbf{v}_s. \tag{19}$$

The output $\mathbf{Atten}_t$ captures the evolution of variables by attending to the most relevant temporal contexts defined by the learned offsets. Finally, the output is reshaped back to the original sequence length L to produce the final representation $\mathbf{z}_t$.

3.1.2. Decoder

The decoder module is designed to synthesize the final prediction $\mathbf{y}_{sp}$ by integrating the learned latent representations with external covariates. Unlike the encoder, which focuses on feature extraction, the decoder employs a cross-attention mechanism to query the encoded historical context based on future-known information (NWP predicted values). To account for temporal or positional variations, a learnable embedding is added to the $\hat{\mathbf{x}}_{nwp}$:

$$\mathbf{z}_{\mathrm{dec}} = \mathrm{MLP}(\hat{\mathbf{x}}_{nwp} + \mathbf{P}_{\mathrm{pos}}(\hat{\mathbf{x}}_{nwp})), \tag{20}$$

where $\mathbf{P}_{\mathrm{pos}}$ represents the positional encoding, and the MLP projects the input into the model's latent dimension. This results in a query sequence that defines the target states we wish to predict.

The core of the decoder is the Multi-Head Cross-Attention (MHCA) [50] layer. This layer aligns the decoder's query states with the historical information encoded by the encoder. For each attention head i, the attention scores are calculated as follows:

$$\mathbf{CAtten}_i = \mathrm{Softmax}\left(\frac{\mathbf{z}_{\mathrm{dec}}\mathbf{w}_q^{(i)} \times \mathbf{z}_{\mathrm{enc}}\mathbf{w}_k^{(i)\top}}{\sqrt{d_k}}\right)\mathbf{z}_{\mathrm{enc}}\mathbf{w}_v^{(i)}. \tag{21}$$

where $\mathbf{W}_Q^{(i)}$, $\mathbf{W}_K^{(i)}$, and $\mathbf{W}_V^{(i)}$ are learnable projection matrices. This mechanism allows the decoder to selectively retrieve relevant historical features ($\mathbf{v}_h$) that correspond to the specific future conditions defined by the covariates ($\mathbf{q}_f$). The outputs from all h heads are concatenated and projected linearly to restore the original dimensions. Following the standard Transformer design, we apply a residual connection followed by Layer Normalization:

$$\mathbf{z}_{\mathrm{attn}} = \mathrm{LN}(\mathbf{z}_{\mathrm{dec}} + \mathrm{Concat}(\mathbf{CAtten}_1, \dots \mathbf{CAtten}_h)\mathbf{w}_o), \tag{22}$$

Subsequently, the representation passes through a Feed-Forward Network (FFN) with non-linear activation functions to refine the features:

$$\mathbf{y}_{sp} = \mathrm{FFN}(\mathbf{z}_{\mathrm{attn}}), \tag{23}$$

where, $\mathbf{y}_{sp}$ represents the final shape prior.

3.1.3. Soft-DTW loss function

Conventional pointwise losses are fundamentally ill-suited for training the shape prior backbone in the presence of temporal phase errors. Day-ahead NWP forecasts are subject to systematic timing displacements arising from cloud-field advection and convective initiation uncertainties [51]. Under pointwise losses, a predicted ramp event that is temporally shifted by even a single time step relative to the observation incurs a double penalty, as the loss simultaneously penalizes the false alarm at the predicted location and the miss at the observed location, even though the two sequences are morphologically identical. This effect steers the optimizer towards a smooth conditional mean that suppresses the operationally critical high-frequency variability. The problem is further compounded by the spectral bias inherent in neural networks [52]. Under ℓ_2-based objectives, the model converges preferentially to the low-frequency diurnal envelope while systematically attenuating the sub-hourly fluctuations residing in higher spectral bands.

To jointly mitigate the double-penalty effect and spectral bias, we employ the Soft-DTW loss [53] to optimize the SPTransformer. Soft-DTW replaces rigid pointwise comparison with a differentiable elastic alignment that evaluates shape similarity irrespective of local temporal distortions. Let $\Delta(\mathbf{y}_{sp}, \hat{\mathbf{y}}) \in \mathbb{R}^{n\times m}$ denote the pointwise cost matrix with entries $\delta(y_{sp}^i, \hat{y}^j) = (y_{sp}^i - \hat{y}^j)^2$, and let $\mathcal{A}_{n,m} \subset \{0,1\}^{n\times m}$ be the set of binary alignment matrices encoding all admissible monotonic warping paths between the two sequences. The Soft-DTW loss is defined as a soft minimum over the cost of every admissible alignment [53]:

$$\mathcal{L}^{\gamma}_{\text{soft-DTW}}(\mathbf{y}_{sp}, \hat{\mathbf{y}}) = -\gamma \log \sum_{A\in\mathcal{A}_{n,m}} \exp\left(-\frac{\langle A, \Delta(\mathbf{y}_{sp}, \hat{\mathbf{y}})\rangle}{\gamma}\right), \qquad \gamma > 0, \tag{24}$$

where $\langle\cdot,\cdot\rangle$ denotes the Frobenius inner product and γ controls the smoothness: as $\gamma \to 0^+$, Eq. (24) recovers the classical (non-differentiable) DTW distance, while larger γ yields a smoother surrogate amenable to gradient-based optimization. In practice, Eq. (24) is evaluated by the dynamic-programming recursion

$$r_{i,j} = \delta(y_{sp}^i, \hat{y}^j) + \mathrm{softmin}_{\gamma}\left(r_{i-1,j-1}, r_{i-1,j}, r_{i,j-1}\right), \tag{25}$$

with $\mathcal{L}^{\gamma}_{\text{soft-DTW}} = r_{n,m}$ and

$$\mathrm{softmin}_{\gamma}(a_1, \dots, a_k) = -\gamma \log \sum_{i=1}^{k} \exp(-a_i/\gamma). \tag{26}$$

In all experiments we set $\gamma = 0.1$ on the standardized irradiance scale. By searching over all admissible warping paths and weighting them through the soft minimum, the loss reduces the excessive double penalty assigned to moderately phase-shifted but morphologically consistent trajectories. Consequently, the backbone is free to generate sharp ramp features, preserving the high-frequency spectral content that would otherwise be suppressed under MSE optimization. The resulting shape prior $\mathbf{y}_{sp}$ thus maintains the overall physical trend of the clear-sky-interpolated baseline while correcting for amplitude errors and temporal delays, providing a high-fidelity deterministic anchor for the subsequent probabilistic flow matching stage.

3.2. Shape priors flow matching

The Conditional Flow Matching (CFM) framework [31] is adopted to jointly refine coarse NWP outputs into high-resolution irradiance trajectories and quantify the aleatoric uncertainty induced by sub-grid cloud dynamics. CFM provides a simulation-free training objective for continuous normalizing flows by learning a time-dependent vector field that transports a base distribution p_0 to the target distribution p_1 of high-resolution GHI. Unlike diffusion-based downscalers, CFM admits deterministic ODE sampling under arbitrary p_0, which is desirable for operational forecasting where inference latency and reproducibility are critical.

The choice of p_0 is decisive for any flow-based downscaler. Vanilla flow matching initializes from $p_0 = \mathcal{N}(\mathbf{0}, \mathbf{I})$, which discards the diurnal cycle, clear-sky envelope, and weather-dependent temporal structure already captured by SPTransformer. The resulting large discrepancy between uninformative Gaussian noise and the GHI trajectory manifold increases the transport difficulty and may require more numerical integration steps. To address this limitation, SPFlow anchors the source distribution around the deterministic shape prior $\mathbf{y}_{sp}$:

$$\mathbf{x}_0 = \mathbf{y}_{sp} + \delta\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{27}$$

where $p_0(\mathbf{x}_0 \mid \mathbf{h}) = \mathcal{N}(\mathbf{y}_{sp}, \delta^2\mathbf{I})$ and $\mathbf{h} = [\hat{\mathbf{x}}_{ns}; \hat{\mathbf{x}}_{nwp}; \mathbf{c}]$. In the context of GHI temporal downscaling, this source distribution represents plausible initial 15-min irradiance trajectories concentrated around the deterministic forecast produced by SPTransformer. Unlike samples drawn from a standard Gaussian source, these trajectories already retain the expected diurnal profile, clear-sky envelope, and weather-dependent temporal structure inferred from the NWP and satellite-derived inputs. The perturbation $\delta\epsilon$ introduces local variability around the shape prior and provides different starting trajectories for probabilistic generation. The Gaussian perturbation also ensures the absolute continuity of p_0 required by the CFM framework. The parameter δ controls the trade-off between determinism and diversity: $\delta \to 0$ collapses the source distribution towards the deterministic shape prior, whereas an excessively large δ weakens the physical anchoring and approaches the Gaussian-prior baseline SPFlow-G.

The Optimal Transport path [54] is employed to couple $\mathbf{x}_0$ with the target $\hat{\mathbf{y}} \sim p_1(\cdot \mid \mathbf{h})$:

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\hat{\mathbf{y}}, \quad t \in [0,1]. \tag{28}$$

Differentiating Eq. (28) at fixed endpoints yields a constant conditional vector field along each trajectory, $u_t(\mathbf{x}_t \mid \mathbf{x}_0, \hat{\mathbf{y}}) = \hat{\mathbf{y}} - \mathbf{x}_0$. Such straight-line paths minimize transport cost and produce low-curvature flows that are numerically stable, an essential property for resolving rapidly varying cloud-induced irradiance ramps without artifacts. The vector field is parameterized as $u_\theta(\mathbf{x}_t, t, \mathbf{h})$ via the structured state-space backbone of [55], which is well suited to the long-range temporal dependencies inherent in multi-horizon NWP downscaling, and is trained by minimizing

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t,\mathbf{h},\mathbf{x}_0,\hat{\mathbf{y}}}\left[\left\| u_\theta(\mathbf{x}_t, t, \mathbf{h}) - (\hat{\mathbf{y}} - \mathbf{x}_0)\right\|^2\right], \tag{29}$$

with $t \sim \mathcal{U}[0,1]$, $\mathbf{x}_0 \sim p_0(\cdot \mid \mathbf{h})$, $\hat{\mathbf{y}} \sim p_1(\cdot \mid \mathbf{h})$. In this formulation, the target distribution $p_1(\hat{\mathbf{y}} \mid \mathbf{h})$ represents the conditional distribution of observed 15-min GHI trajectories under the same forecasting context $\mathbf{h}$. During training, each corresponding SURFRAD trajectory $\hat{\mathbf{y}}$ provides a sample from this target distribution. The learned vector field $u_\theta(\mathbf{x}_t, t, \mathbf{h})$ can be interpreted as a time-dependent trajectory-correction rule. Given the current provisional trajectory $\mathbf{x}_t$, the transport time t, and the forecasting context $\mathbf{h}$, it determines the direction and magnitude of the correction required to reduce residual errors in irradiance amplitude, ramp timing, and sub-hourly variability.

Although standard CFM theory assumes a fixed p_0, an input-dependent source remains admissible: conditioned on $\mathbf{h}$, $\mathbf{x}_0$ and $\hat{\mathbf{y}}$ are sampled independently, so the marginalization theorem (Theorem 1 in [31]) carries over conditionally—the optimal regressor of Eq. (29) recovers $u_t(\mathbf{x}_t \mid \mathbf{h}) = \mathbb{E}[\hat{\mathbf{y}} - \mathbf{x}_0 \mid \mathbf{x}_t, \mathbf{h}]$ and the induced ODE transports $p_0(\cdot \mid \mathbf{h})$ to $p_1(\cdot \mid \mathbf{h})$. Beyond admissibility, the shape-prior source yields three quantitative advantages. First, the squared 2-Wasserstein distance between source and target satisfies

$$W_2^2(p_0, p_1 \mid \mathbf{h}) \le \left\| \mathbf{y}_{sp} - \mathbb{E}[\hat{\mathbf{y}} \mid \mathbf{h}]\right\|^2 + \mathrm{tr}(\mathrm{Cov}[\hat{\mathbf{y}} \mid \mathbf{h}]) + d\,\delta^2, \tag{30}$$

where d denotes the dimensionality. The deterministic gap thus contracts from the full GHI magnitude $\|\mathbb{E}[\hat{\mathbf{y}} \mid \mathbf{h}]\|^2$ under a Gaussian prior to the SPTransformer's residual error. Second, as the Euler local truncation error scales as $\mathcal{O}(\Delta t^2 \cdot \mathrm{Var}[\hat{\mathbf{y}} - \mathbf{x}_0 \mid \mathbf{x}_t, \mathbf{h}])$, this contraction permits accurate sampling with markedly fewer ODE steps. Third, the training-target variance shrinks from $\mathrm{Var}[\hat{\mathbf{y}} \mid \mathbf{h}] + \mathbf{I}$ to $\mathrm{Var}[\hat{\mathbf{y}} \mid \mathbf{h}] + \delta^2\mathbf{I}$, reducing gradient noise and accelerating convergence [54] on limited NWP–observation paired records. Substituting Eq. (27) into Eq. (29) further shows that u_θ implicitly learns the residual mapping $\hat{\mathbf{y}} - \mathbf{y}_{sp}$ (up to $\delta\epsilon$), aligning SPFlow with the classical deterministic-mean-plus-stochastic-residual paradigm while preserving the full generative capacity of CFM.

