Contents lists available at ScienceDirect

Applied Energy

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

Self-attention mechanism to enhance the generalizability of data-driven time-series prediction: A case study of intra-hour power forecasting of urban distributed photovoltaic systems

Hanxin Yu [a], Shanlin Chen [a], Yinghao Chu [a,*], Mengying Li [b], Yueming Ding [c], Rongxi Cui [c], Xin Zhao [d]

[a] Department of Systems Engineering, City University of Hong Kong, Hong Kong Special Administrative Region, China
[b] Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong Special Administrative Region, China
[c] State Grid Rizhao Power Supply Company, State Grid Shandong Electric Power Company, Rizhao, China
[d] Southern Institute of Industrial Technology (Shenzhen), Shenzhen, China

ARTICLE INFO

Keywords:
Model generalizability
Distributed photovoltaic
Solar forecast
Data-driven models
Attention mechanism

ABSTRACT

The emergence of small-scale urban distributed solar generation (DSG) has urged the exploration of site-adaptive forecasting models designed to accurately predict future power outputs for unseen DSGs. In such scenarios, with numerous DSGs spread across utility-scale cities and a lack of historical data, it is not economically viable to use conventional approaches that develop individual models for each DSG. Therefore, this work aims to tackle this real-world challenge by adapting the state-of-the-art, attention-based temporal fusion transformer (TFT) model to 188 real-world operational DSG data, thereby validating the generalizability of self-attention mechanism for multi-step time series forecasting. When adapted to unseen DSGs without training data, the experiment results demonstrate that the proposed solar TFT (STFT) improves by 11.07%, 17.58%, and 22.76% over the persistence model at the 10-, 20-, and 30-minute forecasts, respectively. Even when compared to representative deep-learning models, such as a long short-term memory model specialized in time series forecasting, STFT has demonstrated improved forecast accuracy, achieving 3.34%, 4.18%, and 5.85% enhancements at the 10-, 20-, and 30-minute forecast horizons, respectively. However, the model architecture of STFT is more complex, and the computational cost associated with it is relatively higher compared to other deep learning models. This trade-off between accuracy and computational efficiency should be considered in practical applications. The forecast performance is analyzed in three typical weather conditions, namely, clear, partly cloudy, and overcast. STFT demonstrates advantages in high variability periods, especially during weather transition periods, where reference models experience lagged predictions yielding relatively large errors.

1. Introduction

With the continuous development of both urbanization and low-carbon society trend, the increasingly limited availability of allocatable land has given rise to a massive number of distributed solar generations (DSGs). DSGs encompass both technical and environmental benefits, including enhanced power supply security and reduced fossil fuel costs [1]. Therefore, the DSG installation rate has recently outpaced that of centralized solar systems, especially in densely populated urban areas with scarce land availability [2–4]. As the penetration of photovoltaic (PV) power generation in the market continues to grow, the issue of the variable and stochastic nature of the power output from solar systems, attributed to the variability of solar irradiation, becomes increasingly apparent [5]. Accurate solar forecasting becomes increasingly essential to maintain stability in grid voltage and frequency, optimize the allocation of power production resources, and maximize economic benefits and resource utilization [6–8].

However, most forecasting models have been developed for centralized solar power plants, often catering to a limited number of specific locations. Their adaptability to new sites beyond the training set remains challenging [9]. Furthermore, even within the domain of distributed PV power forecasting, the focus tends to be on training and forecasting within the confines of individual DSGs with relatively large-scale capacities [10–12]. Consequently, the surge in the number of DSGs or distributed PV systems poses new challenges for solar

[*] Corresponding author.
E-mail address: yinghchu@cityu.edu.hk (Y. Chu).

https://doi.org/10.1016/j.apenergy.2024.124007
Received 29 January 2024; Received in revised form 19 June 2024; Accepted 22 July 2024
Available online 31 July 2024

power forecasting and integration techniques [13]. Firstly, the inherent decentralization in DSGs means that each site has unique system configurations and parameters. These variations include installed system size, operational parameters, panel orientation, panel tilt angle, local shading, etc. [14]. The unique configurations and parameters associated with each system and location render it difficult to construct a highly generalizable model. Furthermore, challenges arise from machine wear and degradation of solar instruments during usage, which can affect the consistency and effectiveness of forecasting models across different DSGs [15]. Compounding the issue, many DSGs lack dedicated personnel to record historical power outputs and meteorological data, further complicating the availability of reliable historical records [16]. Additionally, newly established DSGs often lack sufficient historical data, which introduces challenges for traditional machine or deep learning algorithms that require substantial past data for effective training before implementation [17]. This lack of appropriate historical data presents a distinct challenge for accurate forecasting of solar PV power generation, necessitating models with a high level of generalization. Addressing power variability and the individual characteristics of DSGs requires the development of a more generalizable solar power forecasting model capable of learning both shared information and the distinctive patterns of DSGs. Therefore, the self-attention mechanism is potential to develop a generalizable intra-hour solar PV power output forecasting model, addressing the challenge of accurately predicting new and unseen urban DSGs, even in the absence of historical data for the specific DSG.

The motivation for this work can be summarized as follows (exhibited in Fig. 1): To address the aforementioned challenges, this work integrates the attention-based temporal fusion transformer (TFT) [18] approach into the solar energy field, proposing the solar TFT (STFT) model tailored for PV power forecasting. STFT is capable of learning patterns separately for different types of features. Moreover, the attention mechanism within STFT facilitates the enhancement of generalizability. This addresses the challenges of generalization and improves the accuracy of solar PV power output forecasts in real-world scenarios. Consequently, this approach aids in conserving human effort and time that would otherwise be expended on manually recording solar power and related meteorological data. However, it is worth noting that due to its increased complexity compared to representative deep learning models, STFT requires more time for both training and inference. This aspect should be carefully considered.

The proposed approach involves training the model on data from surrounding DSGs with available historical records, rather than relying solely on available data from the same DSG. Fig. 1 visually illustrates the necessity of addressing the issues tackled by this work and depicts the industry scenario this work addresses. Fig. 1(a) demonstrates the variations in PV power output for DSGs located in different areas (depicted as DSG A, B, and C with distinct longitudes and latitudes). These variations can be significant, even when the clear sky global horizontal irradiance (GHI) is the same in different DSGs. This implies that neighboring DSGs, despite experiencing similar weather conditions, can exhibit substantial variations in PV power outputs. Consequently, the development of a highly generalizable model that takes into account diverse system characteristics becomes crucial. Fig. 1(b) demonstrates that the proposed approach of this work leverages historical data from different DSGs, achieving a highly generalizable model. This approach enables accurate forecasting for unseen DSGs. The main contributions of this work include but not limited to:

- *Exploring the potential of the self-attention mechanism to meet the demands of solar forecasting for urban distributed PV systems.* This work investigates the self-attention mechanisms within existing models and investigates their potential applicability in solar PV power forecasting scenarios. The experiment results demonstrate that the proposed approach exhibits significant generalization capabilities in comparison to other popular models.
- *Demonstrating that the STFT approach has shown enhanced accuracy and generalizability in intra-hour power forecasting for unseen systems.* Through experiments and validation using real operational data, it has been demonstrated that training models with data from surrounding DSGs and directly applying them to unseen systems yields promising results. This indicates strong practicality and competitiveness in the market. During the training process, this work maximizes the use of time-series data by incorporating the time feature itself as an input. The time feature, representing temporal variations and anticipated in the future, is combined with relevant meteorological data to investigate temporal patterns. When applying the model to unseen DSGs, including the time information enables the model to extract specific temporal patterns, facilitating accurate forecasting across different DSGs.
- *Developing an effective method to enhance forecasting performance under highly variable weather conditions.* In the domain of solar energy, researchers typically focus on the forecasting accuracy of the model under common weather conditions: clear, partly cloudy, and overcast. Most machine learning models, numerical weather prediction (NWP) models and persistence models can provide accurate prediction under clear and overcast conditions since these two weather types are relatively stable and predictable. However, when it comes to the weather with high variability, such as partly cloudy conditions, forecasting accuracy of these traditional models tend to be lower. After learning patterns from multiple DSGs, STFT model can accurately forecast PV power even in these challenging conditions.
- *Validating advanced methods using data from real-world urban DSGs.* This work incorporates both real-world operational data and publicly available data. The utilization of real-world operational data, obtained directly from the functioning of distributed systems, ensures that the developed method performs effectively under operational conditions. Additionally, the inclusion of public data serves to further validate the models used in this work.

This work is organized as follows. Section 2 introduces preliminaries on solar PV power forecasting. Section 3 details the STFT model. Section 4 describes the experimental data and setup. Section 5 comprehensively analyzes the experimental results. The concluding remarks are given in Section 6.

2. Related work

Solar power, as an eco-friendly energy source, has the ability to produce a sufficient amount of electricity in an environmentally sustainable manner. Motivated by worldwide policies or incentives, the global solar capacity is growing rapidly to fight climate change, reduce pollution, and mitigate the dependence on traditional fossil fuels. For example, according to the International Energy Agency, the global PV capacity is expected to supply more than 20% of global electricity demand in 2050 [8]. However, solar generation is highly variable due to the complex atmospheric processes. It may impose severe challenges to the grid integration of this weather-dependent power.

Over the past several decades, solar PV power forecasting has advanced significantly, with applications in physical models [19–21], statistical models [22,23], and hybrid models that integrate different techniques [8,24,25]. The rise of AI, particularly in the field of deep learning [26], has generated notable interest within the research community. This surge in interest has led researchers to explore increasingly powerful data-driven models [27,28], as they strive to leverage the potential of AI and deep learning for various applications [29]. These models are typically built on convolutional neural networks (CNNs) [30], recurrent neural networks (RNNs) [31], and the specialized RNN called long short-term memory (LSTM) [32]. These methods have been proven to be highly effective for tackling PV power forecasting tasks across various time horizons [33], and they exhibit superior

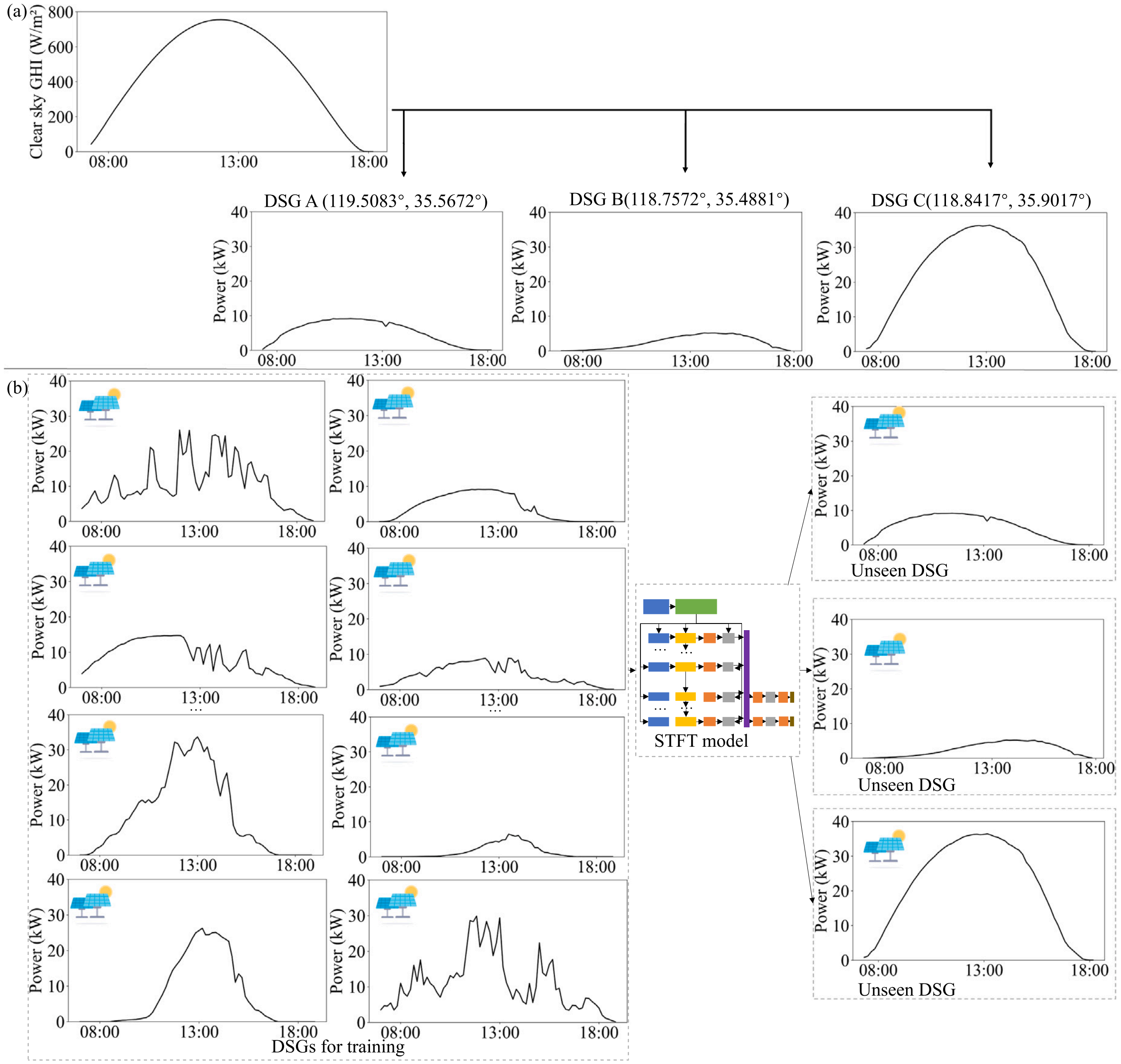

Fig. 1. (a) The challenge that this work aims to address. PV power outputs from different DSGs that are in proximity (depicted as DSG A, B, and C with distinct longitudes and latitudes) can show considerable variations. (b) This work aims to address the challenge of lacking historical data from specific distributed systems, and endeavors to develop a highly generalized forecasting model STFT, which is trained using data from multiple surrounding DSGs to predict solar PV power for newly installed DSGs.

performance in comparison with conventional physical models or traditional machine learning approaches. A more detailed review of solar forecasting methods can be found in [8,34].