During inference, multiple independent perturbations are sampled around the same forecast-specific shape prior, producing source trajectories $\{\mathbf{x}_0^{(m)}\}_{m=1}^{M}$. Each source trajectory represents a plausible initial 15-min GHI realization consistent with the deterministic NWP-based forecast. The samples are subsequently propagated from $t = 0$ to $t = 1$ through the learned ordinary differential equation. Using N forward-Euler steps with $\Delta t = 1/N$, the numerical update is

$$\mathbf{x}_{t+\Delta t} = \mathbf{x}_t + u_\theta(\mathbf{x}_t, t, \mathbf{h})\Delta t, \quad t = 0, \Delta t, \ldots, 1 - \Delta t. \tag{31}$$

The resulting ensemble represents alternative 15-min GHI evolutions that are consistent with the same day-ahead NWP and meteorological conditions. The ensemble mean provides the deterministic downscaled forecast, while the empirical ensemble distribution is used to construct prediction intervals and quantify forecast uncertainty. Algorithms 1 and 2 summarize the corresponding two-stage training and probabilistic inference procedures.

Algorithm 1 Two-stage training of SPFlow.

Input: $\mathcal{D}$, f_ϕ, u_θ, noise scale δ, EMA decay ρ
Output: Trained parameters ϕ^*, $\bar{\theta}^*$
Stage I: Shape-prior learning
while not converged **do**
  $\mathcal{B} = (\mathbf{x}_{ns}, \mathbf{x}_{cl}, \mathbf{x}_{nwp}, \mathbf{c}, \mathbf{y}) \sim \mathcal{D}$
  $(\hat{\mathbf{x}}_{ns}, \hat{\mathbf{x}}_{nwp}, \hat{\mathbf{y}}) \leftarrow \mathrm{Mask}_{85^\circ}(\mathcal{B})$
  $\mathbf{y}_{sp} \leftarrow f_\phi(\hat{\mathbf{x}}_{ns}, \hat{\mathbf{x}}_{nwp}, \mathbf{x}_{cl}, \mathbf{c})$ ▷ *shape prior*
  $\mathcal{L}_{sp} \leftarrow \mathcal{L}^{\gamma}_{\text{soft-DTW}}(\mathbf{y}_{sp}, \hat{\mathbf{y}})$ ▷ *Eq.* (24)
  $\phi \leftarrow \mathrm{AdamW}(\phi, \nabla_\phi \mathcal{L}_{sp})$
end while
$\phi^* \leftarrow \phi$; freeze f_{ϕ^*}; $\bar{\theta} \leftarrow \theta$
Stage II: Conditional flow matching
while not converged **do**
  $\mathcal{B} = (\mathbf{x}_{ns}, \mathbf{x}_{cl}, \mathbf{x}_{nwp}, \mathbf{c}, \mathbf{y}) \sim \mathcal{D}$
  $(\hat{\mathbf{x}}_{ns}, \hat{\mathbf{x}}_{nwp}, \hat{\mathbf{y}}) \leftarrow \mathrm{Mask}_{85^\circ}(\mathcal{B})$
  $\mathbf{y}_{sp} \leftarrow \mathrm{sg}\left[f_{\phi^*}(\hat{\mathbf{x}}_{ns}, \hat{\mathbf{x}}_{nwp}, \mathbf{x}_{cl}, \mathbf{c})\right]$
  $\mathbf{h} \leftarrow [\hat{\mathbf{x}}_{ns}; \hat{\mathbf{x}}_{nwp}; \mathbf{c}]$
  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $t \sim \mathcal{U}(0,1)$
  $\mathbf{x}_0 \leftarrow \mathbf{y}_{sp} + \delta\epsilon$ ▷ *Eq.* (27)
  $\mathbf{x}_t \leftarrow (1-t)\mathbf{x}_0 + t\hat{\mathbf{y}}$
  $\mathbf{u}_t^* \leftarrow \hat{\mathbf{y}} - \mathbf{x}_0$ ▷ *Eq.* (28)
  $\mathcal{L}_{\mathrm{CFM}} \leftarrow \left\| u_\theta(\mathbf{x}_t, t, \mathbf{h}) - \mathbf{u}_t^*\right\|_2^2$ ▷ *Eq.* (29)
  $\theta \leftarrow \mathrm{AdamW}(\theta, \nabla_\theta \mathcal{L}_{\mathrm{CFM}})$
  $\bar{\theta} \leftarrow \rho\bar{\theta} + (1-\rho)\theta$ ▷ *EMA*
end while
$\bar{\theta}^* \leftarrow \bar{\theta}$
return $\phi^*, \bar{\theta}^*$

Algorithm 2 Probabilistic inference of SPFlow.

Input: Forecast context $\mathcal{C} = (\mathbf{x}_{ns}, \mathbf{x}_{cl}, \mathbf{x}_{nwp}, \mathbf{c})$; f_{ϕ^*}, $u_{\bar{\theta}^*}$; δ, M, N
Output: $\widehat{\mathcal{Y}} = \{\widehat{\mathbf{y}}^{(m)}\}_{m=1}^{M}$
$(\hat{\mathbf{x}}_{ns}, \hat{\mathbf{x}}_{nwp}) \leftarrow \mathrm{Mask}_{85^\circ}(\mathcal{C})$
$\mathbf{y}_{sp} \leftarrow f_{\phi^*}(\hat{\mathbf{x}}_{ns}, \hat{\mathbf{x}}_{nwp}, \mathbf{x}_{cl}, \mathbf{c})$ ▷ *shape prior*
$\mathbf{h} \leftarrow [\hat{\mathbf{x}}_{ns}; \hat{\mathbf{x}}_{nwp}; \mathbf{c}]$
$\Delta t \leftarrow 1/N$
for $m = 1, \ldots, M$ **do**
  $\epsilon^{(m)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mathbf{x}_0^{(m)} \leftarrow \mathbf{y}_{sp} + \delta\epsilon^{(m)}$ ▷ *source sample*
  for $n = 0, \ldots, N-1$ **do**
    $t_n \leftarrow n\Delta t$
    $\mathbf{x}_{n+1}^{(m)} \leftarrow \mathbf{x}_n^{(m)} + \Delta t\, u_{\bar{\theta}^*}(\mathbf{x}_n^{(m)}, t_n, \mathbf{h})$ ▷ *Euler refinement*
  end for
  $\widehat{\mathbf{y}}^{(m)} \leftarrow \mathrm{Denorm}(\mathbf{x}_N^{(m)})$
end for
$\widehat{\mathcal{Y}} \leftarrow \{\widehat{\mathbf{y}}^{(m)}\}_{m=1}^{M}$
$\bar{\mathbf{y}} \leftarrow M^{-1}\sum_{m=1}^{M}\widehat{\mathbf{y}}^{(m)}$
$(\mathbf{q}_{0.05}, \mathbf{q}_{0.95}) \leftarrow \mathrm{Quantile}_{0.05,0.95}(\widehat{\mathcal{Y}})$
return $\widehat{\mathcal{Y}}, \bar{\mathbf{y}}, \mathbf{q}_{0.05}, \mathbf{q}_{0.95}$

4. Performance evaluation

4.1. Training and inference procedure

The experimental dataset is temporally partitioned into two subsets: data spanning 2019–2021 serves as the training set, and data from 2022 is reserved for final testing. To ensure operational and leakage-free evaluation, all models use only information available at the forecast origin. Historical SURFRAD and NSRDB data are restricted to the period before initialization, while the forecast window uses only 00Z HRRR forecasts, clear-sky GHI, calendar variables, and fixed geographical metadata. Target-period SURFRAD GHI is used only for training supervision and post-forecast evaluation, and all normalization statistics are computed from the training set and fixed thereafter.

Notably, cross-learning is employed to enhance generalization during the training process, and data from all five stations are jointly utilized to train a single unified model. All experiments are conducted on a single NVIDIA A6000 GPU. The training procedure for SPFlow is executed in two distinct stages: the SPTransformer is optimized first, followed by the training of the conditional flow matching component. Model optimization is performed using the AdamW optimizer with $\beta_1 = 0.95$, $\beta_2 = 0.999$, and a weight decay of 1×10^{-5}. The δ is set

Table 1
Hyperparameter configurations for the SPTransformer and Flow Matching components.

SPTransformer architecture		Flow matching & optimization	
Input Length (L_{in})	288	Flow Hidden Dim	256
Pred Length (L_{out})	96	Flow Residual Blocks	8
Patch Length	6	Time Step Embedding	50
Stride (S)	3	Inference Steps	50
Hidden Dim	64	Optimizer	AdamW
Attention Heads (H)	4	Learning Rate	2×10^{-4}
Enc/Dec Layers	2	Weight Decay	1×10^{-5}
Dropout	0.1	Betas (β_1, β_2)	(0.95, 0.999)
Batch Size	64	EMA Decay	0.9999
Soft-DTW	$\gamma = 0.1$		

to 0.1. An initial learning rate of 2×10^{-4} is employed, regulated by a cosine annealing scheduler with restarts, which includes 500 warm-up steps and 2 restart cycles. To enhance training stability, the Exponential Moving Average (EMA) mechanism [56] is applied to the flow matching parameters. Detailed hyperparameter configurations are provided in Table 1. Finally, for probabilistic inference, an ensemble size of 10 is generated for all downscaling models. On a single NVIDIA A6000 GPU, the two-stage training of SPFlow takes approximately 1.5 h, and generating a 10-member ensemble for one day-ahead forecast window requires around 1.2 s, roughly four times faster than CSDI (4.8 s). The acceleration arises from the straight-line optimal-transport probability path and the shape-prior warm-start, which together yield low-curvature trajectories that are accurately integrated with a single Euler update per step, in contrast to the multi-step score-based denoising required by CSDI.

4.2. Benchmarks

To rigorously evaluate SPFlow, we consider a diverse set of benchmark models, including classical statistical methods, deterministic deep learning models, and probabilistic generative approaches. Specifically, the compared baselines include Smart Persistence (SP), ARIMA, Gaussian Processes (GP), LSTM, iTransformer [57], CSDI [29], and two internal variants, namely SPFlow-G and SPFlow-C. In addition, we include CSI-NWP as a reference baseline, which directly converts the hourly HRRR forecasts into 15-min resolution using the clear-sky-index interpolation. These benchmarks were selected because they represent standard references for solar forecasting, temporal downscaling, and probabilistic time-series generation. Detailed descriptions, implementation settings, and model rationales are provided in Appendix.

4.3. Assessment metrics

To quantify down-scaling performance, we constructed a comprehensive assessment framework combining probabilistic and deterministic metrics to investigate the reliability and sharpness of the prediction ensemble [58,59]. All models generate the complete 96-step trajectory corresponding to HRRR lead times of 1–24 h. Following established solar forecast verification practice [45,60,61], only samples satisfying SZA $\leq 85°$ are retained for metric calculation. Thus, the nominal forecast horizon remains 1–24 h, while the evaluated lead times correspond to the site- and season-dependent daylight subset determined by solar geometry.

The Continuous Ranked Probability Score (CRPS) [59] is used to evaluate the overall quality of probabilistic downscaling. CRPS integrates the characteristics of reliability and sharpness of the probabilistic forecasting results by calculating the distance between the cumulative distribution function $\hat{F}^i$ of the ensemble forecast $\hat{Y}_i$ and the Heaviside function H centered on the observation Y_i:

$$\mathrm{CRPS} = \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\left(\hat{F}^i(x) - H(x - Y^i)\right)^2 dx, \quad (32)$$

where N denotes the number of samples in the ensemble. A lower magnitude of CRPS indicates better performance of probabilistic forecasts. Statistical significance of the CRPS differences is assessed using a paired day-level bootstrap with 10,000 resamples. For each test day, CRPS is first averaged over the 96 forecast steps, after which the paired difference between SPFlow and the comparison model is calculated. Resampling is performed at the day level to preserve the correspondence between models and the temporal dependence within each forecast window. The mean paired differences and their 95% bootstrap confidence intervals are reported, and an improvement is considered statistically significant when the confidence interval excludes zero.