More recently, transformers and attention mechanisms proposed by Vaswani et al. [35] have profoundly influenced the field of deep learning. Consequently, studies exploring the application of attention mechanisms in PV power forecasting have emerged as a popular research area. Numerous studies [36–39] have validated the effectiveness of utilizing attention mechanisms in PV power forecasting. Table 1 provides a summary of recent literature related to different methods of PV power forecasting.

In addition to delivering promising results in general time-series forecasting, the self-attention mechanism reduces dependence on external information and excels in characterizing internal correlations among input features. This enhances the generalizability of the model [40], helping to address the aforementioned problem. As stated by Zhao et al. [41] and Tian et al. [42], the self-attention mechanism enhances the capacity of the model to capture the global-level context, resulting in improved performance and generalization.

However, the self-attention mechanism alone encounters difficulties in handling features of different types, TFT model [18] emerges as a solution. TFT model, as an attention-based model, employs different temporal mechanisms for features with distinct characteristics and introduces specialized mechanisms for handling various features, including meteorological and time data. Then, López Santos et al. [43] introduced the adoption of TFT in PV power forecasting. TFT outperforms the results of several other methods, including auto-regressive integrated moving average, LSTM, MLP, and extreme gradient boosting (XGB), showcasing its potential in handling multivariate solar forecasting problems. Moreover, Mazen et al. [44] combined the gated recurrent unit with TFT to predict PV power generation, demonstrating the superior accuracy of their model compared to commonly used time-series algorithms and prediction models in the solar field. TFT has been proved to be a viable solution, even in the context of forecasting residential electricity consumption, particularly at the substation level [45]. This scenario shares comparable recurrent profiles and local influences with PV power generation in urban settings, further highlighting the effectiveness of TFT. However, it is important to note that the available work did not consider the prediction of power generation in distributed PV systems under the constraint of limited historical data. Therefore, there is still a need for comprehensive evidence to assess

Table 1
Summary of recent literature related to different methods for PV power forecasting.

Category	Method	Horizon/Resolution	Input
Physical model	Physical model chains [19]	Day-ahead, intra-day/15-min	NWP data and weather data
	Physical environmental parameter prediction model [21]	Day-ahead/Hourly	Geographical data, meteorological data and circuit data
Statistical model	Motion estimation model [11]	30-min-ahead/30-min	PV power data
	Auto-regressive integrated moving average [22]	Day-ahead/Daily	PV power data
	Caputo derivative [23]	1-, 5-, and 10-min-ahead/1-min	PV power data
Hybrid model	Hybrid gated recurrent unit (GRU) model [10]	Day-ahead/5-min	PV power data and weather data
	Hybrid wavelet packet decomposition and LSTM model [24]	1-h-ahead/5-min	PV power data and weather data
	Hybrid salp swarm algorithm, RNN and LSTM model [25]	1-h-ahead/5-min	Weather data
Deep learning CNN-based model	CNN and variational mode decomposition model [33]	1 to 3-h-ahead/Hourly	PV power data and weather data
	Multi-column CNN model [46]	2-h-ahead/5-min	PV power data, satellite image and weather data
Deep learning RNN-based model	LSTM and GRU model [32]	1 to 5-h-ahead/Hourly	PV power data, seasonal data and weather data
	RNN-based multi-task learning model [31]	30-min to 7-h-ahead/30-min	PV power data
Deep learning attention-based model	Interpretable temporal-spatial graph attention model [12]	4 to 6-h-ahead/15-min	PV power data, geographical data and clear-sky irradiance
	Attention-based multi-task learning model [36]	1-h-ahead/Hourly	PV power data and weather data
	Convolutional and channel attention-based model [37]	30-min-ahead/30-min	PV power data and weather data
	Sequence to sequence and attention-based model [38]	1-h-ahead/Hourly	PV power data, NWP data and weather data
	LSTM and self-attention based model [39]	24-h-ahead/Hourly	PV power data, weather data and weather forecast data

Note: Weather data comprises ambient temperature, atmospheric pressure, solar irradiation, solar radiation, elevation angle, ambient temperature, daily rainfall, wind speed, wind direction and relative humidity.

the effectiveness of TFT in addressing the generalization issue across multiple systems within distributed PV systems.

3. Methodology

3.1. Problem formulation

The objective of this work is to formulate a generalizable solar PV power forecasting model. In the training phase, let the total DSGs be X, the ratio of the training set to the whole set be ρ, and denote the training set as X_{train} and the testing set as X_{test}:

$$\begin{aligned} X &= \{X_1, X_2, \ldots, X_m\}, \\ X_{\text{train}} &= \{X_i\} \quad \text{with ratio } \rho, \\ X_{\text{test}} &= X - X_{\text{train}}, \end{aligned} \tag{1}$$

where m represents the number of DSGs. Each X_i has its own unique data distribution. Therefore, a model that fits X_i effectively may not be suitable for X_j, where $i \neq j$. For specific DSG X_l at the time t_1 to t_n in the training set:

$$X_l = \begin{bmatrix} x_1^l(t_1) & \ldots & x_1^l(t_n) \\ \ldots & \ldots & \ldots \\ x_k^l(t_1) & \ldots & x_k^l(t_n) \end{bmatrix}, \tag{2}$$

where k denotes the number of input dimensions. In the first row, x_1^l corresponds to the PV power, while the subsequent entries represent meteorological data, including clear sky GHI, clear sky direct normal irradiance (DNI), clear sky diffuse horizontal irradiance (DHI) and clear sky beam horizontal irradiance (BHI). n stands for the number of time instances. The model f can simultaneously output PV forecasts:

$$\begin{bmatrix} \hat{x}_1^l(t_{n+1}) & \ldots & \hat{x}_1^l(t_{n+j}) \end{bmatrix} = f\left(\begin{bmatrix} x_1^l(t_1) & \ldots & x_1^l(t_n) \\ \ldots & \ldots & \ldots \\ x_k^l(t_1) & \ldots & x_k^l(t_n) \end{bmatrix} \right), \tag{3}$$

where j represents j-step-ahead forecast. The training objective is to minimize the error between the observed value (ground truth) and the predicted result:

$$\epsilon_T = \frac{1}{L} \sum_{l=1}^{L} \left| \hat{x}_1^l(t_{n+1}) - x_1^l(t_{n+1}), \ldots, \hat{x}_1^l(t_{n+j}) - x_1^l(t_{n+j}) \right|. \tag{4}$$

The specific error definition formula may be different depending on the chosen metric. After the training, f will be applied on the test set X_{test} for inference. The ultimate goal of the model is to minimize prediction errors on the testing set:

$$\epsilon_R = \left| \hat{x}_1^{test}(t_{n+1}) - x_1^{test}(t_{n+1}), \ldots, \hat{x}_1^{test}(t_{n+j}) - x_1^{test}(t_{n+j}) \right|. \tag{5}$$

For simplicity, the observed value will be expresses as $y(t_{n+1})$ to $y(t_{n+j})$ to replace the expression $x_1(t_{n+1})$ and $x_1(t_{n+j})$. In this work, for each sequence instance, the generation occurs in a sliding window manner with n set to 6 as the sequence input, k set to 5 as the number of inputs, and j set to 3, covering 30-minute-ahead multi-step PV power forecasts from t_1 to t_3 with a step size of 10 min.

3.2. STFT for solar forecasting scenario

This study leverages TFT [18] as its base model and introduces STFT to generate accurate multi-horizon forecasts for solar PV power. As a model specialized in the field of solar energy, the architecture of STFT employed in this study is depicted in Fig. 2. Considering the diverse nature of real-world solar time series data, STFT is designed to handle multiple data sources in solar power forecast simultaneously, ranging from time-dependent to stationary, past-observed, or known future, and utilizes various techniques to effectively capture temporal dynamics. Specifically, in Fig. 2, it can be observed that the nameplate parameters of DSGs serve as static data, DSG power is denoted as both past-observed and target data, while clear sky GHI, clear sky DNI, clear sky DHI, and clear sky BHI are also considered as past-observed data. Additionally, calendar date serves as known future data, as it can be retrieved in advance. The data undergo

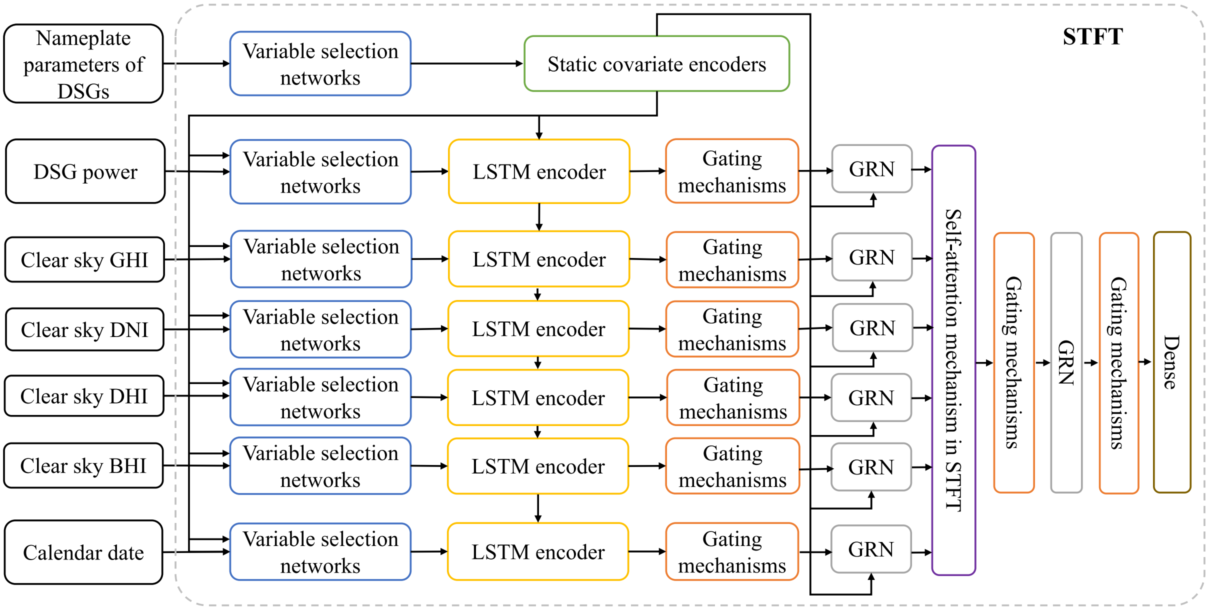

Fig. 2. Model architecture of STFT. It consists of variable selection networks, static covariate encoders, LSTM encoders, gating mechanisms, GRNs, self-attention mechanisms. STFT selects relevant variables, captures temporal dependencies, applies self-attention, and produces final predictions.

variable selection networks to identify relevant features, ensuring the recognition of the most correlated variables on a per-instance basis. Moreover, variable selection networks effectively remove redundant features during training, resulting in the optimal selection of clear sky meteorological data. Static covariate encoders seamlessly integrate static features into temporal dynamics for static data. In this study, the nameplate parameters of DSGs are considered as categorical static identifiers in scenarios involving multiple DSGs. When handling unseen DSGs, STFT has the ability to capture the most similar temporal context patterns from the existing DSGs, thereby enabling accurate forecasting. For the remaining data, a well-defined structure with key components, including LSTM encoder, gating mechanisms, and gated residual network (GRN), is employed for flexible model component selection, enabling the detection of the more relevant meteorological information at specific past time frames. With the clear sky historical meteorological data, especially GHI and DNI, gating mechanisms selectively chooses the relevant historical features. Simultaneously, it filters out irrelevant clear sky historical meteorological data to predict the PV power at a specific time. The self-attention mechanism effectively captures long-term dependencies and enhances generalizability. Lastly, the inclusion of the quantile loss function enhances the robustness of STFT against outliers. Given the presence of extreme weather conditions, the quantile loss function aids in reducing sensitivity to extreme weather outliers by providing a probabilistic forecast of a quantile of PV power. Further details about gating mechanisms, variable selection networks, static covariate encoders and the corresponding quantile loss can be found in Appendix A.