Furthermore, to explicitly evaluate the quality of the prediction intervals, we utilize the Prediction Interval Coverage Probability (PICP) and the Mean Prediction Interval Width (MPIW). Throughout this study, these metrics are reported at the nominal confidence level of $(1 - \alpha) = 90\%$, with the lower and upper bounds constructed from the 5^{th} and 95^{th} empirical quantiles of the 10-member ensemble forecast. For the ith sample, let $\hat{L}_i$ and $\hat{U}_i$ denote the corresponding interval bounds [59]. At this nominal level, a PICP close to 0.9 indicates that the empirical coverage of the 90% prediction interval is close to its nominal value, while a smaller MPIW indicates greater sharpness conditional on adequate coverage. Because PICP is evaluated at only one nominal level, it is interpreted here as a measure of empirical interval reliability rather than evidence of full distributional calibration. With $M = 10$ ensemble members, the empirical 5th and 95th percentiles are obtained by linear interpolation between adjacent order statistics. The PICP measures empirical reliability:

$$\mathrm{PICP} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{L}_i \leq y_i \leq \hat{U}_i), \quad (33)$$

where $\mathbb{I}(\cdot)$ is the indicator function, which equals 1 if the condition holds and 0 otherwise. Complementing this, the MPIW assesses the sharpness of the forecasts by calculating the average width of the intervals:

$$\mathrm{MPIW} = \frac{1}{N}\sum_{i=1}^{N}(\hat{U}_i - \hat{L}_i). \quad (34)$$

In addition to evaluating probabilistic predictions, the root mean square error (RMSE), mean absolute error (MAE), mean bias error (MBE), coefficient of determination (R^2), normalized RMSE (nRMSE), and normalized MBE (nMBE) are used to evaluate the deterministic error between the predicted ensemble mean and the observed values:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}, \quad (35)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|, \quad (36)$$

$$\mathrm{MBE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i), \quad (37)$$

$$\mathrm{nRMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2/\bar{y}_i}, \quad (38)$$

$$\mathrm{nMBE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)/\bar{y}_i, \quad (39)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}, \quad (40)$$

where $\hat{y}_i$ and y_i denote the forecast and observation for the ith sample, respectively, $\bar{y}$ denotes the mean of the observations, and n is the total number of evaluated samples.

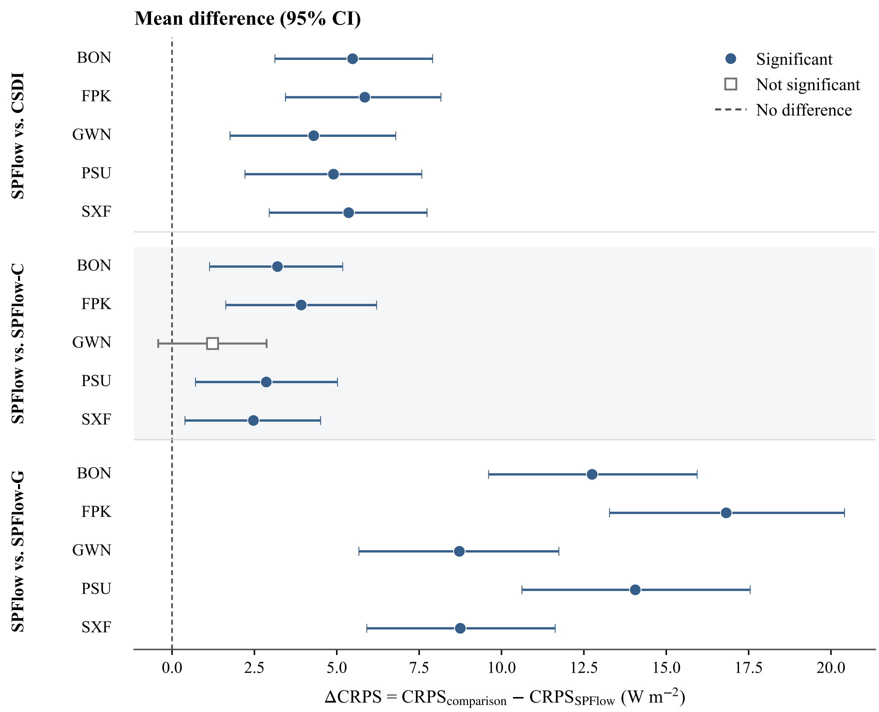


Fig. 4. Paired day-level bootstrap analysis of CRPS differences between SPFlow and the probabilistic comparison models. Points denote the mean paired differences, defined as $\Delta\mathrm{CRPS} = \mathrm{CRPS}_{\mathrm{comparison}} - \mathrm{CRPS}_{\mathrm{SPFlow}}$, and horizontal error bars indicate the 95% bootstrap confidence intervals. Positive values favor SPFlow, while confidence intervals crossing zero indicate that the difference is not statistically significant.

5. Results and discussion

5.1. Comparison with benchmarks

Table 2 presents the probabilistic forecasting performance across five sites (BON, FPK, GWN, PSU, and SXF). Overall, the proposed SPFlow model achieves the lowest mean CRPS at all five stations. At BON, for example, SPFlow obtains a CRPS of 44.93 W/m^2, corresponding to reductions of approximately 22.5% relative to GP (57.97 W/m^2) and 10.9% relative to CSDI (50.41 W/m^2). The paired day-level bootstrap analysis in Fig. 4 further shows that the CRPS reductions relative to CSDI are statistically significant at all five stations, as the corresponding 95% confidence intervals lie entirely above zero. This confirms that the observed ranking is robust to test-day sampling variability.

Beyond overall accuracy, the superiority of SPFlow is further demonstrated by examining the trade-off between reliability (PICP) and sharpness (MPIW). The GP model attains high PICP values but tends to over-cover the nominal 0.9 level, doing so at the cost of excessively wide prediction intervals, with MPIW values ranging from 444 to 488 W/m^2. In contrast, SPFlow maintains MPIW values largely between 177 and 208 W/m^2, reducing the prediction interval width by more than 55% while keeping PICP close to the nominal 0.9 level. Compared with CSDI, SPFlow produces sharper intervals in most cases and achieves PICP values that are consistently closer to the nominal coverage level. Overall, these results demonstrate that SPFlow provides a more favorable reliability–sharpness trade-off than the probabilistic baselines.

Table 3 presents a comprehensive deterministic evaluation of the proposed SPFlow model against multiple benchmark methods across five SURFRAD stations, demonstrating its superior performance in day-ahead NWP GHI temporal downscaling. SPFlow consistently achieves the lowest error metrics and highest goodness-of-fit scores (R^2), significantly outperforming traditional methods and advanced deep learning competitor models. Specifically, at the BON station, SPFlow reduced RMSE by approximately 36.1% compared with ARIMA, by 34.5% relative to CSI-NWP, and by 38.1% relative to SP; at the GWN station, its MAE decreased by 8.2% relative to CSDI. Compared with the deterministic SPTransformer backbone, the complete SPFlow framework reduces RMSE by approximately 7.7%–11.2% across the five stations, demonstrating the additional contribution of the conditional flow-matching refinement stage. Notably, Smart Persistence ranks as the weakest baseline at all five stations, confirming that simple clear-sky scaling of a single past observation is insufficient to reconstruct sub-hourly cloud-driven variability over day-ahead horizons; its relatively lower nMBE compared with ARIMA indicates that while clear-sky normalization partially corrects the diurnal bias, the high RMSE and low R^2 (e.g., 0.605 at PSU) reveal its inability to capture stochastic irradiance fluctuations. Furthermore, SPFlow demonstrates robust generalization capabilities, maintaining R^2 values above 0.84 in most scenarios. Similarly, in terms of bias, the model demonstrates minimal systematic error, with nMBE values remaining tightly clustered near zero (e.g., 0.251% at SXF), whereas baselines such as GP exhibit larger directional biases (e.g., −13.0% at SXF). These results quantitatively validate the effectiveness of SPFlow in capturing complex temporal dependencies and generating accurate downscaled irradiance trajectories.

5.2. Performance under different weather conditions

To validate the model's predictive performance under varying weather conditions, this study employs the weather condition classification method proposed in [7] to categorize results from five observation stations. Samples are subsequently analyzed under two distinct scenarios: clear and cloudy. Fig. 5 illustrates the performance comparison of SPFlow against benchmark models (CSDI, iTransformer, and ARIMA) across five observation stations (BON, FPK, GWN, PSU, and SXF) under varying meteorological conditions. Overall, across all evaluation scenarios, the proposed SPFlow consistently outperformed the comparison methods in both RMSE and MAE metrics, demonstrating its consistent performance advantage in day-ahead NWP GHI downscaling. Specifically, in the cloudy scenario characterized by high uncertainty, SPFlow substantially reduced errors. The RMSE decreased by approximately 10%–45% compared to the statistical model ARIMA, while

Table 2
Probabilistic temporal downscaling performance evaluation across five stations. PICP and MPIW are reported at the 90% nominal confidence level. CRPS is independent of the chosen interval level. Bold values mark the best result per column: minimum CRPS, PICP closest to the nominal 0.9, and minimum MPIW among models whose PICP attains at least 0.88; models failing this coverage threshold are excluded from MPIW ranking.

Models	BON			FPK			GWN			PSU			SXF		
	CRPS	PICP	MPIW	CRPS	PICP	MPIW	CRPS	PICP	MPIW	CRPS	PICP	MPIW	CRPS	PICP	MPIW
GP	57.97	0.945	461.62	60.24	0.939	451.74	62.90	0.954	488.27	65.98	0.918	444.50	56.73	0.957	467.31
CSDI	50.41	0.873	246.97	53.74	0.860	248.15	56.53	0.866	249.44	60.02	0.829	243.51	50.01	0.857	247.02
SPFlow-C	48.13	0.853	191.65	51.81	0.832	188.78	53.46	0.866	190.47	57.98	0.808	190.49	47.12	0.838	193.59
SPFlow-G	57.68	0.795	209.93	64.71	0.787	206.50	60.95	0.823	216.94	69.18	0.752	207.16	53.40	0.815	210.79
SPFlow	**44.93**	**0.913**	**189.27**	**47.89**	**0.906**	**202.13**	**52.23**	**0.892**	**207.69**	**55.12**	**0.889**	**177.42**	**44.65**	**0.908**	**179.82**

Table 3
Performance evaluation of SPFlow and baseline models at five stations. For probabilistic models, deterministic evaluation metrics are calculated using the mean of 10 samples as predictions.

Station	Metrics	Models										
		SPFlow	SPFlow-C	SPFlow-G	CSDI	GP	SPTransformer	CSI-NWP	iTransformer	LSTM	ARIMA	SP
BON	RMSE	**96.14**	97.22	113.0	100.0	113.0	107.5	146.8	106.7	121.1	150.5	155.2
	nRMSE (%)	**26.36**	26.66	30.96	27.51	30.85	29.47	40.25	29.25	33.20	41.27	42.56
	MBE	1.764	−1.488	−22.34	0.357	−45.95	0.484	20.35	−0.579	5.698	28.17	15.43
	nMBE (%)	0.484	−0.408	−6.127	0.133	−12.59	0.134	5.580	−0.159	1.563	7.725	4.231
	MAE	**57.10**	59.36	68.09	61.51	77.26	70.46	87.90	69.75	79.76	90.50	94.17
	R^2	**0.885**	0.882	0.841	0.875	0.842	0.856	0.732	0.858	0.817	0.718	0.700
FPK	RMSE	**103.0**	**103.0**	123.0	107.0	117.0	113.0	141.9	113.1	126.2	145.1	149.8
	nRMSE (%)	**29.29**	**29.29**	35.09	30.43	33.30	32.23	40.47	32.26	35.98	41.40	42.74
	MBE	2.844	0.408	−31.52	3.743	−45.27	3.538	17.82	3.734	8.614	24.85	13.62
	nMBE (%)	0.811	0.116	−8.989	1.065	−12.91	1.009	5.080	1.065	2.457	7.088	3.884
	MAE	**61.14**	62.15	74.78	65.07	78.31	73.54	82.70	73.73	83.80	85.65	89.43
	R^2	**0.855**	**0.855**	0.792	0.844	0.813	0.825	0.724	0.824	0.782	0.711	0.693
GWN	RMSE	**116.0**	117.0	128.0	119.0	127.0	125.7	159.2	125.6	136.4	164.3	168.7
	nRMSE (%)	**28.19**	28.39	31.02	28.99	30.94	30.53	38.66	30.49	33.13	39.90	40.97
	MBE	8.219	−0.033	−10.66	5.254	−41.64	5.029	21.73	5.784	9.959	26.45	14.28
	nMBE (%)	1.955	−0.008	−2.588	1.276	−10.11	1.221	5.280	1.404	2.418	6.423	3.468
	MAE	**62.22**	67.26	70.92	67.79	83.42	76.72	92.10	76.82	86.35	95.62	99.35
	R^2	**0.855**	0.853	0.824	0.847	0.826	0.830	0.728	0.831	0.800	0.710	0.694
PSU	RMSE	**110.0**	113.0	130.0	115.0	128.0	122.1	166.1	121.1	136.6	172.7	177.4
	nRMSE (%)	**31.91**	32.63	37.70	33.36	36.92	35.26	48.01	34.99	39.48	49.89	51.25
	MBE	−6.609	−7.349	−26.06	−7.545	−51.36	−7.944	34.86	−7.527	−0.199	41.73	20.16
	nMBE (%)	−1.909	−2.123	−7.531	−2.175	−14.84	−2.295	10.08	−2.180	−0.057	12.06	5.826
	MAE	**67.80**	72.08	80.10	72.08	85.88	80.04	103.4	79.38	89.36	108.6	112.8
	R^2	**0.847**	0.840	0.787	0.833	0.795	0.813	0.654	0.816	0.766	0.626	0.605
SXF	RMSE	**92.12**	94.07	101.8	97.31	108.4	103.7	134.7	103.8	117.5	138.7	142.5
	nRMSE (%)	**24.83**	25.36	27.44	26.23	29.23	27.95	36.31	27.99	31.67	37.38	38.41
	MBE	0.933	−2.968	−26.04	0.194	−48.24	0.061	20.12	0.224	6.141	25.81	12.97
	nMBE (%)	0.251	−0.800	−7.018	0.052	−13.0	0.016	5.420	0.060	1.655	6.957	3.496
	MAE	**56.43**	58.82	63.37	61.47	75.00	69.66	78.30	69.89	79.24	80.72	84.16
	R^2	**0.889**	0.884	0.865	0.876	0.846	0.860	0.763	0.859	0.820	0.749	0.734