STFT adopts and modifies the self-attention mechanism [35] to capture long-term relationships in the context of multi-horizon forecasting scenarios. Given that historical information can be viewed as a sequence, STFT leverages attention mechanisms to capture dependencies across various time steps, effectively learning and integrating temporal patterns. When deployed in a new and unseen DSG, which may exhibit distinct characteristics compared to the initially trained DSG, the attention mechanism aids the model in uncovering relevant temporal relationships specific to the new DSG. This enables the model to apply these insights to make accurate predictions. The attention mechanism consists of values $\boldsymbol{V} \in \mathbb{R}^{N\times d_V}$, keys $\boldsymbol{K} \in \mathbb{R}^{N\times d_{attn}}$, and queries $\boldsymbol{Q} \in \mathbb{R}^{N\times d_{attn}}$. The scaled dot-product attention method is as follows:

$$\mathrm{Attention}(Q,K,V) = \mathrm{Softmax}(Q\cdot K^T/\sqrt{d_{attn}})\cdot V. \tag{6}$$

To capture the multiple patterns simultaneously, multi-head attention is proposed by Vaswani et al. [35]:

$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1,\dots,\mathrm{head}_h)\cdot W, \tag{7}$$

where $\mathrm{head}_i = \mathrm{Attention}(Q\cdot W_i^Q, K\cdot W_i^K, V\cdot W_i^V)$ and weights are head-specific. To assess the importance of each feature, STFT adjusts the original multi-head attention mechanism and incorporates additive aggregation of all heads to create a shared head:

$$\mathrm{STFTHead}(Q,K,V) = \widetilde{head}\cdot W, \tag{8}$$

where $\widetilde{head} = \frac{1}{H}\sum_{i=1}^{h}\mathrm{Attention}(Q\cdot W_i^Q, K\cdot W_i^K, V\cdot W^V)$.

Generalization in unseen DSG. Time-related features can be broadly categorized into short-term dependency features and long-term dependency features. STFT leverages LSTM for local enhancement, capturing specific and distinguishable information for each DSG. As for the long-term features, which encompass the overall generalized characteristics of multiple DSGs, they are considered complementary. The inclusion of the attention mechanism aids in capturing long-term temporal patterns, thereby enhancing the generalizability. By incorporating this attention mechanism, the model can selectively focus on the most informative aspects of past meteorological data and power generation data, particularly in the context of long-term dependencies.

4. Experiment, data and setup

The methodology applied in this study encompasses the following steps. Firstly, data pre-processing including data cleaning and normalization is performed to convert raw data into a format that is more suitable for modeling. Subsequently, during the model training phase, STFT is trained and the hyperparameters are tuned. Finally, the results are evaluated under two scenarios. The overall flowchart is depicted in Fig. 3, while the procedures are summarized in Table 2.

4.1. Experiment data

The experiment utilizes data from 188 DSGs located in the Rizhao area, with the central coordinates at 35.5°N, 119.2°E, in Shandong province, China. Fig. 4 illustrates the position of the Rizhao area within the global PV power potential map and the distribution of investigated

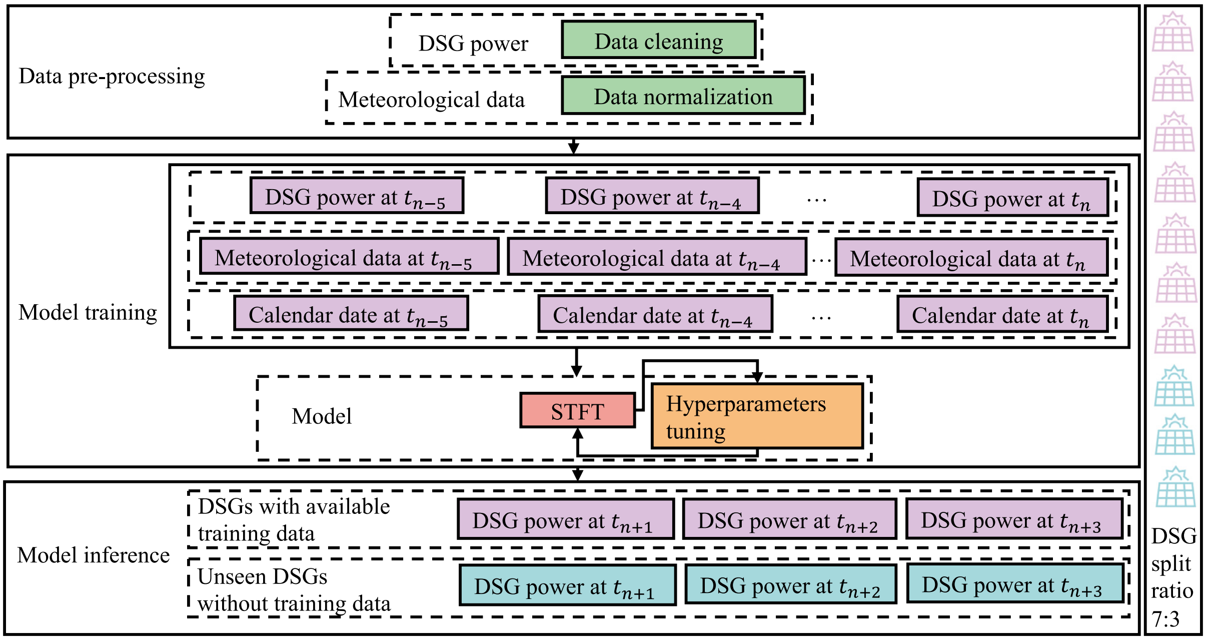

Fig. 3. The overall flowchart of STFT development and inference. DSGs in purple represents 70% of randomly selected DSGs and are utilized for both training and inference purposes. The remaining 30% of DSGs, marked in blue, represent unseen DSGs and are exclusively used for inference purposes. The specific details of the dataset split method can be found in Section 5.

Table 2
STFT development and inference procedures. It includes data pre-processing, model training, and inference.

Step	Procedure	Details
1	Data pre-processing	After splitting the multiple DSGs dataset, data preprocessing involves data cleaning on the power data and min–max normalization of the meteorological data (see Section 4.2).
2	Model training	STFT is constructed using historical DSG power, meteorological data, and the calendar date as inputs, with the future multi-horizon DSG power as the output. The detailed methodology can be found in Section 3.2. The quantile loss function is defined, and its mathematical expression can be found in Appendix A.4. STFT is trained using the AdamW optimizer [47]. The hyperparameters are tuned using Optuna [48] to find suitable combinations (see Appendix C).
3	Model inference	After the training phase, STFT is utilized for inference in two scenarios: DSGs with available training data and unseen DSGs without any prior training data. The details of the evaluation metrics can be found in Section 4.3, while the results are elaborated upon in Section 5.

DSGs in Rizhao. The PV power potential of the Rizhao area is in the middle level on a global scale. Only with proper management and accurate forecasting, Rizhao can reap significant benefits from PV power generation. The dataset covers the period from January 1, 2020, to December 31, 2020. The data consists of DSG power, clear sky GHI, clear sky DNI, clear sky DHI, clear sky BHI and date, all with 10-min temporal resolution. This work is focused on newly established urban distributed PV installations. These PV installations often lack historical data or have limited records but require accurate predictions. Additionally, each PV installation has unique characteristics. Taking all these factors into account, the aim of this work is to develop a highly generalizable model for distributed PV systems within a specific region. This enables the swift deployment of newly distributed PV installations, even in situations where training data is scarce during the initial deployment phase. To maintain the focus on the generalizability of STFT and minimize interference from other factors, this work excludes the incorporation of certain elements such as NWP and external inputs like satellite imagery or weather data. Future research will explore these aspects in more depth. By prioritizing the development of a robust and generalizable model for distributed PV systems, this work lays the groundwork for more effective deployment strategies in urban areas.

Meteorological data. For predicting PV power output across a multi-step horizon, it is essential to take into account specific relevant meteorological features. Among these features, clear sky GHI, DNI, DHI, and BHI from McClear model [50] are used. McClear [50] model is a widely-used physical solar radiation model that estimates clear sky solar radiation under various atmospheric and geographical parameters. Clear sky GHI represents the total available solar energy received on a horizontal surface in clear sky condition. Clear sky DNI refers to the solar radiation that comes directly from the disk of the sun. It represents the intensity of solar radiation in a beam that strikes a surface normal to the sun, directly affecting the electrical output of solar panels. Higher clear sky DNI generally leads to increased PV power production, especially in conditions of ample direct sunlight. Clear sky DHI represents the solar radiation that reaches a horizontal surface after scattering during the clear sky weather. DHI can impact PV power as it contributes to a portion of the radiation received by PV panels. Clear sky BHI refers to the direct solar radiation incident on a horizontal surface in clear sky conditions. Clear sky BHI is essential for evaluating the energy output of PV systems when the panels are tilted to follow the path of the sun, not when they are oriented horizontally.

DSG power. Given the objective of predicting future PV power in each DSG, it is imperative to incorporate past PV power. The DSG power in this work refers to the solar power generated by each DSG, serving

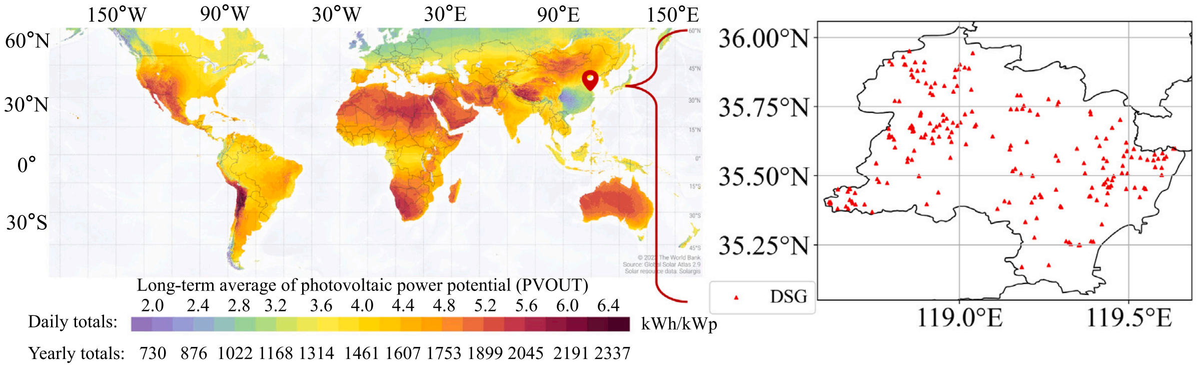

Fig. 4. The Rizhao area, situated at 35.5°N, 119.2°E, ranks in the midrange for PV power potential globally [49]. In Rizhao, DSGs are utilized and distributed across the entire area.

as both a key feature and the primary focus of research. As detailed in Section 3.1, n is set as the number of past PV power outputs to be included in each instance, recognizing its significant impact on PV power forecasts for subsequent time steps.

Calendar date. The time series dataset includes specific calendar date information, encompassing the month, day, and minute as known future, time-varying features. This inclusion aims to provide calendar information for predictions in unseen DSGs, as solar energy exhibits periodic patterns and date information can serve as a valuable reference.

4.2. Data pre-processing

Data pre-processing plays an important role in PV power forecasting. It is defined as the transformation of raw data into a form suitable for modeling. In this work, the following pre-processing tasks are included: data cleaning and data normalization.

Data cleaning. Data points with a solar zenith angle exceeding 80° are excluded to mitigate the high airmass effect. This exclusion is based on the understanding that such angles typically correspond to nighttime or periods when the sun is situated significantly below the horizon, leading to no sunlight available for solar power generation and resulting in DSG power and other relevant meteorological data being close to zero.

Data normalization. For the collected data, separate normalization is performed for each DSG. The meteorological data is normalized using the min–max normalization method. The min–max normalization method [51] has been widely employed in literature focused on solar forecasting, thus equally weighting different types of data with different magnitude [46]. In the training set:

$$x_e^l\hat{(t_n)} = \frac{x_e^l(t_n) - \min(x_e^l)}{\max(x_e^l) - \min(x_e^l)}, \tag{9}$$

where $x_e^l(t_n)$ denotes the irradiance value for a specific meteorological feature e at the time t_n in the DSG x^l, $\min(x_e^l)$ and $\max(x_e^l)$ denote the minimum and maximum irradiance values for certain feature e within the range 2 to k in the certain DSG.

4.3. Experiment setup

Model training environment. DSG power outputs are forecasted using STFT, and the model performance is compared to that of reference models, including the persistence model, MLR, MLP, LSTM, GRU, XGB and gradient boosting regression (GBR). The persistence model serves as a fundamental approach, assuming that PV power for the next time period remains unchanged. Moving on to MLR, it functions as a linear predictor, capturing relationships between historical PV powers and future PV powers through least squares estimation. MLP, on the other hand, is a feedforward neural network leveraging nonlinear activation functions and learning optimal weights to transform historical powers into predictions for future time steps. LSTM, a specialized type of RNN, excels in processing sequential data and is generally effective for multi-step PV forecasting tasks. GRU is also a specialized RNN, but it has a simpler structure and lower computational burden compared to LSTM. Additionally, XGB stands out as an implementation of gradient boosted trees, iteratively focusing on errors and utilizing gradient descent for tuning. Meanwhile, GBR sequentially combines weak learners, often decision trees, to minimize the loss function and provide accurate PV power predictions. Further details can be found in Appendix B. STFT model is trained using Pytorch Forecasting [52]. For specific details on STFT hyperparameter settings, please refer to Appendix C. The other models utilize various libraries and packages, such as scikit-learn [53] for MLR, basic Pytorch [54] framework for LSTM, GRU and MLP, and XGB's [55] own package for XGB.

Evaluation metrics. Two statistical metrics for the models' error are used to assess their performance as recommended in Yang et al. [56] and Chu et al. [57], namely root mean square error (RMSE) and forecast skill. The persistence model [58] in Appendix B.1 is used as a benchmark for all other models.

- RMSE, which measures the average square error in the forecast, with smaller RMSE values indicating higher prediction accuracy:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y(t_n) - \hat{y}(t_n))^2} \quad (\mathrm{kW}). \tag{10}$$

- Forecast skill, which measures the improvement of the investigated forecast over the reference persistence model:

$$s = 1 - \frac{\mathrm{RMSE}}{\mathrm{RMSE_P}}, \tag{11}$$

where s denotes the forecast skill and P denotes the persistence model.

5. Results and discussion

The models are evaluated in two scenarios: DSGs with available training data and unseen DSGs without training data. The dataset is divided into three sets. Initially, 70% of the DSGs are randomly selected to form a primary subset. The training set is then constructed using the data spanning the first three weeks of each month throughout the year within this subset of DSGs. Subsequently, the data from the remaining weeks of each month is employed for model assessment when training data is available, as discussed in Section 5.1. The remaining DSGs are kept as an independent set for assessing unseen DSGs, as detailed in Section 5.2.

Table 3
Forecasting performance when training data is available. STFT model is the top-performing model with the smallest RMSE values and the largest s values.

Step	Metric	STFT	MLR	MLP	LSTM	GRU	XGB	GBR	Persistence
10-min	RMSE (kW)	0.064	0.101	0.066	0.065	0.065	0.073	0.069	0.071
	s	10.14%	−43.28%	6.67%	8.46%	8.29%	−2.68%	2.53%	–
20-min	RMSE (kW)	0.081	0.146	0.083	0.082	0.083	0.099	0.093	0.094
	s	14.08%	−55.06%	11.54%	12.51%	11.82%	−5.13%	1.31%	–
30-min	RMSE (kW)	0.089	0.187	0.095	0.094	0.094	0.126	0.110	0.111
	s	19.79%	−68.19%	15.13%	15.41%	15.35%	−13.28%	1.03%	–

5.1. Model assessment when training data is available

Table 3 displays an evaluation of the model's performance when training data is available. Both the training and assessment data are derived from the same DSGs. To make a general performance comparison, the RMSEs of all models are calculated for each DSG. Subsequently, the average RMSE and the corresponding forecast skills are obtained.

Based on the results in Table 3, the improvement in the forecast skill of all models compared to benchmark models ranges from −43.28% to 10.14% in the 10-min forecast, from −55.06% to 14.08% in the 20-min forecast, and from −68.19% to 19.79% in the 30-min forecast. Additionally, it is notable that the RMSE of all models tends to increase as the forecast time extends, which is logical because as the horizon increases, the models face greater challenges in predicting the future trends of PV power. Delving deeper into model performance, the top-performing model, STFT, exhibits an improvement in forecast skill as the prediction horizon extends, peaking at 19.79% in the 30-min forecast. This consistent outperformance of the deep learning models, including STFT, MLP, LSTM and GRU, over the benchmark models is a noteworthy trend. Note that the GRU model and LSTM model may appear to have the same RMSE values in 10- and 30-min forecast when rounding the experimental results to three decimal places. However, upon calculating the results using a more precise forecast skill evaluation, it is observed that LSTM performs slightly better than GRU. Among the machine learning models in the reference set, only GBR shows relatively good performance. This superior performance can be attributed to the adaptability and capability of deep learning models to learn complex, non-linear relationships.

The model assessment when the training data is available reveals that all deep learning models exhibit strong performance. Notably, STFT outperforms other models across all horizons, showcasing its proficiency in capturing long-term temporal patterns through the self-attention mechanism and utilizing LSTM for local enhancement [18]. Moreover, as the static covariate encoders treat the DSG name as a categorical static identifier, STFT can achieve more precise forecasts when abundant historical data from a specific DSG is available.

5.2. Models assessment on unseen DSGs

5.2.1. Experimental results

In this section, the ultimate goal is to evaluate the models in the new and unseen DSGs so that the generalizability and robustness of the models can be assessed. The evaluation of the all models' performance for unseen DSGs is presented in Table 4. The results indicate that the STFT model outperforms all other models in 10-, 20-, and 30-min forecasts. Specifically, RMSE and forecast skill exhibit similar trends across all models. The improvement in the forecast skill of all models compared to benchmark models ranges −38.43% to 11.07% in 10-min forecast, from −49.79% to 17.58% in 20-min forecast, and from −63.41% to 22.76% in 30-min forecast. The top-performing model, STFT, reveals a 22.76% higher forecast skill than the benchmark model in the 30-min forecast. As the forecast horizon increases, the forecasting accuracy of STFT improves. Additionally, LSTM and GRU achieve their highest forecast skills of 16.91% and 16.82%, respectively, in the 30-min forecast for unseen DSGs.

While the forecasting accuracy of the persistence model substantially declines with the increase in forecast horizon, deep learning models such as STFT, LSTM, GRU, and MLP show significant improvements LSTM, GRU and MLP show less effectiveness with available training data, whereas STFT, known for its high generalizability, demonstrates superior forecasting skills. For LSTM and GRU, a similar phenomenon to model assessment when training data is available also arises here. Specifically, these two models demonstrate very similar performance, indicated by identical RMSE values across 10-, 20-, and 30-min forecasts. Nonetheless, LSTM demonstrates marginally superior forecasting accuracy in terms of forecast skill. In addition, the forecast capabilities of traditional machine learning models are found to be inferior to those of the persistence model. This may be explained by the fact that traditional deep learning and machine learning models have difficulty generalizing effectively when there are changes to the DSG, environment, and location. They rely on learning patterns from the seen data. For instance, Srivastava and Lessmann [59] have proved that LSTM is hardly generalizable. Additionally, due to limitations in complexity, traditional machine learning models only have a limited capacity to capture complex relationships within multiple DSGs. When applied to unseen DSGs, the relationships they capture are often inaccurate. However, for STFT, the gating mechanisms and variable selection network extract relevant historical meteorological and power data while suppressing unnecessary ones. The static covariate encoders handle static known future date information separately. This allows STFT to capture periodic information and make predictions based on the known date for unseen DSGs. In addition, the self-attention mechanism, as explored by Niu et al. [40], delves into long-term dependencies and weighs the importance of different parts of the input sequence. This exploration is aimed at enhancing the generalizability of the model when making predictions in multiple DSGs. When data from unseen DSGs is introduced, the global context in the self-attention mechanism enhances the generalizability of the model. To further evaluate the performance of STFT, the evaluation based on the public dataset has also been explored. The results show that STFT performs better than the benchmarks, which again confirms the higher generalizability of STFT, more details are presented in Appendix D.

5.2.2. Comparison between top performers

This section provides a detailed comparison of the top two performing models, STFT and LSTM. Fig. 5, as the Tukey box [60], provides a concise summary of key statistical properties for RMSE results, including the median, quartiles, and potential outliers, for STFT, LSTM and persistence model. Across all forecast horizons, similar to the results shown in Table 4, STFT outperforms the other two models with the lowest median RMSE. Additionally, STFT demonstrates more consistency, as all predicted values fall within a certain range without any extreme outliers. When the results are compared across all three time horizons, it is clear that the advantages of STFT grow more prominent as the forecast horizon increases. For the 10-min forecast, the prediction results of STFT closely resemble those of LSTM, with the sole difference being that LSTM predicts some values with considerably larger deviations, whereas STFT has fewer of such deviations. The wider box in the persistence model indicates that its predicted values display greater variability compared to the observed results. Meanwhile, for the 20-

Table 4
Forecasting performance when evaluated using data of unseen DSGs. In each category, the smallest RMSE values and the biggest s values are in bold to highlight best performance. STFT model outperforms all other models in 10-, 20-, and 30-min forecasts, exhibiting both the smallest RMSE values and the largest s values.

Step	Metric	STFT	MLR	MLP	LSTM	GRU	XGB	GBR	Persistence
10-min	RMSE (kW)	**0.066**	0.103	0.069	0.069	0.069	0.079	0.073	0.075
	s	**11.07%**	−38.43%	7.59%	7.73%	7.60%	−6.65%	1.73%	–
20-min	RMSE (kW)	**0.081**	0.147	0.085	0.085	0.085	0.103	0.097	0.098
	s	**17.58%**	−49.79%	13.05%	13.40%	13.38%	−4.55%	1.49%	–
30-min	RMSE (kW)	**0.089**	0.187	0.096	0.095	0.095	0.131	0.113	0.115
	s	**22.76%**	−63.41%	16.45%	16.91%	16.82%	−14.09%	−14.09%	–

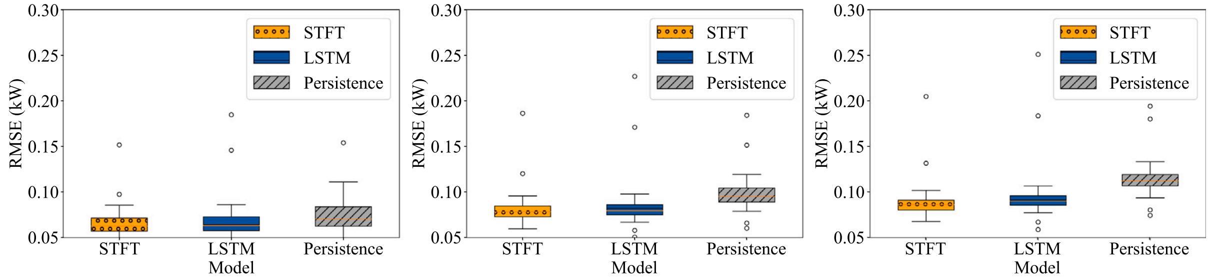

Fig. 5. Tukey box plot for visualization for 10-(left), 20-(middle) and 30-min(right) forecasts of three selected models: STFT, LSTM, and reference persistence model. Across all intra-hour forecast horizons, STFT outperforms the other two models with the lowest Q1, Q2, and Q3 RMSE.

and 30-min forecasts, RMSE values of STFT for Q1, Q2 (median), and Q3 are all smaller than those from the other two models, highlighting its significant advantage in long-term forecasting.

To enable a comprehensive examination of their forecasting capabilities, a series of illustrative figures are presented following the recommendation of Murphy and Winkler [61]. The joint and marginal distributions for different forecast steps are depicted in Fig. 6, illustrating the performance of STFT, LSTM, and the persistence model. On the top and right margins of the sub-figures, the marginal distributions are demonstrated in the form of histograms.

To scrutinize the distribution of predictions from both models more thoroughly, the joint distributions of the observed and predicted power are explored across intervals of 10-, 20-, and 30-min for STFT, LSTM, and the persistence model. For joint distributions, a closer clustering of points along the 45-degree line (the line of perfect agreement between predictions and actual values) suggests better forecast accuracy. Overall, STFT consistently demonstrates superior forecasting accuracy with denser clusters and a more controlled spread compared to LSTM and the persistence model. Specifically, regarding the marginal distribution, the forecasted PV power outputs of these three models exhibit a similar skewed distribution. This suggests that all three models maintain a certain degree of forecasting accuracy for unseen DSGs. However, when considering the joint distribution, STFT shows a higher concentration of points along the diagonal, indicating a stronger alignment between forecasted values and measurements. The forecasted results of LSTM tend to be smaller than those of STFT and the persistence model when the PV power is high (greater than 1.0). The possible reason could be that high PV power generation may indicate significant fluctuations in weather conditions, and LSTM may encounter difficulties to capture these fluctuations, this underprediction is more obvious when LSTM is applied to unseen DSGs without historical training data. It has been proven that LSTM struggles to accurately capture underlying patterns due to their limited memory learned from other DSG [62]. This highlights the limitations of LSTM and reveals its dependency on both the quality and quantity of training data [63].

For the 10-min forecast, STFT and LSTM exhibit a dense clustering of data points along the diagonal, signifying commendable forecasting accuracy. Points in the persistence model are more dispersed, indicating lower accuracy than that of the deep learning models. This dispersion may be due to inherent characteristics or limitations of the persistence model in capturing temporal patterns within dynamic systems. Moving to the 20-min forecast, all models demonstrate a wider spread of data points from the diagonal. STFT's spread increases compared to the 10-min forecast but still maintains a tighter cluster than the other two models. There is a noticeable increase in the dispersion of points away from the diagonal in the LSTM plot, surpassing that observed in the STFT plot. This suggests that the accuracy of the LSTM model may diminish more rapidly than that of the STFT model as the forecast interval increases. The persistence model exhibits the widest spread, indicating significant variability in its predictions. By comparing the 30-min forecast, this challenge intensifies, with all models' scatter points displaying an even broader divergence from the diagonal, emphasizing the increasing unpredictability with longer forecast durations. The marginal histograms suggest that while forecasts are generally normally distributed, there is an increasing spread and potential skewness as the forecast horizon extends, reflecting the inherent increase in variability and prediction challenge over longer timescales.

Some examples to show the time series PV power predictions with error distributions and accumulations are presented as follows. Fig. 7 presents the results for 10-, 20-, and 30-min horizons on clear days. Fig. 8 showcases the comparison for the same horizons during partly cloudy weather conditions, while Fig. 9 illustrates results under overcast weather conditions. The timestamps have been adjusted to the local time for enhanced clarity.