maintaining a performance advantage of about 2%–8% over the deep learning-based probabilistic model CSDI and the deterministic model iTransformer. This indicates SPFlow's superior ability to characterize underlying temporal dependencies and uncertainty structures under complex cloud-variation conditions. These advantages persist in both clear sky conditions and all-sample scenarios. SPFlow consistently exhibits the lowest error levels across all observation stations, with its RMSE bars and MAE curves remaining consistently below those of the baseline models, demonstrating robust stability and consistency. These results indicate that SPFlow not only excels under specific meteorological conditions but also possesses strong generalization capabilities and robustness across diverse weather scenarios and spatial locations, effectively addressing the challenges posed by weather variability to downscaling forecasts.

5.3. Performance under different seasons

To systematically evaluate the model's predictive performance under different seasonal conditions, the samples were divided by season: December to February is defined as winter, March to May as spring, June to August as summer, and September to November as autumn. Fig. 6 compares the performance of six models in day-ahead NWP GHI downscaling across five observation sites and four seasonal conditions. Results indicate that SPFlow achieves the lowest or near-lowest RMSE and MAE across all sites and seasons, demonstrating consistently better overall performance than the benchmark models. Although all methods exhibit higher errors during summer due to intense convective activity and rapid changes in cloud cover, SPFlow maintained the strongest robustness, reducing RMSE and MAE by approximately 25%–45% and 20%–40%, respectively, compared to traditional ARIMA models. It further achieved error reductions of approximately 5%–20% relative to CSDI and iTransformer at most sites and seasons. In scenarios with relatively stable irradiation conditions, such as winter and autumn, SPFlow maintained a consistent lead with significantly lower error fluctuations than models like LSTM and GP. These results demonstrate that SPFlow effectively captures the nonlinear temporal dependencies and seasonal variations inherent in NWP errors. It exhibits strong generalization capabilities and prediction stability across diverse meteorological conditions and spatial locations, making it suitable for high-precision GHI downscaling applications.

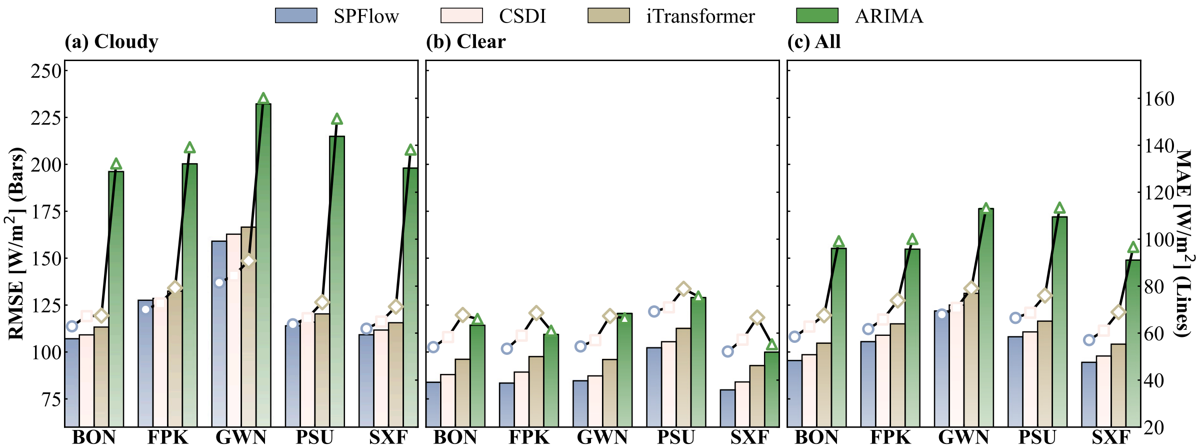


Fig. 5. Performance comparison of models under different weather conditions. Bars indicate RMSE (left axis), and lines indicate MAE (right axis). Panels (a)–(c) correspond to Cloudy, Clear, and All scenarios, respectively.

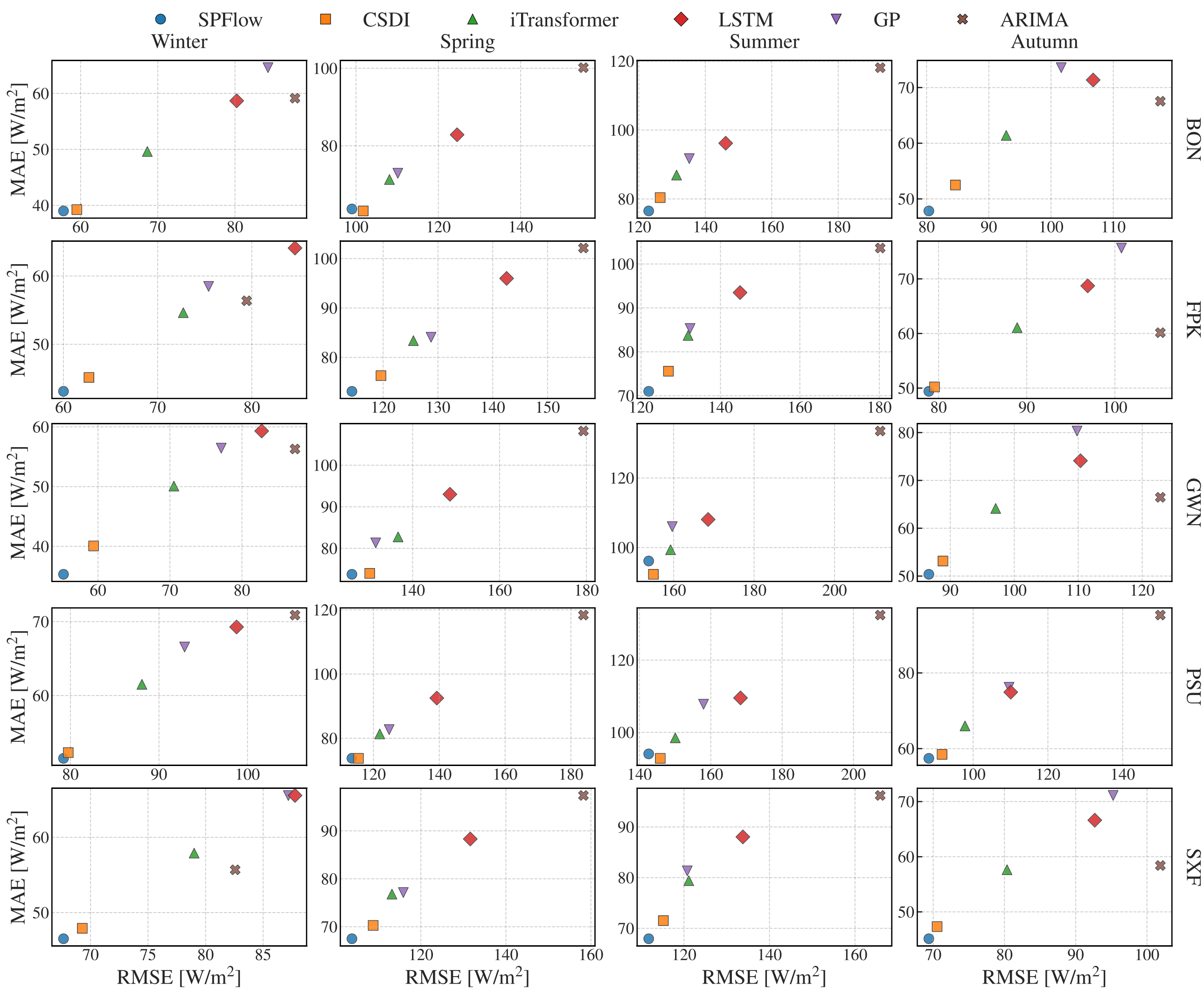


Fig. 6. Seasonal downscaling error comparison of six models across five stations. Models situated towards the bottom-left corner demonstrate superior performance with smaller errors.

5.4. Quantifying the effectiveness of the shape prior

Having established SPFlow's superiority over the benchmark models, we conduct an ablation study to quantify the effectiveness of the learned shape prior. Specifically, SPFlow is compared with two variants using different source distributions: SPFlow-C and SPFlow-G, as shown in Fig. 7. SPFlow-G, which adopts a standard Gaussian source, consistently yields the highest CRPS and the lowest PICP among

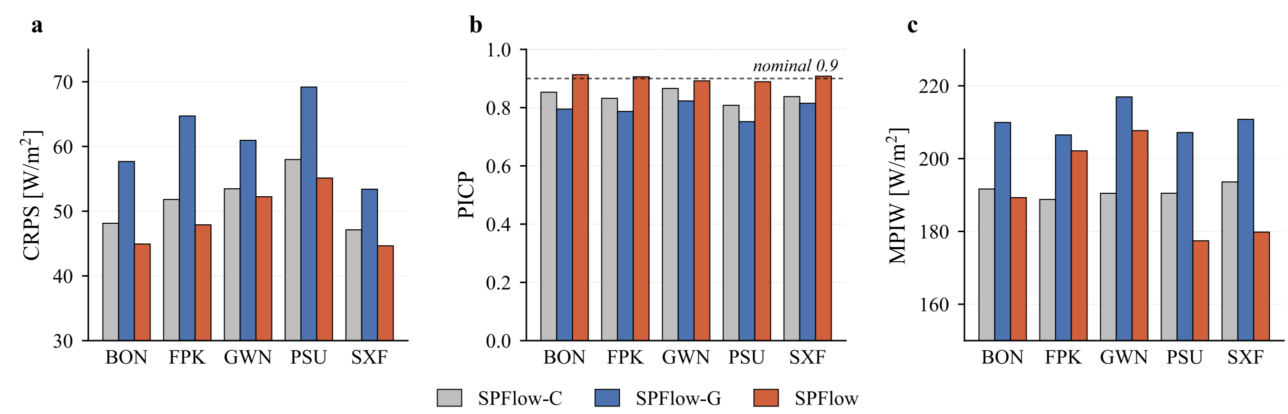


Fig. 7. Performance evaluation of SPFlow, SPFlow-C, and SPFlow-G models across five SURFRAD stations.