A clear period is defined as a time when clouds do not obscure the sun, resulting in high PV power as DSGs absorb a significant amount of solar energy. On a clear day, PV power variability is minimal, and all three models provide accurate forecasts. The 10-min forecast in Fig. 7 reveals an impressive synchronization among all models with actual measurements, illustrating a smooth parabolic curve peaking around midday. This harmony extends to the 20- and 30-min forecasts, although minor deviations become evident, especially in the predictions of LSTM that slightly deviate from the peak measurements. Among all three models, regardless of the time horizon, STFT consistently exhibits the smallest accumulated error, indicating that STFT possesses the best forecast capability. As the sun sets, the PV power decreases, a change that makes it challenging for the models to predict. Both the persistence model and LSTM exhibit some lag and, as a result, perform less effectively in these conditions. STFT outperforms the others and delivers the most accurate forecasts of future PV power. This can be

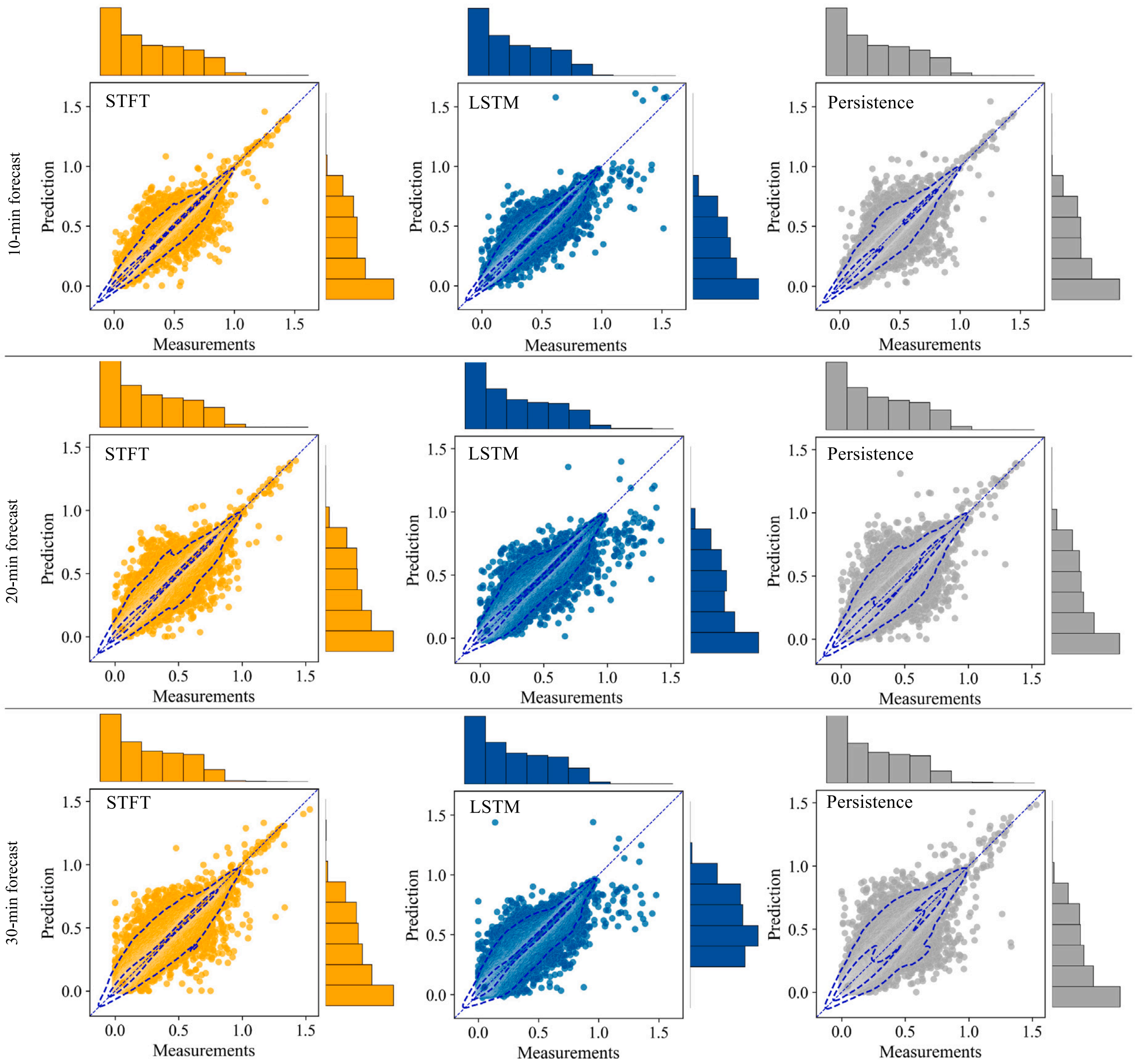

Fig. 6. Joint and marginal distribution of measured (x-axis) and predicted (y-axis) PV power for 10-, 20-, and 30-min ahead horizons. This work randomly selects 1% of the results for visual clarity. Across all intra-hour forecast horizons, STFT consistently demonstrates superior forecasting accuracy with denser clusters and a more controlled spread compared to LSTM and the persistence model.

attributed to STFT learning general characteristics from multiple DSGs, where it recognized the time-varying pattern of PV power under clear-sky conditions since PV power does not vary much under clear sky conditions and applied it to this unseen DSG.

Partly cloudy conditions typically involve a mix of thick and thin clouds, leading to the high variability in PV power outputs. On the example day in Fig. 8, significant fluctuations in PV power output are observed, particularly around 9:00 AM, 1:00 PM, and 2:30 PM. For the 10-min forecast, all models attempt to capture the intermittent variability. However, due to the inherent time lag characteristic in the persistence model, it consistently struggles to provide relatively accurate predictions. LSTM, while identifying trends, introduces relatively large errors compared to actual results, indicating a substantial disparity in specific data. In contrast, STFT demonstrates a more adaptive response to the fluctuating power levels when compared to the benchmark persistence model and LSTM. The 20- and 30-min forecasts continue this trend, with an increase in model disparity, which highlights the challenges in predicting such volatile conditions. STFT, in particular, seems to offer a smoother representation, possibly indicating a more versatile prediction approach. In comparison to clear conditions, the performance of LSTM is not as robust when dealing with minor weather variability. It is noticeable that, as the fluctuations persist, the gap between LSTM and the benchmark model gradually narrows. However, as the forecast horizon increases, the disparity between the predicted results of all models and the actual results also gradually increases, as can be observed in the accumulated error graph.

Overcast conditions are defined as a period when the sun is obscured by clouds, and the total sky cloud coverage exceeds 90%. Therefore, on such a day, solar energy is weak, resulting in overall low PV power outputs. In this gloomier scenario depicted in Fig. 9, the forecasting accuracy of all three models remain considerably low. It can be observed that LSTM consistently produces results higher than the actual values, regardless of the forecast horizon. This can be attributed to the challenge of predicting the fluctuating PV power due to its low values. LSTM can only forecast new results rely on historical learned data, making it challenging to provide accurate predictions for rare or new situations, such as less common overcast days. In contrast, the persistence model and STFT perform better in the face of continuous fluctuations, such as the 10-min forecast around 2:00 PM. In this case, STFT can offer a reasonably accurate prediction. However, as the forecast horizon increases, STFT also struggles to predict accurately when the sun is completely obscured. With the self-attention mechanism,

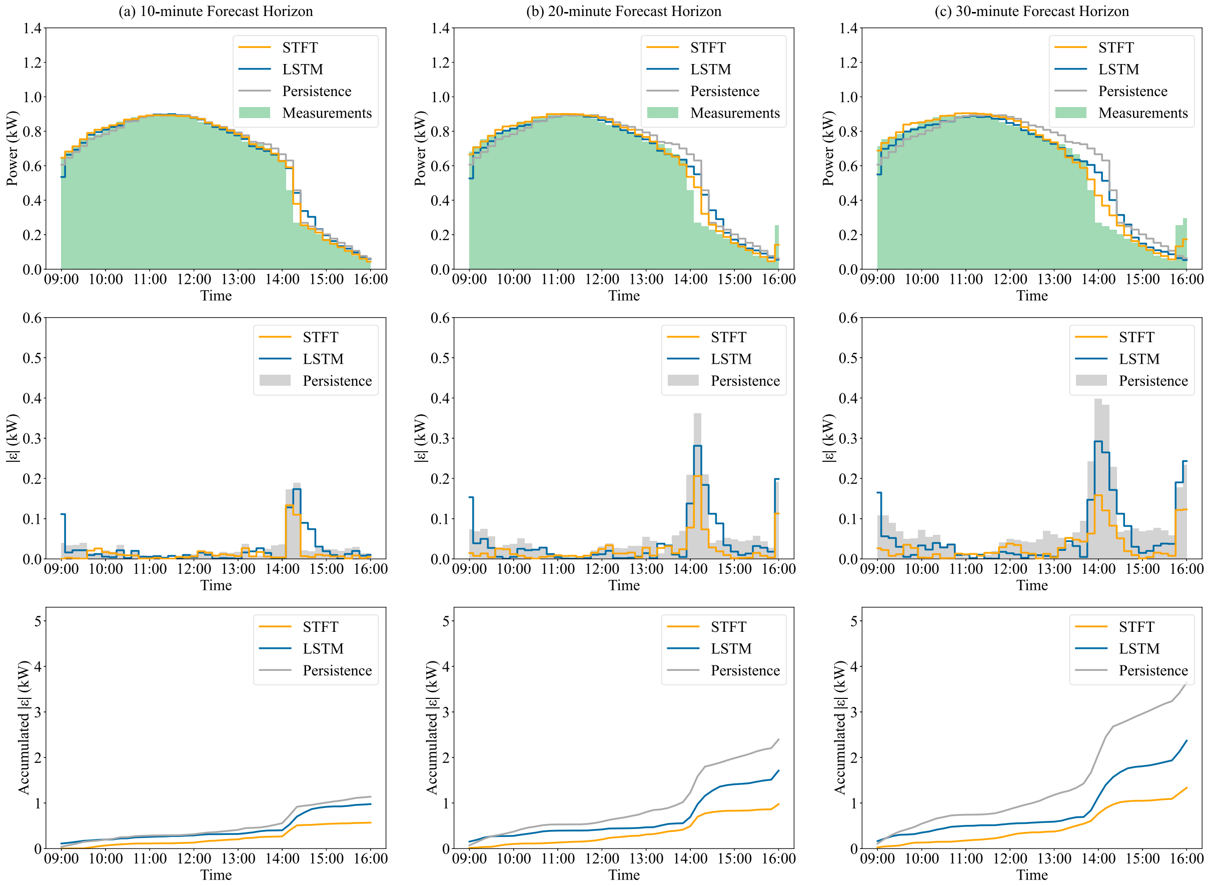

Fig. 7. Sample predictions, absolute error, and accumulated error time series of STFT, LSTM, and the persistence model on a representative clear day (2020-01-08), which has a weather transition in the afternoon. (a), (b), and (c) columns represent forecasts for 10-, 20-, and 30-min horizon, respectively. When there is a weather transition, STFT adapts more quickly and exhibits smaller errors, resulting in a larger disparity in accumulated errors compared to the clear day.

time periods more relevant to past data are better retained, which can help improve predictions for STFT. Overall, STFT still provides better prediction results compared to the other two models.

Across all environmental conditions, STFT consistently shows a strong alignment with actual measurements, demonstrating its reliable and highly accurate forecasting capability in various weather conditions. Furthermore, its ability to generalize and provide substantial assistance in unseen DSGs is evident. On the other hand, LSTM appears to struggle with the high variability and less common scenarios. The persistence model tends to provide a simplified and linear perspective, reflecting its inherent modeling characteristics and limitations in adaptability.

5.2.3. Computational cost

All the experiments are implemented on the platform MAC OS Apple M1 Pro chip 16 GB RAM. Table 5 depicts the computational cost of the proposed and reference models in this work. Among the models used, STFT stands out as the most complex model, requiring more time for both training and inference. As shown in Table 5, STFT takes approximately 0.144 s per data instance, making it well-suited for short-term forecasting with a 10-min temporal resolution. However, when considering the use of STFT for forecasting tasks, it is important to consider the computational cost. STFT is the most sophisticated model compared to other alternatives like LSTM, resulting in significantly longer training and inference times. Specifically, the training time for STFT is roughly 12.93 times longer than LSTM, and the inference time is approximately 20.57 times longer. This increased time requirement may pose challenges, particularly in scenarios where real-time or near-real-time predictions are crucial. Nevertheless, the training and inferencing computational cost could be further reduced using high-performance computers to meet the requirement of practical applications.

Table 5
The computational cost of neural network models. STFT is the most complex model among those used, but it requires only approximately 0.144 s per data instance for inference.

Model	Training time per epoch (s)	Inference time per data instance (s)
STFT	2928.00	0.144
MLP	22.56	0.004
LSTM	226.38	0.007
GRU	191.60	0.006

5.3. Limitations and directions for future research

Considering the high computational cost and model complexity associated with STFT, this work acknowledges these limitations and offers potential solutions to address them. Furthermore, to enhance the transparency and generalizability of the model, future research directions are outlined accordingly.