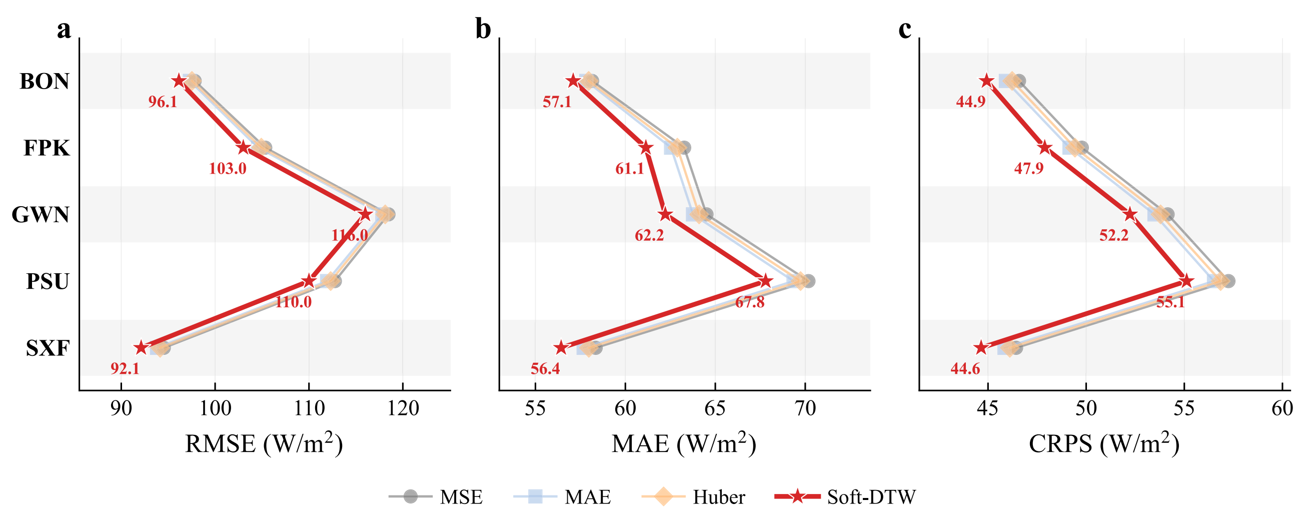


Fig. 8. Comparison of RMSE, MAE, and CRPS across five SURFRAD stations for four loss functions used to train the SPTransformer backbone. Numerical annotations indicate the Soft-DTW results.

the flow-based models. At PSU, for example, its PICP decreases to 0.752, substantially under-covering the nominal 0.9 level compared with SPFlow (0.889). Using clear-sky GHI as the source prior in SPFlow-C improves the results relative to Gaussian initialization. At BON, SPFlow further reduces CRPS by 6.6% relative to SPFlow-C and increases PICP from 0.853 to 0.913. The paired bootstrap results in Fig. 4 provide statistical support for these improvements. SPFlow significantly reduces CRPS relative to SPFlow-G at all five stations. Relative to SPFlow-C, the reductions are statistically significant at BON, FPK, PSU, and SXF, whereas the confidence interval at GWN crosses zero, indicating only a numerical improvement at this station. The SPTransformer, therefore, provides a more informative deterministic estimate of the irradiance shape, while the flow-matching module acts as a probabilistic residual corrector that learns the remaining mismatch between the shape prior and the target trajectory. Overall, the statistically significant improvements at four of the five stations relative to SPFlow-C, together with significant gains over SPFlow-G at all stations, demonstrate that the performance improvement arises not only from the flow matching model itself, but also from the learned informative prior and its subsequent residual refinement.

5.5. Ablation study on the loss function

Fig. 8 compares four candidate loss functions across five SURFRAD sites under three evaluation metrics. The three pointwise objectives (MSE, MAE, Huber) yield nearly indistinguishable results, confirming that the choice among standard regression losses exerts a limited influence on the learned shape prior. By contrast, Soft-DTW attains the lowest error at every site under all three metrics. Relative to the strongest pointwise baseline (MAE), the CRPS gain ranges from 2.8% to 3.9%, and is most pronounced at GWN and PSU, where irradiance variability and short-term ramping are stronger. The performance gap further widens when moving from RMSE and MAE to CRPS, indicating that elastic temporal alignment improves not only deterministic reconstruction but also the quality of the predictive distribution produced by the downstream flow-matching model.

Spectral diagnostics at the BON station, which exhibits the most pronounced sub-hourly variability among the five SURFRAD sites, are further employed to verify the advantage of Soft-DTW over pointwise losses in capturing high-frequency dynamics. For each test day, the one-sided power spectral density (PSD) of the 96-step day-ahead forecast is obtained via the discrete Fourier transform ($\Delta t = 15$ min, $N = 96$) and averaged across all test days to suppress sample-specific stochasticity. The frequency axis is partitioned into three operationally meaningful bands: low ($f < 1/6\,\mathrm{h}^{-1}$, diurnal envelope), mid ($1/6 \le f < 1\,\mathrm{h}^{-1}$, synoptic cloud variability), and high ($f \ge 1\,\mathrm{h}^{-1}$, sub-hourly ramps). Two diagnostics quantify spectral fidelity: the spectral ratio $\mathcal{R}(f) = S_{\hat{y}}(f)/S_y(f)$, which measures band-wise energy preservation, and the high-frequency energy retention

$$\eta_{\mathrm{HF}} = \frac{\int_{f_h}^{f_{\max}} S_{\hat{y}}(f)\,df}{\int_{f_h}^{f_{\max}} S_y(f)\,df}, \quad f_h = 1\,\mathrm{h}^{-1}. \tag{41}$$

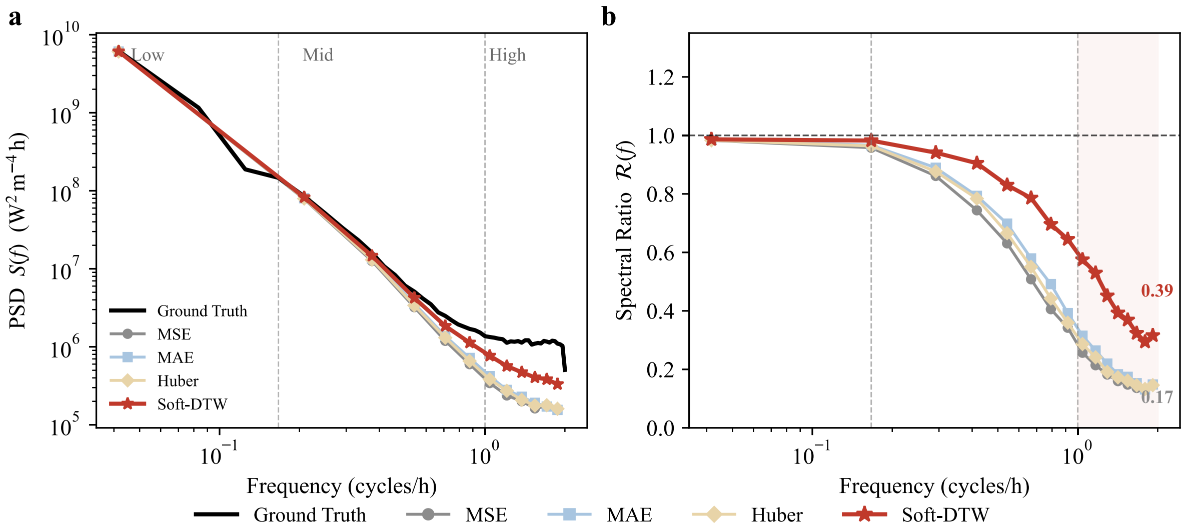


Fig. 9. Frequency-domain diagnostics at the BON station. **a**, Mean PSD of ground truth and of SPTransformer predictions trained with four loss functions. Vertical dashed lines delineate low-, mid-, and high-frequency bands. **b**, Spectral ratio $\mathcal{R}(f) = S_{\hat{y}}(f)/S_y(f)$; the dashed line marks perfect reconstruction $\mathcal{R} = 1$. Annotated values (0.39 for Soft-DTW and 0.17 for MSE) report $\mathcal{R}(f)$ at the Nyquist frequency $f_{\max} = 2\ \mathrm{h}^{-1}$, whereas the band-integrated metric η_{HF} reported in the main text aggregates spectral energy over the entire high-frequency band $f \geq 1\ \mathrm{h}^{-1}$.

Values approaching unity indicate faithful reconstruction, whereas values substantially below unity signal the over-smoothing characteristic of spectrally biased models.

The diagnostics reported in Fig. 9 reveal distinct behaviors across the three frequency bands. In the low-frequency band, all four loss functions reconstruct the diurnal envelope almost perfectly, with $\mathcal{R} \approx 1$ for $f < 1/6\,\mathrm{h}^{-1}$, indicating that the diurnal cycle is not the primary source of the observed performance differences. In the high-frequency band, the three pointwise losses exhibit pronounced spectral attenuation, retaining only $\eta_{\mathrm{HF}} = 0.46$ under MSE, 0.52 under MAE, and 0.49 under Huber. This systematic suppression of high-frequency energy provides empirical evidence of spectral bias, which may be associated with the double-penalty effect imposed by pointwise losses on temporally misaligned ramps. In contrast, Soft-DTW retains $\eta_{\mathrm{HF}} = 0.64$, corresponding to a 39% relative improvement over MSE. This result indicates that Soft-DTW reduces spectral attenuation and better preserves the magnitude of sub-hourly irradiance variability, rather than fully recovering all ramp-related information. The largest gain is concentrated within $f \in [1/(2\,\mathrm{h}),\ 1/(0.5\,\mathrm{h})]$, demonstrating that the advantage of Soft-DTW is primarily associated with short-timescale variations. By relaxing strict point-to-point correspondence in favor of trajectory-level alignment, Soft-DTW reduces the excessive penalty assigned to moderately phase-shifted but morphologically similar trajectories, thereby mitigating the tendency to oversmooth rapid transitions. Nevertheless, high-frequency energy retention characterizes spectral fidelity rather than the timestamp-level alignment of individual ramp events. Because frequency-domain analysis does not preserve temporal localization and Soft-DTW permits elastic alignment during training, these results should not be interpreted as direct evidence of improved ramp-timing accuracy. Instead, they provide complementary evidence that Soft-DTW better retains the spectral content of rapid irradiance variations than the pointwise loss functions.

5.6. *Traceable trajectory analysis*

To elucidate how SPFlow optimizes the prediction mechanism, Fig. 10 visualizes the trajectory evolution of the flow matching process at the BON station. Unlike conventional flow-matching models initialized from an uninformative Gaussian distribution, SPFlow employs the deterministic output of SPTransformer, and subsequently injects Gaussian perturbations as an informative prior (initial condition x_0). As depicted in the top row, the inference process is modeled as an ODE integration problem. The SPTransformer provides warm-starting, capturing the global trend of GHI. Subsequently, the flow matching model constructs a time-varying vector field $v_t(x)$, guiding the state trajectory to evolve from the shape prior towards the target data distribution. The blue gradient streamlines clearly illustrate the steps of Euler integration. During this process, the model iteratively corrects biases and local fluctuations present in the SPTransformer output. The quantified metrics in the figure demonstrate that this correction process significantly reduces errors. On 2022-11-08, the RMSE decreased from 193.11 W/m^2 for SPTransformer to 103.56 W/m^2 for SPFlow. The cases shown in Fig. 10 were intentionally selected to illustrate the refinement behavior of SPFlow under challenging conditions and are not representative of its average performance; the overall improvements should instead be assessed from the aggregate results reported in the benchmark tables.

Velocity heatmaps provide a granular analysis of the correction dynamics, providing a detailed interpretation of the correction dynamics for the flow matching process. These visualizations spatially and temporally map the magnitude of the model's rectification efforts. Specifically, high-velocity zones (indicated in deep red) align with temporal intervals where the SPTransformer's prior significantly deviates from the true distribution. Within these high-error regimes, the learned vector fields exhibit pronounced activation, imposing substantial corrective gradients to realign the trajectories. In contrast, low-velocity areas (lighter hues) indicate that the prior predictions maintain high fidelity, requiring only negligible fine-tuning via the ODE solver. Consequently, these findings provide evidence that SPFlow operates as an adaptive residual corrector, allocating dynamic adjustments exclusively where the SPTransformer's initial estimation lacks sufficient precision.