Model computational efficiency and complexity. The strength of STFT lies in its ability to learn distinct patterns from different types of features and uncover generalizable patterns for unseen DSGs. However, the adoption of STFT significantly increases computational requirements and the model complexity compared to other representative deep learning models. Table 5 quantifies the differences between STFT and other models. To enhance computational efficiency and reduce model

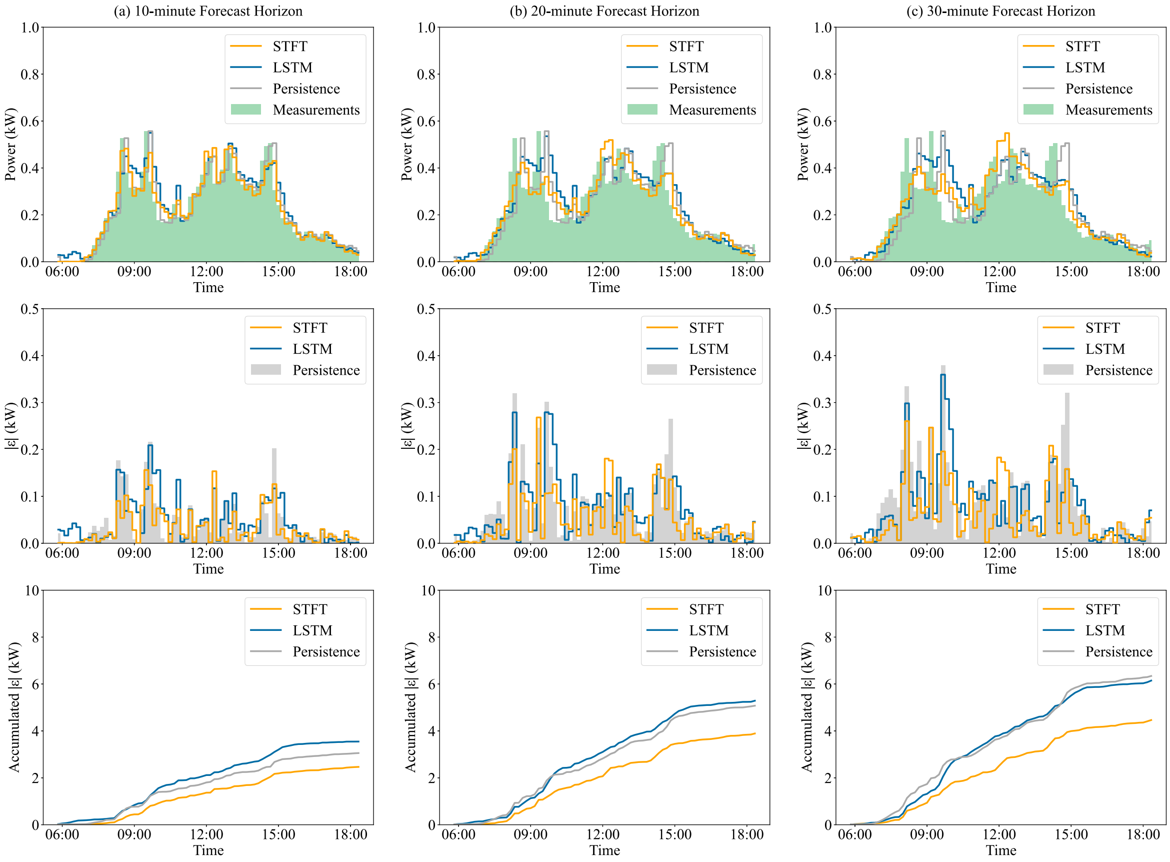

Fig. 8. Sample predictions, absolute error, and accumulated error time series of STFT, LSTM, and the persistence model on a typical partly cloudy day (2020-06-30). (a), (b), and (c) columns refer to forecasts for 10-, 20-, and 30-min horizon, respectively. The ground irradiance of this day exhibits high variability. When encountering high fluctuation, STFT exhibits the most adaptive response, with the slowest increase in accumulated errors, resulting in the best performance.

complexity, future research can consider adopting the following model compression techniques:

- Pruning: Pruning can be implemented to decrease the storage overhead of the model by removing the unimportant units using a suitable pruning strategy. As STFT is a transformer-based model, learnable pruning-related parameters can be used to adaptively adjust the depth and width of the transformer [64]. Following the pruning, fine-tuning can be conducted to restore the performance of the model.
- Precision truncation: Precision truncation can be applied during model training by converting the training precision from high-precision numbers to low-precision numbers [65], such as from 32-bit floating-point to 16-bit floating-point in STFT. This conversion improves efficiency by reducing the computational workload and memory requirements associated with precision truncation.
- Model quantization: Model quantization reduces the memory footprint and computational requirements of deep learning models. By quantizing the model parameters, such as weights and activations, from higher precision (e.g., 32-bit floating-point) to lower precision (e.g., 8-bit integers), the amount of memory needed to store STFT can be significantly reduced [66], leading to a substantial reduction in computational costs. The difference between precision truncation and model quantization lies in their respective focuses. Precision truncation primarily aims to reduce the precision of numerical values during model training, while model quantization primarily focuses on representing model parameters in a lower precision format.
- Distillation: By leveraging distillation technology [67], the model can be optimized by utilizing a teacher model to transfer knowledge to the student model. STFT, on the specific solar domain, can be adopted based on a pretrained teacher model from general tasks. The computational cost of STFT can be reduced by leveraging the knowledge embedded in the pretrained teacher model.

Exploring these approaches further holds promise as a potential avenue for future research.

Model interpretability. This study employs the self-attention mechanism to unveil temporal dependencies in distributed PV systems. The attention weights learned by the self-attention model signify the importance of each input token, contributing to enhance model interpretability. Future research can utilize the attention weights to improve the transparency of the model.

Incorporating more real-world data. To enhance the generalizability of the model, future research can incorporate a broader range of real-world data from diverse regions and climates. By utilizing data from multiple locations with distinct climatic characteristics, forecasting capabilities of the model to predict PV power in unseen DSG scenarios can be further solidified. This approach will open up avenues for deeper exploration and investigation in this field.

6. Conclusions

This work aims to develop intra-hour multi-step power forecasting methods for urban distributed solar generations, taking into account the distinct characteristics of different distributed systems. The proposed STFT involves utilizing an attention-based deep learning methods to achieve high generalizability for unseen DSGs. This work also explores the performance of seven reference models and compares them with

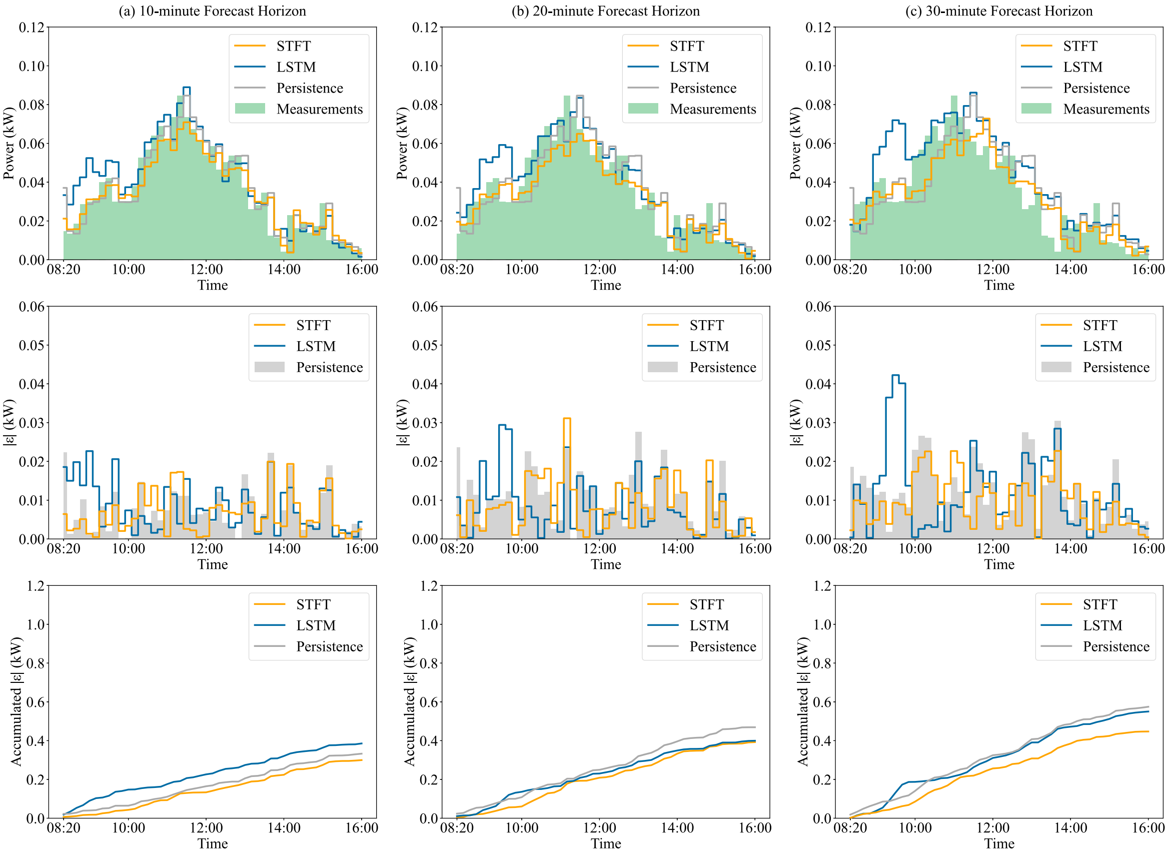

Fig. 9. Sample predictions, absolute error, and accumulated error time series of STFT, LSTM, and the persistence model were analyzed on a typical overcast day (2020-01-05) characterized by a low average PV power output. (a), (b), and (c) columns refer to 10-, 20-, and 30-min forecast horizon, respectively. All three models exhibit low forecasting accuracy. STFT provides the best prediction results compared to LSTM and the persistence model.

the proposed STFT model using data collected from 188 real-world DSG systems.

The experiment results reveal that the STFT model outperforms both traditional machine learning and deep learning models significantly when assessed using data of unseen distributed solar PV installations, achieving an average RMSE of 0.066 kW, 0.081 kW, and 0.089 kW for 10-, 20-, and 30-min forecasts, respectively, and exhibits a forecast skill improvement of 11.07%, 17.58%, and 22.76% against the persistence model. Compared to LSTM, which is specialized in time series forecasting, TFT demonstrates improvements of approximately 3.34%, 4.18%, and 5.85% at the 10-, 20-, and 30-min forecast horizons, respectively. These results consistently highlight the superior forecasting accuracy of STFT over conventional deep-learning methods, such as LSTM, across varied intra-hour forecasting horizons. The proposed STFT demonstrates both higher generalizability and accuracy when dealing with complex real-world environments, particularly for high-variability weather like partly cloudy and weather transition periods that are challenging for other models to predict. However, it is necessary to acknowledge that the architectural complexity of STFT leads to a comparatively high computational cost. Striking a balance between accuracy and computational efficiency becomes crucial in real-world application. In conclusion, this work proposes an attention-based model STFT to predict solar PV power for new distributed solar generations when historical data is unavailable.

CRediT authorship contribution statement

Hanxin Yu: Writing – original draft, Software, Methodology, Investigation, Data curation. **Shanlin Chen:** Writing – original draft, Validation, Software, Methodology. **Yinghao Chu:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **Mengying Li:** Writing – review & editing, Supervision, Resources, Conceptualization. **Yueming Ding:** Writing – review & editing, Supervision, Resources. **Rongxi Cui:** Writing – review & editing, Supervision, Resources. **Xin Zhao:** Writing – review & editing, Supervision, Investigation.

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.

Data availability

Data will be made available on request.

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

During the preparation of this work the authors used ChatGPT in order to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgments

This work is substantially supported by a grant from the Guangdong Basic and Applied Basic Research Foundation (Project No. 2024A1515010117) and a grant from Research Grants Council of Hong Kong (Project No. 21200424).

The authors also gratefully acknowledge the support and data provided by the State Grid Shandong Electric Power Company.

Appendix A. Module within STFT

In this section, this work provides detailed introductions to specific modules related to gating mechanisms, variable selection networks, static covariate encoders, and the quantile loss within STFT.

A.1. Gating mechanisms in time series solar forecasting

The gating mechanisms, serving as a fundamental component of STFT, aim to provide the model with the flexibility to apply non-linear processing, focusing on the relevant parts. In specific time series solar forecasting, gating mechanisms selectively chooses relevant historical features. During clear sky conditions, historical data—especially GHI and DNI—often has the most significant impact on PV power. Simultaneously, it filters out irrelevant past data to predict the PV power at a specific time.

The detailed implementation of the key gate, the gated residual network (GRN), is formulated:

$$\begin{aligned}\mathrm{GRN}_{\omega}(x_1^l(t_n), h^l(t_{n-1})) &= \mathrm{LayerNorm}(x_1^l(t_n) + \mathrm{GLU}_{\omega}(\eta_1)),\\ \eta_1 &= W_{1,\omega}\eta_2 + b_{1,\omega},\\ \eta_2 &= \mathrm{ELU}(W_{2,\omega}x_1^l(t_n) + W_{3,\omega}\cdot h^l(t_{n-1}) + b_{2,\omega}),\end{aligned} \tag{A.1}$$

where ELU refers to the Exponential Linear Unit activation function [68], $h^l(t_{n-1})$ represents the historical meteorological representation, specifically the PV power and clear sky irradiances in the DSG, at time t_{n-1}. If no historical PV power data or clear sky irradiances are available for the same DSG, they are considered as 0. LayerNorm is the layer normalization technique [69], $\eta_1 \in \mathbb{R}^{d_\mathrm{f}}$ and $\eta_2 \in \mathbb{R}^{d_\mathrm{f}}$ correspond to intermediate layers, while W and b represent the weights and biases. For the gated linear unit (GLU) [70]:

$$\mathrm{GLU}_{\omega}(\gamma) = \sigma(W_{4,\omega}\gamma + b_{4,\omega}) \odot (W_{5,\omega}\gamma + b_{5,\omega}), \tag{A.2}$$

where σ represents the Sigmoid activation function, $\odot$ denotes the element-wise Hadamard product, and W and b still refer to the weights and bias.