5.7. *Interpretability of deformable attention modules*

To assess the effectiveness of VAU and TAU, we compare the learned attention maps with those produced by standard self-attention. The attention matrices in Fig. 11 are averaged over 200 randomly sampled test samples from the BON station to improve statistical robustness. In the variable dimension, standard self-attention (Fig. 11(a)) produces a relatively diffuse weighting pattern (0.14–0.24), with limited discrimination among variables and no clearly identifiable physical interaction structure. In contrast, VAU (Fig. 11(b)) yields sparse and interpretable attention concentrations, highlighting three coherent interaction groups: a radiation-related block associated with $\mathrm{GHI}_{\mathrm{lag}}$ and

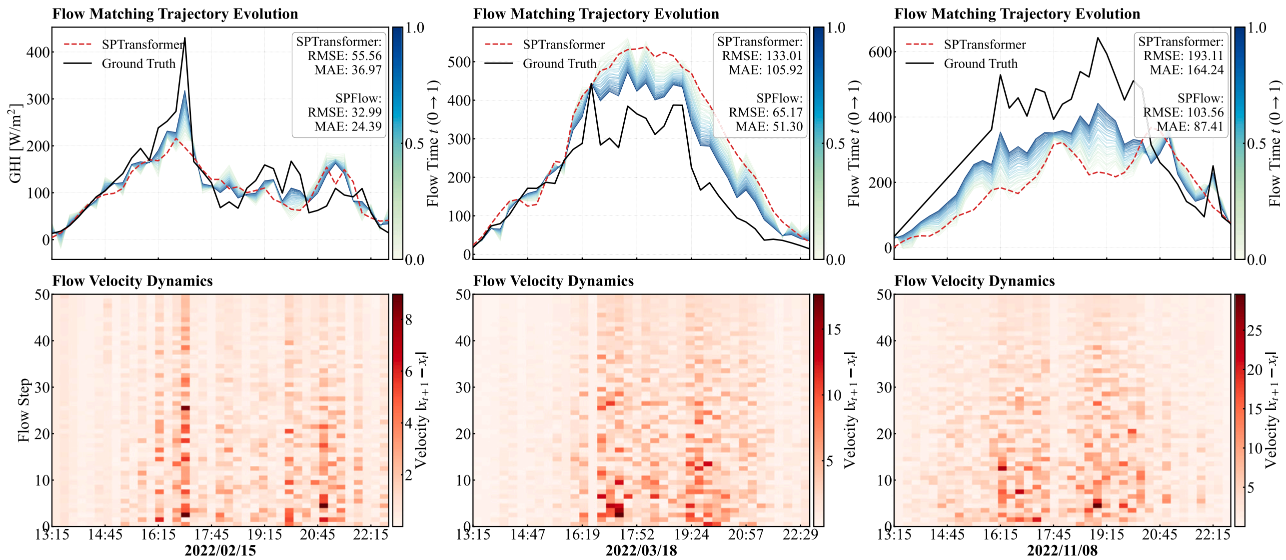


Fig. 10. Visualization of the traceable inference process based on SPFlow. This diagram illustrates the generative trajectories of three representative samples under cloudy weather conditions. Top row: Trajectory evolution of GHI values during the Euler integration process ($t = 0 \rightarrow 1$). Note that the initial distribution at $t = 0$ is intentionally misaligned with the SPTransformer prior due to the injection of Gaussian noise (Eq. (27)); this perturbation is designed to enhance the diversity of the inference samples. Bottom row: The heatmap visualizes the magnitude of the velocity vector $|v_t(\mathbf{x}_t)| = |\mathbf{x}_{t+1} - \mathbf{x}_t|$ at each flow step, highlighting the temporal and magnitude distribution characteristics of the corrective effort required to bridge the gap between prior and target.

CS-GHI, a thermo-hygrometric block associated with Temp and RH, and a wind-related block associated with WD and WS. In the temporal dimension, standard attention (Fig. 11(c)) mainly forms a broad diagonal band, indicating only weak locality and limited periodic structure. TAU (Fig. 11(d)), however, produces a much sharper diagonal concentration together with a distinct secondary ridge at approximately 12-step lag, suggesting that it captures both short-term temporal dependence and diurnal periodicity more effectively. Overall, these visualizations indicate that VAU and TAU learn more structured and physically meaningful dependencies than standard self-attention, providing the explanation for the forecasting gains.

5.8. Sensitivity analyses

The source-noise scale δ controls the dispersion of the conditional source distribution around the deterministic shape prior. This locally perturbed source is consistent with conditional and stochastic flow matching [31,62]. To assess sensitivity of forecasting performance on the value of δ, we evaluate $\delta \in \{0.02, 0.05, 0.10, 0.15, 0.20, 0.30\}$ at BON. For each setting, the pretrained SPTransformer is fixed and only the conditional flow-matching component is retrained under identical settings. As shown in Fig. 12, PICP increases smoothly from 0.872 at $\delta = 0.02$ to 0.921 at $\delta = 0.20$. Over $\delta \in [0.05, 0.20]$, coverage remains close to the nominal 0.9 level, while CRPS varies only between 44.93 and 45.24 W/m^2. At $\delta = 0.30$, the higher PICP of 0.934 is accompanied by a wider MPIW of 218.9 W/m^2 and a higher CRPS of 46.07 W/m^2. Conversely, $\delta = 0.02$ yields narrower intervals but insufficient coverage. These results show that $\delta = 0.10$ lies within a broad stable region of near-nominal coverage and low CRPS, rather than representing a narrowly tuned optimum. Therefore, $\delta = 0.10$ is adopted for further analysis.

Fig. 13 evaluates the sensitivity of CRPS and PICP to the number of ensemble members at the BON station. Increasing the ensemble size from 5 to 10 reduces CRPS from 46.12 to 44.93 W/m^2 and increases PICP from 0.887 to 0.913. Beyond 10 members, both metrics approach a stable plateau. At 50 members, CRPS is 44.71 W/m^2, representing a relative difference of only 0.49% compared with the 10-member result, while PICP increases by only 0.005, from 0.913 to 0.918. These small differences indicate that 10 members are sufficient to provide stable estimates of the principal probabilistic metrics in this study.

5.9. Zero-shot performance

Fig. 14 presents a comprehensive visualization of the joint and marginal distributions for the downscaled day-ahead NWP GHI prediction, evaluating the zero-shot performance of SPFlow against three benchmark models (CSDI, iTransformer, and ARIMA) at the TBL and DRA stations. To ensure a robust assessment of deterministic accuracy within a probabilistic framework, predictions for both SPFlow and CSDI are derived from the means of 10 ensemble members. The kernel density plots provide a clear visualization of the significant differences in model fidelity. SPFlow demonstrates exceptional capability in capturing the data distribution, characterized by the tightest clustering of scatter points along the diagonal. Unlike ARIMA, which exhibits significant dispersion and overestimation in the moderate irradiance range, SPFlow's high-density core region aligns precisely with the diagonal, indicating minimal downscaling bias. Quantitative metrics consistently demonstrate SPFlow's superiority over benchmark models. At DRA, SPFlow achieves the best performance among all evaluated models, with an RMSE of 81.04 W/m^2, an MAE of 45.80 W/m^2, and an R^2 of 0.926. In comparison, CSDI obtains an RMSE of 92.15 W/m^2, an MAE of 59.47 W/m^2, and an R^2 of 0.905. These results indicate that SPFlow transfers effectively to DRA despite the absence of station-specific training. The TBL site exhibited a similar trend, with SPFlow consistently maintaining the highest accuracy ($R^2 = 0.796$). The representative cases in Fig. 15 further demonstrate that SPFlow reliably tracks dynamic irradiance variations while maintaining coherent predictive intervals at both unseen stations. Collectively, these results demonstrate SPFlow's exceptional generalization capability and precision in capturing dynamic irradiance variations caused by cloud dynamics, even when encountering unseen geographical data in zero-sample scenarios.

6. Conclusion

This study proposes SPFlow, a shape-prior conditional flow matching framework for improving day-ahead probabilistic solar irradiance forecasting from coarse NWP products. The method addresses a key operational limitation in renewable energy forecasting: hourly NWP forecasts provide useful large-scale atmospheric information but cannot directly resolve the sub-hourly irradiance ramps required for PV

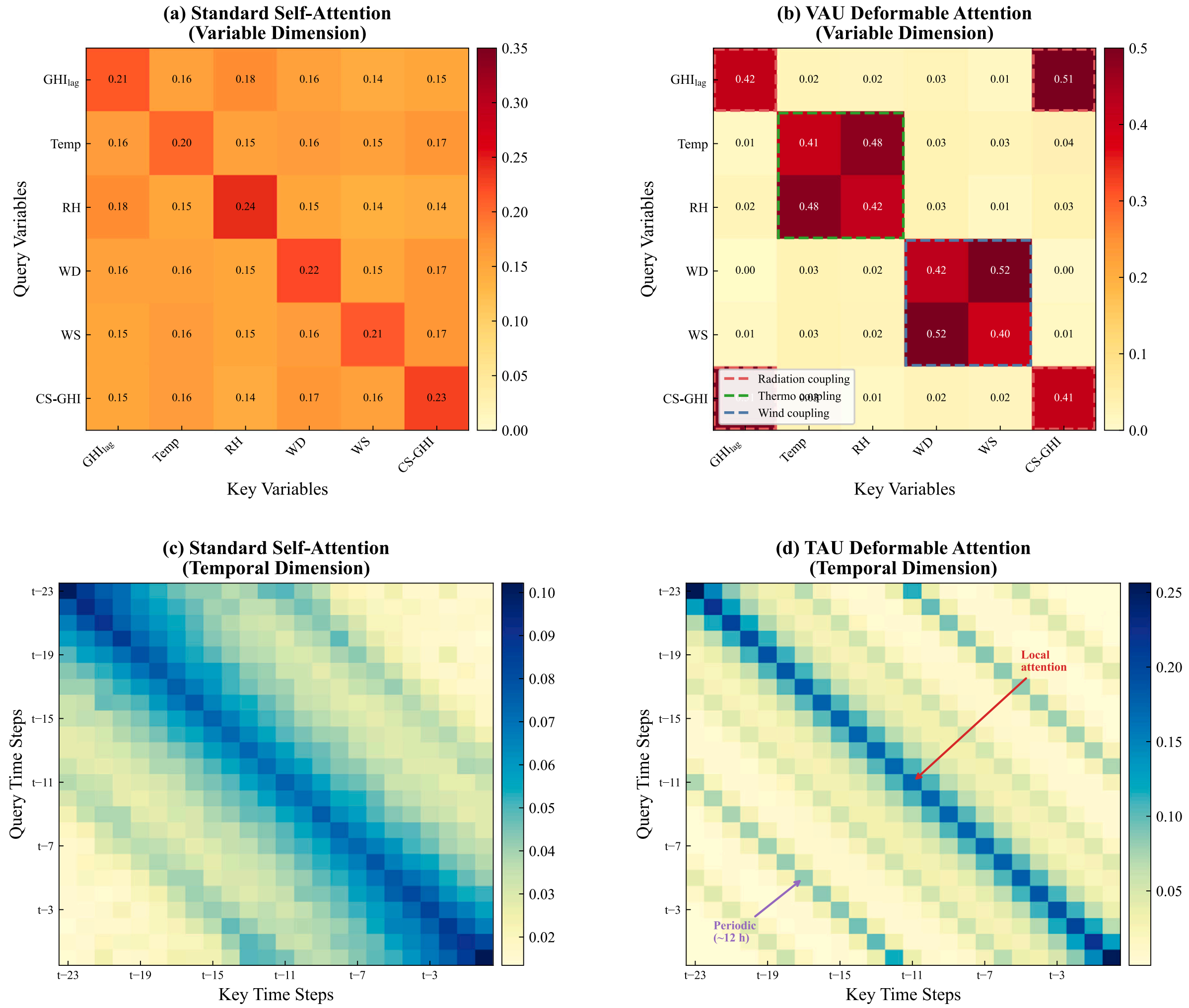


Fig. 11. Comparison of attention weight matrices between standard self-attention and the proposed deformable attention modules. (a) Standard self-attention in the variable dimension. (b) VAU deformable attention in the variable dimension; dashed boxes highlight automatically discovered physical couplings (red: radiation, green: thermo-hygrometric, blue: wind). (c) Standard self-attention in the temporal dimension. (d) TAU deformable attention in the temporal dimension, exhibiting a local attention band along the diagonal and a periodic ridge at ~12-step lag.

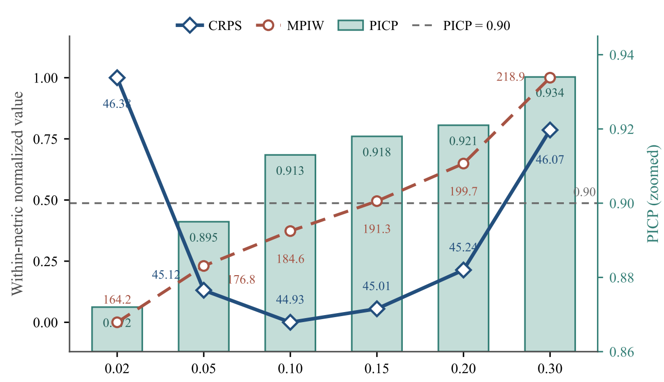


Fig. 12. Sensitivity of SPFlow to the source-noise scale δ at the BON station. CRPS and MPIW are normalized within each metric for joint visualization, while the right axis reports the original PICP values. The horizontal dashed line indicates the nominal PICP of 0.9.

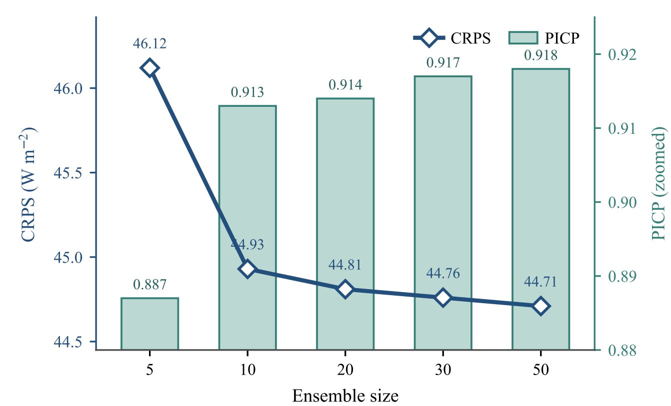


Fig. 13. Sensitivity of CRPS and PICP to ensemble size at the BON station. The line plot shows CRPS, while the bars show PICP for ensemble sizes of 5, 10, 20, 30, and 50. Both metrics approach a stable plateau after 10 ensemble members.

integration, storage scheduling, and grid operation. SPFlow overcomes this limitation by first constructing a physically meaningful shape prior using a Soft-DTW-trained SPTransformer and then refining this prior through conditional flow matching to generate probabilistic forecast trajectories with near-nominal empirical coverage at the evaluated 90% prediction-interval level. Extensive experiments on SURFRAD, NSRDB,

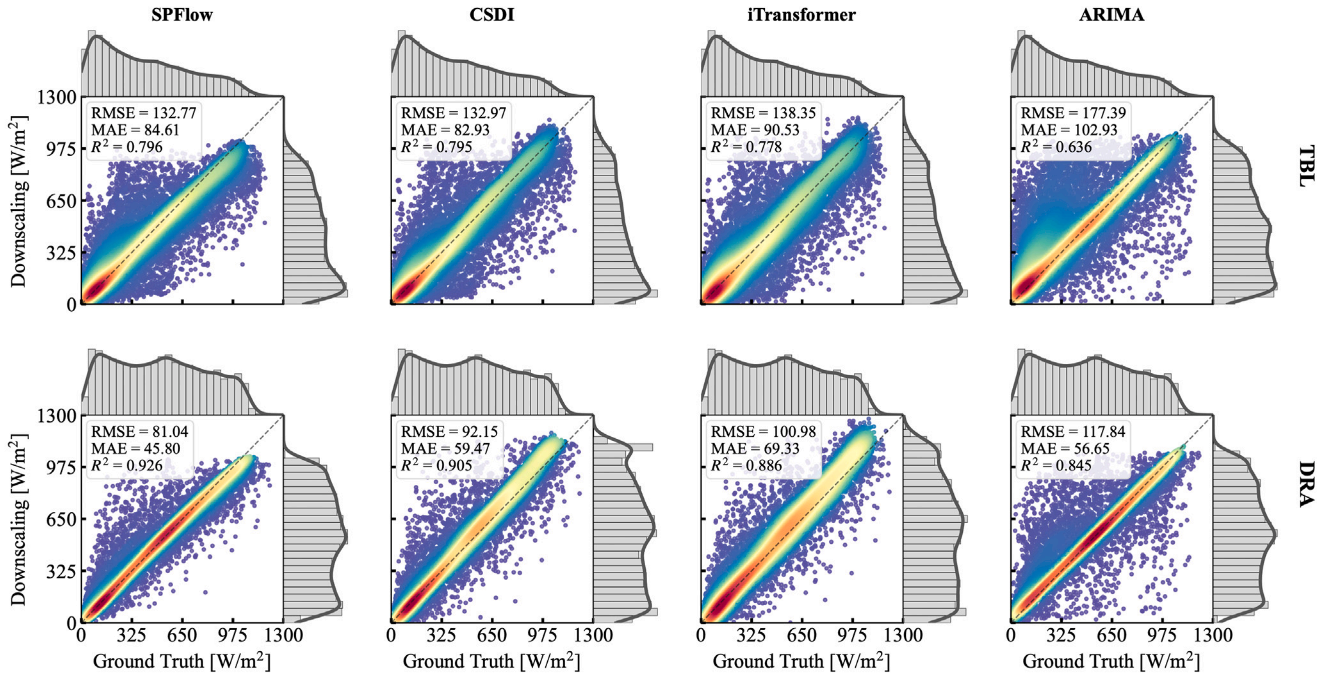


Fig. 14. Joint and marginal distributions of measured and downscaled day-ahead NWP GHI prediction using SPFlow and three benchmarks when evaluated at TBL and DRA stations. The heatmap shows the two-dimensional kernel density estimates.

and HRRR data demonstrate that SPFlow consistently outperforms statistical, deterministic deep learning, and probabilistic generative baselines. It achieves the lowest CRPS across all test stations and improves CRPS by up to 10.9% relative to the strongest probabilistic baseline. The model also produces prediction intervals that are close to the nominal coverage level while remaining substantially sharper than conventional probabilistic baselines. Ablation studies confirm that the learned shape prior and Soft-DTW training both contribute to the forecasting improvement, while spectral diagnostics show that SPFlow better preserves sub-hourly ramp-related high-frequency components. Zero-shot experiments further demonstrate its transferability to unseen locations without site-specific retraining. Overall, SPFlow provides an interpretable and operationally relevant solution for converting coarse day-ahead NWP solar forecasts into sub-hourly probabilistic forecasts. Its traceable ODE-based refinement process and empirically reliable uncertainty estimates make it suitable for renewable energy applications requiring reliable short-interval solar information, including PV operation, ramp-risk assessment, and storage scheduling.

CRediT authorship contribution statement

Tao Jing: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Conceptualization. **Shanlin Chen:** Writing – review & editing, Visualization, Validation. **Jinqing Peng:** Writing – review & editing, Validation. **Mengying Li:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Funding acquisition.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used GPT-5 to improve language and readability. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is substantially supported by the Innovation and Technology Commission (Project No. ITS/188/23) of the Hong Kong Special Administrative Region, China, and Shanghai Yi Xu Guan Technology Ltd. Co. (Project No. C2026J008).

Appendix. Detailed descriptions of benchmark models

To provide a rigorous evaluation of SPFlow, we curated a comprehensive benchmark comprising classical statistical approaches, deterministic deep learning architectures, and state-of-the-art generative frameworks. The selected models are widely recognized as standard baselines in solar energy forecasting and time-series generation.

- **Smart Persistence (SP):** Smart Persistence is the standard minimum-skill benchmark in solar irradiance forecasting [63]. Unlike naive persistence, which simply repeats the last observed value, SP scales the most recent GHI observation by the ratio of clear-sky irradiance between the target and observation times, i.e., $\hat{G}(t+k) = G(t) \cdot \frac{G_{cs}(t+k)}{G_{cs}(t)}$. This implicitly accounts for the deterministic diurnal cycle, making it a stringent baseline that any proposed model must surpass to demonstrate genuine forecasting skill.
- **AutoRegressive Integrated Moving Average (ARIMA):** ARIMA is adopted as a classical linear forecasting baseline [4,63]. For each forecast instance, it is fitted exclusively to historical 15-min GHI values available before the forecast origin and recursively generates the complete 96-step day-ahead trajectory. No target-period observation is used during model fitting or inference.

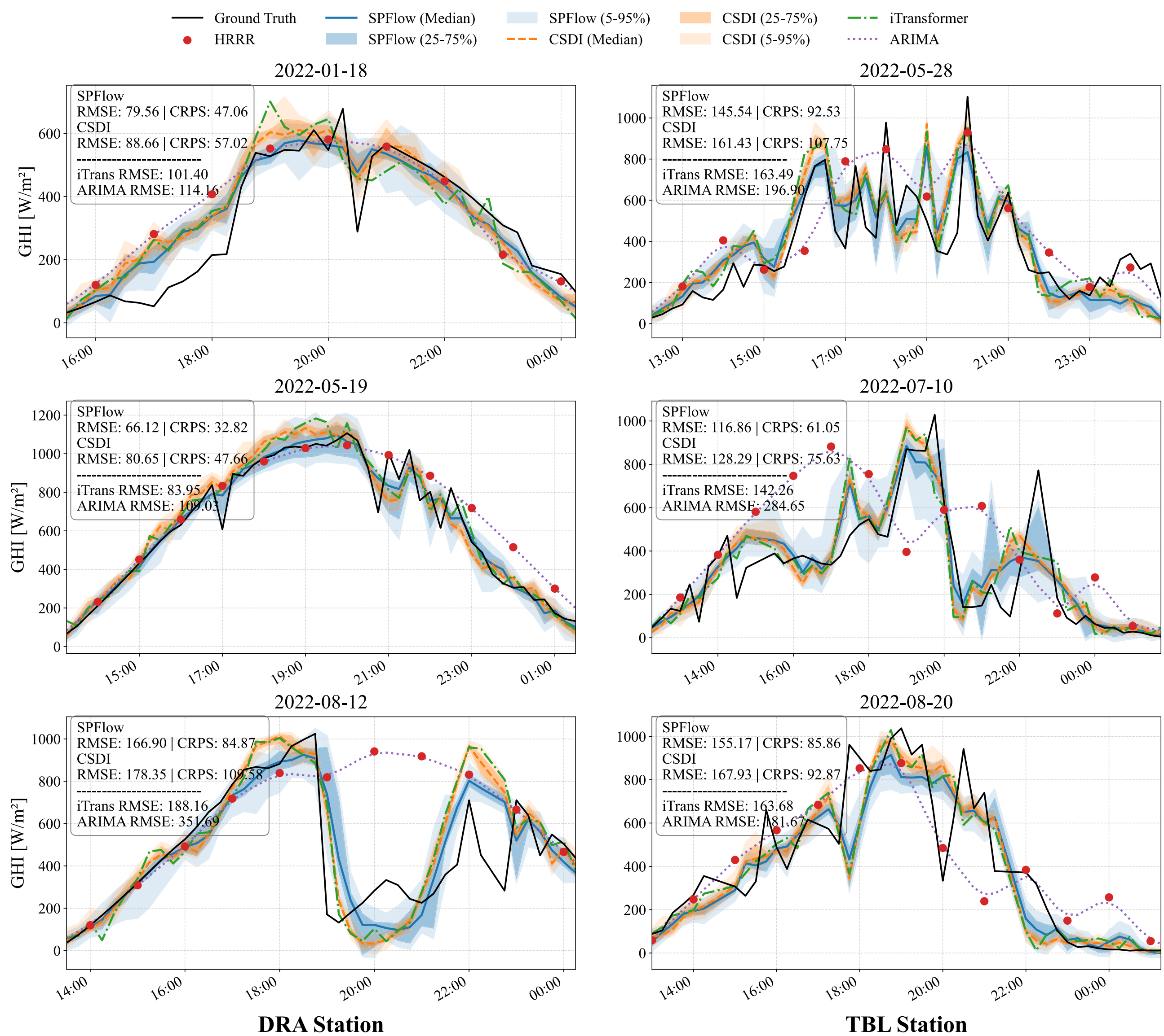


Fig. 15. Visual comparison of NWP downscaling performance across two zero-shot stations during three randomly selected time periods. The plots display the ground truth against probabilistic models (SPFlow in blue, CSDI in orange) and deterministic baselines (iTransformer in green, ARIMA in purple). For probabilistic models, the solid lines represent the median forecast, while the dark and light shaded regions correspond to the 50% (25th–75th quantiles) and 90% (5th–95th quantiles) prediction intervals, respectively. Evaluation metrics (RMSE and CRPS) are annotated within each panel.

- **Gaussian Processes (GP):** GP is widely adopted in solar resource assessment for their ability to perform probabilistic downscaling. By modeling the target variable as a distribution rather than a point estimate, GP provides essential uncertainty quantification. It is often employed in downscaling studies where data scarcity is a concern, serving here as a benchmark for probabilistic calibration and reliability [64]. In our implementation, a Matérn-5/2 kernel is adopted with automatic relevance determination (ARD), and hyperparameters are optimized by maximizing the marginal log-likelihood. The input features are identical to those used by the deep-learning baselines to ensure a fair comparison.
- **Long Short-Term Memory (LSTM):** As a representative of recurrent neural networks, LSTM is extensively utilized in solar irradiance forecasting due to its capacity to model long-term temporal dependencies [4,65]. In this study, it serves as a deterministic deep learning baseline.
- **iTransformer:** Representing a strong Transformer-based benchmark for multivariate time-series forecasting, iTransformer has demonstrated superior performance in capturing variable-wise dependencies [57]. Although originally developed for forecasting tasks, it is included here to benchmark the capability of modern deterministic architectures in reconstructing high-frequency irradiance dynamics.
- **Conditional Score-based Diffusion Imputation (CSDI):** CSDI is selected as the primary generative benchmark [29]. Among diffusion-based time-series models, it remains one of the most widely adopted conditional generation frameworks for probabilistic imputation and downscaling, making it an appropriate reference for evaluating SPFlow against diffusion-based probabilistic generation.
- **SPFlow-G (Standard Gaussian Prior):** This variant initializes the flow matching process with a standard Gaussian distribution and is used to evaluate the effectiveness of standard flow matching without informative prior guidance.
- **SPFlow-C (Clear-Sky Prior):** This variant replaces the learned shape prior with clear-sky GHI in the initial distribution. It is

introduced to quantify whether a physically informed but non-learned prior can provide benefits over Gaussian initialization, and to isolate the contribution of the proposed SPTransformer.