A.2. Variable selection networks in multivariate solar forecasting

Variable selection networks are an instance-wise selection mechanism in STFT. All static and time-dependent inputs use separate variable selection networks. It is applied on a per-instance basis, taking into account all past, present, and known future or unknown inputs. This approach allows the model to evaluate whether periodicity influences its predictions through known future variables, while also offering insights for removing the impact of redundant features. For instance, GHI, DNI, and DHI are highly correlated; hence, utilizing all three features simultaneously may lead to redundancy. The variable selection network aids in reducing redundancy per instance.

Variable selection networks consist of two main components: variable selection weights and processed variables. The variable selection weights are assigned:

$$v_{\chi}(t_n) = \mathrm{Softmax}\left(\mathrm{GRN}_{v_\chi}\left(P(t_n), c_s\right)\right), \tag{A.3}$$

where $P(t_n)$ represents the representations of all past meteorological information including DSG power at time t_n, and c_s is to serve as a temporal meteorological information selection using the static covariate encoder (see Appendix A.3). The Softmax function serves as the normalization. Overall, variable selection weights are used to select relevant values based on the time dimension. For each processed feature:

$$P_j\hat{(t_n)} = \mathrm{GRN}_{P_j}(P_j(t_n)), \tag{A.4}$$

where $P_j(t_n)$ represents the jth meteorological information at time t_n, which undergoes a transformation to meet the input requirement. Finally, the formula of variable selection networks obtained through the multiplication of two key elements is illustrated below:

$$P\hat{(t_n)} = \sum v_{\chi}(t_n) \cdot P_j\hat{(t_n)}. \tag{A.5}$$

A.3. Static covariate encoders in generalizing solar forecasting

STFT includes a specialized encoder designed specifically for static data. This is crucial because the encoder handles both time-dependent and time-independent features, and the static feature, being time-invariant, should be processed separately. In this work, a comprehensive model is trained by combining data from multiple DSGs. To predict PV power output within the same DSG, the system name serves as the categorical static variable. This static variable functions as a key feature for predictions within the same DSG, acting as an identifier that distinguishes data across various DSGs. Consequently, when predicting for an unseen DSG, the model endeavors to identify highly relevant temporal context patterns. Similar to time-dependent features, continuous and categorical data are also processed differently. Continuous data is transformed using linear transformations, whereas categorical data is transformed using entity embeddings [71]. As shown in Eq. (A.1), separate GRN encoders are utilized to produce a static context vector.

A.4. Loss function

Unlike traditional regression models, which usually employ loss functions focused on predicting the deterministic value of a variable given a feature vector, STFT utilizes quantile loss [72] as its loss function to provide a probabilistic forecast for a quantile of DSG power. During the training, the objective is to minimizing the summing quantile loss:

$$\begin{aligned}L(X_{\mathrm{train}}, W) &= \sum_{y(t_n)\in X_{\mathrm{train}}} \sum_{q\in Q} \sum_{j=1}^{j_{\max}} \frac{L_q(y(t_{n+j}), \hat{y}(q, t_{n+j}, j))}{M \cdot j_{\max}},\\ L_q(y, \hat{y}) &= \begin{cases}(q-1)|y-\hat{y}| & \text{if } y \le \hat{y}\\ q|y-\hat{y}| & \text{if } y > \hat{y},\end{cases}\end{aligned} \tag{A.6}$$

where W represents the weights of the STFT, j stands for the j-step ahead prediction of future DSG power, M denotes the number of samples in the training DSG X_{train}. Specifically, $\hat{y}(q, t_{n+j}, j)$ means the predicted DSG power at a specific quantile q, at a particular intra-hour horizon t_{n+j}. Given the potential for outliers due to extreme weather conditions or anomalies, the quantile loss function enhances the robustness of predictions by reducing sensitivity to such outliers in the PV power forecasting.

Appendix B. Selected data-driven algorithms for solar forecasting

Except for STFT in Section 3.2, this work employs seven reference models: the persistence model, multivariate regression (MLR), multilayer perceptron (MLP), long short-term memory (LSTM) networks, gated recurrent units (GRU), extreme gradient boosting (XGB), and gradient boosting regression (GBR). These models are utilized to assess and compare the performance of multi-step PV power forecasting.

B.1. Persistence model

The persistence model stands as the most elementary forecast model, yet its simplicity does not undermine its accuracy, especially when it comes to forecasting PV power. In this work, the persistence model is regarded as the benchmark comparing to other models. The persistence model assumes that the PV power remains constant between t_n and t_{n+j}, resulting in the forecasting:

$$\hat{y}(t_{n+j}) = y(t_n), \tag{B.1}$$

where j can be any forecast step.

B.2. Multivariate regression

Multivariate regression (MLR) captures the relationship between the input $\boldsymbol{x}$ and the output $\hat{\boldsymbol{y}}$ using linear predictor functions, and estimates unknown parameters through the least squares method [73]. It can be regarded as a single-layer MLP. The mathematical formula is:

$$\left[y^l(t_{n+1}) \quad \cdots \quad y^l(t_{n+j})\right] = \begin{bmatrix} x_1^l(t_1) & \ldots & x_1^l(t_n) \\ \ldots & \ldots & \ldots \\ x_k^l(t_1) & \ldots & x_k^l(t_n) \end{bmatrix} \cdot W + B, \tag{B.2}$$

where W denotes the weight and B denotes the bias.

B.3. Multilayer perceptron

Multilayer perceptron (MLP) is a class of feed-forward artificial neural networks that employ nonlinear and differentiable activation functions and consist of multiple layers, including at least one hidden layer that contains multiple neurons [74]. Neurons within each layer are interconnected by weights, forming a highly connected network structure. During the training, the input $x(t_n)$ is propagated through the MLP network, and the network learns the optimal weights by minimizing the error between the predicted output and the observed value. With a simple 2-layer neural network, the feed-forward operation transforms historical input data into predictions for future time steps:

$$\begin{aligned} y^{(1)} &= \sigma(W^{(1)} \cdot \begin{bmatrix} x_1^l(t_1) & \cdots & x_1^l(t_n) \\ \vdots & \ddots & \vdots \\ x_k^l(t_1) & \cdots & x_k^l(t_n) \end{bmatrix} + b^{(1)}), \\ y^{(s)} &= \sigma(W^{(s)} \cdot y^{(s-1)} + b^{(s)}), \end{aligned} \tag{B.3}$$

where W and b are the weights and biases for the certain layer, s refers to the number of layers in the model and σ denotes the non-linear activation function. This learning process is achieved using the back propagation algorithm [75], which applies gradient descent to evaluate the influence of errors in the hidden layers on the output layer and subsequently propagates the error corrections back to earlier layers.

In this work, all past and present features will be initially treated equally with the same weight in the MLP without considering any temporal dynamics, allowing the training process to adjust these weights to assign greater importance to features that are more relevant for predicting future PV power.

B.4. Long short-term memory

Long short-term memory (LSTM) networks are a special class of RNN, capable of learning long-term dependencies [76]. The LSTM introduces a unique memory cell structure with gating mechanisms, including the input gate, forget gate, and output gate. These gates are responsible for controlling the flow of information within the memory cell, allowing the LSTM network to selectively learn, store, and retrieve information over extended time periods.

The LSTM cell computation procedures with the input $x(t_n)$ at the certain time t_n are as follows [77]:

$$\begin{aligned} i_t &= \sigma(W_{ii}x(t_n) + b_{ii} + W_{hi}h_{t-1} + b_{hi}), \\ f_t &= \sigma(W_{if}x(t_n) + b_{if} + W_{hf}h_{t-1} + b_{hf}), \\ g_t &= \tanh(W_{ig}x(t_n) + b_{ig} + W_{hg}h_{t-1} + b_{hg}), \\ o_t &= \sigma(W_{io}x(t_n) + b_{io} + W_{ho}h_{t-1} + b_{ho}), \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t, \\ h_t &= o_t \odot \tanh(c_t), \end{aligned}$$

where i_t, f_t, o_t denote the input gate, forget gate, and output gate, g_t represents the candidate values that could be added to the cell state, c_t refers to the cell state, and σ denotes the sigmoid activation function.

LSTM excels in processing sequential data and has the capacity to remember past information. In multi-step forecasting, LSTM treats both the input and output as sequences, with the objective of discovering temporal patterns between these sequence pairs.

B.5. Gated recurrent units

Gated recurrent units (GRU) [78] is also a specialized RNN that excels at capturing temporal features from hidden time series. It achieves this with a simpler structure and lower computational burden when compared to the LSTM. Considering the large volume of historical power data from multiple PV systems that needs to be inputted into the training model, the high-efficiency GRU model is utilized.

The computation procedures of the GRU cell are denoted as follows:

$$\begin{aligned} r_t &= \sigma(W_{ir}x(t_n) + b_{ir} + W_{hr}h_{t-1} + b_{hr}), \\ z_t &= \sigma(W_{iz}x(t_n) + b_{iz} + W_{hz}h_{t-1} + b_{hz}), \\ n_t &= \tanh(W_{in}x(t_n) + b_{in} + r_t \odot W_{hn}h_{t-1} + b_{hn}), \\ h_t &= (1 - z_t) \odot n_t + z_t \odot h_{t-1}, \end{aligned}$$

where h_t denote the hidden state at time t_n, r_t, z_t, n_t are the reset, update, and new gates, respectively.

B.6. Extreme gradient boosting

Extreme gradient boosting (XGB) is an efficient implementation of gradient boosted trees, a popular machine learning technique that combines weak learners iteratively to create a stronger learner by focusing on the errors made at each step [79]:

$$y^l(t_{n+1}) = \sum_{i=1}^{N} T_i(x_1^l(t_{n+1}), \ldots, x_k^l(t_{n+1})), \tag{B.4}$$

where N denotes the total number of decision trees in the XGB ensemble and $T_i(x_1^l(t_{n+1}), \ldots, x_k^l(t_{n+1}))$ represents the output of the ith decision tree.

The boosting iterations are based on the functional gradient descent approach. Notably, the loss function is approximated using the second-order Taylor expansion to handle the optimization problem. The loss function for the target values $y(t_n)$ and the predicted values $\hat{y}(t_n)$:

$$\begin{aligned} L(y(t_n), \hat{y}(t_n)) \approx \sum_{i=1}^{n} \Big[& l(y_i(t_n), \hat{y}_i(t_n)) + g_i(t_n)(\hat{y}_i(t_n) - y_i(t_n)) \\ & + \frac{1}{2} h_i(t_n)(\hat{y}_i(t_n) - y_i(t_n))^2 \Big] + \Omega(f), \end{aligned} \tag{B.5}$$

where $l(y_i(t_n), \hat{y}_i(t_n))$ denotes the loss function, $g_i(t_n)(\hat{y}_i(t_n) - y_i(t_n))$ represents the gradient of the loss function, $h_i(t_n)(\hat{y}_i(t_n) - y_i(t_n))$ represents the second derivative (Hessian) of the loss function and $\Omega(f)$ denotes the regularization term in the model f.

B.7. Gradient boosting regression

Gradient boosting regression (GBR) is an ensemble learning technique that builds a strong predictive model by combining multiple weak learners, typically decision trees, in a sequential manner [80]. The method aims to minimize the loss function by iteratively adding trees that address the residuals or errors of the previous trees. GBR can capture patterns in the data and provide robust and accurate regression predictions [81]. Different from other boosting regression algorithm, GBR uses the gradient descent approach for model tuning during the boosting iterations.

GBR is an iterative ensemble learning technique that begins by fitting a simple model, such as linear regression, to make an initial prediction using input $x^l(t_n)$, and then calculates the residual errors between the initial output $\hat{y}^l(t_{n+j})$ and the observed $y^l(t_{n+j})$. Subsequently, another model is created to predict the residual errors of the previous model, with the goal of minimizing these residual errors. The predicted

Table C.1
Selected hyperparameters for STFT model. Hyperparameters are tuned using Optuna.

Hyperparameters	Value
Attention head size	2
LSTM layers	1
Hidden continuous size	19
Hidden size	64
Learning rate	0.003
Optimizer	AdamW [47]
Dropout	0.246
Batch size	64

residual errors from this subsequent model are added to the previous model, resulting in an improved prediction and updated residual errors. This process is repeated, fitting a new subsequent model and updating the residual errors accordingly. Ultimately, the predicted values from all the models are combined to produce the final prediction, which offers a more accurate and robust estimation of the target variable.

Appendix C. Hyperparameter setting

The power outputs of distributed PV systems are predicted using different models. Considering time constraints, except for the STFT model, which was trained for 10 epochs, all other neural network models are trained for 30 epochs. The hyperparameters of STFT are tuned with the automatic hyperparameter tuning algorithm Optuna [48]. The fundamental idea is to first set the ranges for each hyperparameter and then utilize Optuna to select suitable combinations that maximize the defined objectives within a specified number of trial iterations. Table C.1 specifies the STFT hyperparameters in detail. For the other reference models, the selection of hyperparameters is based on empirical experience. The MLP model consists of three layers, including a hidden layer with 128 dimensions. In the case of LSTM, a sequence-to-sequence paradigm is employed, where the input of the encoder is determined by the number of features, and the output of the encoder is 128, matching the input of the encoder. A final linear layer is added to ensure the output dimension aligns with the desired configuration. Regarding the GRU, for the alignment, a hidden layer with 128 dimensions and 1 layer dimension is chosen.