Data availability

The data used in this work are publicly available. SURFRAD measurements are available at https://gml.noaa.gov/aftp/data/radiation/surfrad/, NSRDB data are accessible via https://nsrdb.nrel.gov/data-viewer, and HRRR forecasts can be accessed at https://registry.opendata.aws/noaa-hrrr-pds/.

References

[1] Chen S, Liang Z, Guo S, Li M. Estimation of high-resolution solar irradiance data using optimized semi-empirical satellite method and GOES-16 imagery. Sol Energy 2022;241:404–15.

[2] Jing T, Chen S, Navarro-Alarcon D, Chu Y, Li M. SolarFusionNet: Enhanced solar irradiance forecasting via automated multi-modal feature selection and cross-modal fusion. IEEE Trans Sustain Energy 2025;16(2):761–73.

[3] Xie Y, Deng N, Jing T, Li M. Real-time estimation of cloud properties and surface downwelling longwave radiation from satellite imagery: A two-stage surrogate modeling approach. Appl Energy 2026;423:128373.

[4] Chu Y, Wang Y, Yang D, Chen S, Li M. A review of distributed solar forecasting with remote sensing and deep learning. Renew Sustain Energy Rev 2024;198:114391.

[5] Chen S, Jing T, Yang D, Chen Y, Li M, Chu Y. The role of spatio-temporal satellite-derived products and numerical weather forecasts in data-driven intra-day solar forecasting. Energy AI 2026;100718.

[6] Mayer MJ, Yang D. Calibration of deterministic NWP forecasts and its impact on verification. Int J Forecast 2023;39(2):981–91.

[7] Chen S, Li C, Stull R, Li M. Improved satellite-based intra-day solar forecasting with a chain of deep learning models. Energy Convers Manage 2024;313:118598.

[8] Omoyele O, Matrone S, Hoffmann M, Ogliari E, Weinand JM, Leva S, Stolten D. Impact of temporal resolution on the design and reliability of residential energy systems. Energy Build 2024;319:114411.

[9] Bistline JET. The importance of temporal resolution in modeling deep decarbonization of the electric power sector. Environ Res Lett 2021;16(8):084005.

[10] Mayer MJ. Effects of the meteorological data resolution and aggregation on the optimal design of photovoltaic power plants. Energy Convers Manage 2021;241:114313.

[11] Frimane Â, Soubdhan T, Bright JM, Aggour M. Nonparametric Bayesian-based recognition of solar irradiance conditions: Application to the generation of high temporal resolution synthetic solar irradiance data. Sol Energy 2019;182:462–79.

[12] Buster G, Rossol M, Maclaurin G, Xie Y, Sengupta M. A physical downscaling algorithm for the generation of high-resolution spatiotemporal solar irradiance data. Sol Energy 2021;216:508–17.

[13] Zhang W, Kleiber W, Florita AR, Hodge B-M, Mather B. A stochastic downscaling approach for generating high-frequency solar irradiance scenarios. Sol Energy 2018;176:370–9.

[14] Omoyele O, Hoffmann M, Weinand JM, Larrañeta M, Linßen J, Stolten D. A high-resolution downscaling approach for solar irradiance using statistical parameter matching. Renew Energy 2026;256:124551.

[15] Ngoko B, Sugihara H, Funaki T. Synthetic generation of high temporal resolution solar radiation data using Markov models. Sol Energy 2014;103:160–70.

[16] Bright JM. The impact of globally diverse GHI training data: Evaluation through application of a simple Markov chain downscaling methodology. J Renew Sustain Energy 2019;11(2).

[17] McDowell T, Kummert M. Estimating sub-hourly solar radiation and effective sky temperature from hourly weather data. In: ASHRAE & IBPSA-USA simBuild conference 2016, vol. 7, ASHRAE/IBPSA-USA; 2016, p. 423–30.

[18] Frimane A, Bright JM, Yang D, Ouhammou B, Aggour M. Dirichlet downscaling model for synthetic solar irradiance time series. J Renew Sustain Energy 2020;12(6).

[19] Zhang W, Kleiber W, Hodge B-M, Mather B. A nonstationary and non-Gaussian moving average model for solar irradiance. Environmetrics 2022;33(3):e2712.

[20] Peruchena CF, Blanco M, Bernardos A. Generation of series of high frequency DNI years consistent with annual and monthly long-term averages using measured DNI data. Energy Procedia 2014;49:2321–9.

[21] Larrañeta M, Fernandez-Peruchena C, Silva-Pérez MA, Lillo-Bravo I. Methodology to synthetically downscale DNI time series from 1-h to 1-min temporal resolution with geographic flexibility. Sol Energy 2018;162:573–84.

[22] Jing T, Chen S, Fang J, Porte-Agel F, Lam MC, Li M. Weakly informative priors flow matching for probabilistic regional solar irradiance forecasting. IEEE Trans Sustain Energy 2026;1–14.

[23] Schreck S, Schroedter-Homscheidt M, Klein M, Cao K-K. Satellite image-based generation of high frequency solar radiation time series for the assessment of solar energy systems. Meteorol Z 2020.

[24] Tang R, Dore J, Ma J, Leong PH. Interpolating high granularity solar generation and load consumption data using super resolution generative adversarial network. Appl Energy 2021;299:117297.

[25] Liu H, Zhang Z. Probabilistic forecasting of multiple plant day-ahead renewable power generation sequences with data privacy preserving. Energy AI 2025;19:100463.

[26] Song Z, Xiao F, Chen Z, Madsen H. Probabilistic ultra-short-term solar photovoltaic power forecasting using natural gradient boosting with attention-enhanced neural networks. Energy AI 2025;20:100496.

[27] Kosmopoulos P, Dhake H, Gakou E, Melita N, Kashyap Y. Similarity-based solar power forecasting: A technical advance for global energy systems. Energy AI 2026;24:100761.

[28] Rasul K, Seward C, Schuster I, Vollgraf R. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In: Proceedings of the 38th international conference on machine learning, vol. 139, PMLR; 2021, p. 8857–68.

[29] Tashiro Y, Song J, Song Y, Ermon S. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. Adv Neural Inf Process Syst 2021;34:24804–16.

[30] Shen L, Kwok J. Non-autoregressive conditional diffusion models for time series prediction. In: Proceedings of the 40th international conference on machine learning, vol. 202, PMLR; 2023, p. 31016–29.

[31] Lipman Y, Chen RT, Ben-Hamu H, Nickel M, Le M. Flow matching for generative modeling. 2022, arXiv preprint arXiv:2210.02747.

[32] Kollovieh M, Lienen M, Lüdke D, Schwinn L, Günnemann S. Flow matching with Gaussian process priors for probabilistic time series forecasting. 2024, arXiv preprint arXiv:2410.03024.

[33] Jing T, Chen S, Du T, Li M. Prior knowledge-guided residual diffusion model for enhanced regional-scale probabilistic solar irradiance forecasting. 2025, Available At SSRN 6113906.

[34] Beck HE, Zimmermann NE, McVicar TR, Vergopolan N, Berg A, Wood EF. Present and future Köppen-Geiger climate classification maps at 1-km resolution. Sci Data 2018;5(1):1–12.

[35] Augustine JA, DeLuisi JJ, Long CN. SURFRAD–A national surface radiation budget network for atmospheric research. Bull Am Meteorol Soc 2000;81(10):2341–58.

[36] Sengupta M, Xie Y, Lopez A, Habte A, Maclaurin G, Shelby J. The National Solar Radiation Database (NSRDB). Renew Sustain Energy Rev 2018;89:51–60.

[37] Chen S, Jing T, Li M, Lee HH, Bu S. Improved satellite-based estimation of direct normal solar irradiance using clearness index with deep learning methods. Sol Energy 2025;302:113991.

[38] Yang D, Bright JM. Worldwide validation of 8 satellite-derived and reanalysis solar radiation products: A preliminary evaluation and overall metrics for hourly data over 27 years. Sol Energy 2020;210:3–19.

[39] Yagli GM, Yang D, Gandhi O, Srinivasan D. Can we justify producing univariate machine-learning forecasts with satellite-derived solar irradiance? Appl Energy 2020;259:114122.

[40] Benjamin SG, Weygandt SS, Brown JM, Hu M, Alexander CR, Smirnova TG, Olson JB, James EP, Dowell DC, Grell GA, Lin H, Peckham SE, Smith TL, Moninger WR, Kenyon JS, Manikin GS. A North American hourly assimilation and model forecast cycle: The Rapid Refresh. Mon Weather Rev 2016;144(4):1669–94.

[41] Yang D, Li W, Yagli GM, Srinivasan D. Operational solar forecasting for grid integration: Standards, challenges, and outlook. Sol Energy 2021;224:930–7.

[42] Yang D. SolarData package update v1. 1: R functions for easy access of baseline surface radiation network (BSRN). Sol Energy 2019;188:970–5.

[43] Zhang G, Yang D, Galanis G, Androulakis E. Solar forecasting with hourly updated numerical weather prediction. Renew Sustain Energy Rev 2022;154:111768.

[44] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Adv Neural Inf Process Syst 2017;30.

[45] Chen S, Li M. Improved turbidity estimation from local meteorological data for solar resourcing and forecasting applications. Renew Energy 2022;189:259–72.

[46] Su J, Ahmed M, Lu Y, Pan S, Bo W, Liu Y. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing 2024;568:127063.

[47] Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable detr: Deformable transformers for end-to-end object detection. 2020, arXiv preprint arXiv:2010.04159.

[48] Shu Y, Lampos V. DeformTime: capturing variable dependencies with deformable attention for time series forecasting. 2024, arXiv preprint arXiv:2406.07438.

[49] Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, Wei Y. Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. 2017, p. 764–73.

[50] Chen C-FR, Fan Q, Panda R. Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 357–66.

[51] Gilleland E, Ahijevych D, Brown BG, Casati B, Ebert EE. Intercomparison of spatial forecast verification methods. Weather Forecast 2009;24(5):1416–30.

[52] Rahaman N, Barber A, Arpit D, Draxler F, Lin M, Hamprecht F, Bengio Y, Courville A. On the spectral bias of neural networks. In: International conference on machine learning. PMLR; 2019, p. 5301–10.

[53] Cuturi M, Blondel M. Soft-dtw: a differentiable loss function for time-series. In: International conference on machine learning. PMLR; 2017, p. 894–903.

[54] Tong A, Fatras K, Malkin N, Huguet G, Zhang Y, Rector-Brooks J, Wolf G, Bengio Y. Improving and generalizing flow-based generative models with minibatch optimal transport. 2023, arXiv preprint arXiv:2302.00482.

[55] Alcaraz JML, Strodthoff N. Diffusion-based time series imputation and forecasting with structured state space models. 2022, arXiv preprint arXiv:2208.09399.

[56] Huang G, Li Y, Pleiss G, Liu Z, Hopcroft JE, Weinberger KQ. Snapshot ensembles: Train 1, get m for free. 2017, arXiv preprint arXiv:1704.00109.

[57] Liu Y, Hu T, Zhang H, Wu H, Wang S, Ma L, Long M. itransformer: Inverted transformers are effective for time series forecasting. 2023, arXiv preprint arXiv:2310.06625.

[58] Yang D, Kleissl J. Solar irradiance and photovoltaic power forecasting. CRC Press; 2024.

[59] Gneiting T, Balabdaoui F, Raftery AE. Probabilistic forecasts, calibration and sharpness. J R Stat Soc Ser B Stat Methodol 2007;69(2):243–68.

[60] Yang D. Operational solar forecasting for the real-time market. Int J Forecast 2019;35(4):1499–519.

[61] Yang D. Making reference solar forecasts with climatology, persistence, and their optimal convex combination. Sol Energy 2020.

[62] Fotiadis S, Brenowitz N, Geffner T, Cohen Y, Pritchard M, Vahdat A, Mardani M. Stochastic flow matching for resolving small-scale physics. 2024, arXiv preprint arXiv:2410.19814.

[63] Yang D, Wang W, Gueymard CA, Hong T, Kleissl J, Huang J, Perez MJ, Perez R, Bright JM, Xia X, van der Meer D, Peters IM. A review of solar forecasting, its dependence on atmospheric sciences and implications for grid integration: Towards carbon neutrality. Renew Sustain Energy Rev 2022;161:112348.

[64] Van der Meer DW, Widén J, Munkhammar J. Review on probabilistic forecasting of photovoltaic power production and electricity consumption. Renew Sustain Energy Rev 2018;81:1484–512.

[65] Qing X, Niu Y. Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. Energy 2018;148:461–8.