Appendix D. Additional experiment on UNISOLAR open dataset

In order to evaluate the generalizability of the proposed model, additional experiments are conducted on public datasets using the same set of sequence inputs and step size. The additional experiment leverages the UNISOLAR dataset [82] encompassing DSGs across campus A and spanning data recorded from 2020 to 2022 at 15-min intervals. Since not all solar sites have records for all the years from 2020 to 2022, the selection of the training set is based on the number of solar sites, prioritizing those with more extensive data to ensure that the model captures a broader range of seasonal patterns. Given that the UNISOLAR dataset comprises distributed PV systems situated on the same campus, all sharing the same longitude, latitude, and irradiance data but varying in panel size, only the normalized PV data is considered. Data pre-processing follows the same paradigm as in previous experiments. Specifically, for unseen DSG assessment, to overcome year-specific findings, PV power output forecasts from 2020 to 2022 are obtained.

As demonstrated in the main text (see Section 5.2.2), LSTM and GRU are top-performing deep learning models aside from STFT. In these additional experiments, STFT, LSTM, GRU and the persistence model are chosen for the evaluation. It is shown in Table D.1 that STFT performs better in forecasting unseen DSGs, surpassing LSTM, GRU and the persistence model. The forecast skill of STFT increases from 31.19% to 43.39% as the forecast horizons expand. LSTM and GRU both demonstrate high forecasting accuracy on unseen DSGs. These models demonstrate accuracy levels that are surpassed only by the STFT method. For the 30- and 45-min forecasts, the accuracy of LSTM slightly exceeds that of GRU, consistent with the results obtained from the main text. The high accuracy of STFT in forecasting unseen DSGs in the UNISOLAR dataset can be attributed to the location of DSGs, as they are situated within the same campus. As a result, even without any training data, the attention mechanism within STFT can effectively capture these common patterns, thereby facilitating accurate predictions. The precise predictions in the UNISOLAR dataset further validate the generalizability of STFT.

Table D.1
Forecasting performance when evaluated using data of unseen DSGs. Solar sites with site keys 15, 16, 17, 18, 19, 20, 22, 24, 25, 27, 31, 33, 34, 35, 36, 37, 38, and 40 are selected for training and solar sites with site keys 14, 21, 23, 26, 30, 32, and 39 are reserved for unseen DSG assessment. STFT model outperforms LSTM, GRU and the persistence model in 15-, 30-, and 45-min forecasts, exhibiting both the smallest RMSE values and the largest s values.

Step	Metric	STFT	LSTM	GRU	Persistence
15-min	RMSE (W)	**4.89**	7.43	7.29	7.33
	s	**31.19%**	−1.59%	0.56%	–
30-min	RMSE (W)	**5.69**	9.10	9.11	9.63
	s	**38.92%**	5.47%	5.29%	–
45-min	RMSE (W)	**6.08**	10.26	10.26	11.12
	s	**43.39%**	7.56%	7.45%	–

References

[1] Ehsan A, Yang Q. Optimal integration and planning of renewable distributed generation in the power distribution networks: A review of analytical techniques. Appl Energy 2018;210:44–59.

[2] Jiang S, Wan C, Chen C, Cao E, Song Y. Distributed photovoltaic generation in the electricity market: status, mode and strategy. CSEE J Power Energy Syst 2018;4(3):263–72.

[3] Mah DN-y, Wang G, Lo K, Leung MK, Hills P, Lo AY. Barriers and policy enablers for solar photovoltaics (PV) in cities: Perspectives of potential adopters in Hong Kong. Renew Sustain Energy Rev 2018;92:921–36.

[4] Haben S, Arora S, Giasemidis G, Voss M, Greetham DV. Review of low voltage load forecasting: Methods, applications, and recommendations. Appl Energy 2021;304:117798.

[5] Chen S, Gooi H, Wang M. Solar radiation forecast based on fuzzy logic and neural networks. Renew Energy 2013;60:195–201.

[6] Inman RH, Pedro HTC, Coimbra CFM. Solar forecasting methods for renewable energy integration. Prog Energy Combust Sci 2013;39(6):535–76.

[7] Lave M, Kleissl J. Solar variability of four sites across the state of Colorado. Renew Energy 2010;35(12):2867–73.

[8] 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.

[9] Liu C, Li M, Yu Y, Wu Z, Gong H, Cheng F. A review of multitemporal and multispatial scales photovoltaic forecasting methods. IEEE Access 2022;10:35073–93.

[10] Qu Y, Xu J, Sun Y, Liu D. A temporal distributed hybrid deep learning model for day-ahead distributed PV power forecasting. Appl Energy 2021;304:117704.

[11] Miyazaki Y, Kameda Y, Kondoh J. A power-forecasting method for geographically distributed PV power systems using their previous datasets. Energies 2019;12(24):4815.

[12] Simeunović J, Schubnel B, Alet P-J, Carrillo RE, Frossard P. Interpretable temporal-spatial graph attention network for multi-site PV power forecasting. Appl Energy 2022;327:120127.

[13] Hong T, Pinson P, Wang Y, Weron R, Yang D, Zareipour H. Energy forecasting: A review and outlook. IEEE Open Access J Power Energy 2020;7:376–88.

[14] De Hoog J, Perera M, Ilfrich P, Halgamuge S. Characteristic profile: improved solar power forecasting using seasonality models. ACM SIGENERGY Energy Inform Rev 2021;1(1):95–106.

[15] Van Gompel J, Spina D, Develder C. Satellite based fault diagnosis of photovoltaic systems using recurrent neural networks. Appl Energy 2022;305:117874.

[16] Mishra S, Anderson K, Miller B, Boyer K, Warren A. Microgrid resilience: A holistic approach for assessing threats, identifying vulnerabilities, and designing corresponding mitigation strategies. Appl Energy 2020;264:114726.

[17] Sobri S, Koohi-Kamali S, Rahim NA. Solar photovoltaic generation forecasting methods: A review. Energy Convers Manage 2018;156:459–97.

[18] Lim B, Arık SÖ, Loeff N, Pfister T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. Int J Forecast 2021;37(4):1748–64.

[19] Mayer MJ, Gróf G. Extensive comparison of physical models for photovoltaic power forecasting. Appl Energy 2021;283:116239.

[20] Ahmed R, Sreeram V, Mishra Y, Arif M. A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. Renew Sustain Energy Rev 2020;124:109792.

[21] Zhi Y, Sun T, Yang X. A physical model with meteorological forecasting for hourly rooftop photovoltaic power prediction. J Build Eng 2023;75:106997.

[22] Atique S, Noureen S, Roy V, Subburaj V, Bayne S, Macfie J. Forecasting of total daily solar energy generation using ARIMA: A case study. In: 2019 IEEE 9th annual computing and communication workshop and conference. CCWC, IEEE; 2019, p. 0114–9.

[23] Lauria D, Mottola F, Proto D. Caputo derivative applied to very short time photovoltaic power forecasting. Appl Energy 2022;309:118452.

[24] Li P, Zhou K, Lu X, Yang S. A hybrid deep learning model for short-term PV power forecasting. Appl Energy 2020;259:114216.

[25] Akhter MN, Mekhilef S, Mokhlis H, Ali R, Usama M, Muhammad MA, Khairuddin ASM. A hybrid deep learning method for an hour ahead power output forecasting of three different photovoltaic systems. Appl Energy 2022;307:118185.

[26] Paletta Q, Terrén-Serrano G, Nie Y, Li B, Bieker J, Zhang W, Dubus L, Dev S, Feng C. Advances in solar forecasting: Computer vision with deep learning. Adv Appl Energy 2023;100150.

[27] Mohammed MA, Ahmed MA, Hacimahmud AV. Data-driven sustainability: Leveraging big data and machine learning to build a greener future. Babylon J Artif Intell 2023;2023:17–23.

[28] Wang H, Lei Z, Zhang X, Zhou B, Peng J. A review of deep learning for renewable energy forecasting. Energy Convers Manage 2019;198:111799.

[29] 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.

[30] Feng C, Zhang J. SolarNet: A sky image-based deep convolutional neural network for intra-hour solar forecasting. Sol Energy 2020;204:71–8.

[31] Song H, Al Khafaf N, Kamoona A, Sajjadi SS, Amani AM, Jalili M, Yu X, McTaggart P. Multitasking recurrent neural network for photovoltaic power generation prediction. Energy Rep 2023;9:369–76.

[32] Lee D, Kim K. PV power prediction in a peak zone using recurrent neural networks in the absence of future meteorological information. Renew Energy 2021;173:1098–110.

[33] Korkmaz D. SolarNet: A hybrid reliable model based on convolutional neural network and variational mode decomposition for hourly photovoltaic power forecasting. Appl Energy 2021;300:117410.

[34] Chu Y, Li M, Coimbra CFM, Feng D, Wang H. Intra-hour irradiance forecasting techniques for solar power integration: A review. Iscience 2021;24(10).

[35] 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;30:5998–6008.

[36] Ju Y, Li J, Sun G. Ultra-short-term photovoltaic power prediction based on self-attention mechanism and multi-task learning. IEEE Access 2020;8:44821–9.

[37] Zhang Z, Wang J, Wei D, Xia Y. An improved temporal convolutional network with attention mechanism for photovoltaic generation forecasting. Eng Appl Artif Intell 2023;123:106273.

[38] Kharlova E, May D, Musilek P. Forecasting photovoltaic power production using a deep learning sequence to sequence model with attention. In: 2020 international joint conference on neural networks. IJCNN, IEEE; 2020, p. 1–7.

[39] Hu Z, Gao Y, Ji S, Mae M, Imaizumi T. Improved multistep ahead photovoltaic power prediction model based on LSTM and self-attention with weather forecast data. Appl Energy 2024;359:122709.

[40] Niu T, Li J, Wei W, Yue H. A hybrid deep learning framework integrating feature selection and transfer learning for multi-step global horizontal irradiation forecasting. Appl Energy 2022;326:119964.

[41] Zhao G, Xue M, Cheng L. A new hybrid model for multi-step WTI futures price forecasting based on self-attention mechanism and spatial–temporal graph neural network. Resour Policy 2023;85:103956.

[42] Tian C, Niu T, Wei W. Developing a wind power forecasting system based on deep learning with attention mechanism. Energy 2022;257:124750.

[43] López Santos M, García-Santiago X, Echevarría Camarero F, Blázquez Gil G, Carrasco Ortega P. Application of temporal fusion transformer for day-ahead PV power forecasting. Energies 2022;15(14):5232.

[44] Mazen FMA, Shaker Y, Abul Seoud RA. Forecasting of solar power using GRU–temporal fusion transformer model and DILATE loss function. Energies 2023;16(24):8105.

[45] Giacomazzi E, Haag F, Hopf K. Short-term electricity load forecasting using the temporal fusion transformer: Effect of grid hierarchies and data sources. In: Proceedings of the 14th ACM international conference on future energy systems. 2023, p. 353–60.

[46] Kumar A, Kashyap Y, Kosmopoulos P. Enhancing solar energy forecast using multi-column convolutional neural network and multipoint time series approach. Remote Sens 2022;15(1):107.

[47] Niu Y, Wang J, Zhang Z, Luo T, Liu J. De-Trend First, Attend Next: A Mid-Term PV forecasting system with attention mechanism and encoder–decoder structure. Appl Energy 2024;353:122169.

[48] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019, p. 2623–31.

[49] Global Solar Atlas 29. Solar resource map: Photovoltaic power potential. 2023, Available at: https://globalsolaratlas.info/. [Accessed 27 January 2024].

[50] Lefevre M, Oumbe A, Blanc P, Espinar B, Gschwind B, Qu Z, Wald L, Schroedter-Homscheidt M, Hoyer-Klick C, Arola A, Benedetti A, Kaiser J, Morcrette J-J. McClear: a new model estimating downwelling solar radiation at ground level in clear-sky conditions. Atmos Meas Tech 2013;6(9):2403–18.

[51] Jo J-M. Effectiveness of normalization pre-processing of big data to the machine learning performance. J Korea Inst Electron Commun Sci 2019;14(3):547–52.

[52] Gridin I. Time series forecasting using deep learning: Combining PyTorch, RNN, TCN, and deep neural network models to provide production-ready prediction solutions. English ed.. BPB Publications; 2021.

[53] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825–30.

[54] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S. PyTorch: An imperative style, high-performance deep learning library. In: Advances in neural information processing systems 32. Curran Associates, Inc.; 2019, p. 8024–35.

[55] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. KDD '16, New York, NY, USA: ACM; 2016, p. 785–94.

[56] Yang D, Gu Y, Mayer MJ, Gueymard CA, Wang W, Kleissl J, Li M, Chu Y, Bright JM. Regime-dependent 1-min irradiance separation model with climatology clustering. Renew Sustain Energy Rev 2024;189:113992.

[57] Chu Y, Yang D, Yu H, Zhao X, Li M. Can end-to-end data-driven models outperform traditional semi-physical models in separating 1-min irradiance? Appl Energy 2024;356:122434.

[58] Marquez R, Coimbra CF. Proposed metric for evaluation of solar forecasting models. J Sol Energy Eng 2013;135(1):011016.

[59] Srivastava S, Lessmann S. A comparative study of LSTM neural networks in forecasting day-ahead global horizontal irradiance with satellite data. Sol Energy 2018;162:232–47.

[60] McGill R, Tukey JW, Larsen WA. Variations of box plots. Amer Statist 1978;32(1):12–6.

[61] Murphy AH, Winkler RL. A general framework for forecast verification. Mon Weather Rev 1987;115(7):1330–8.

[62] Gers FA, Schraudolph NN, Schmidhuber J. Learning precise timing with LSTM recurrent networks. J Mach Learn Res 2002;3:115–43.

[63] Hu Y, Liu H, Wu S, Zhao Y, Wang Z, Liu X. Temporal collaborative attention for wind power forecasting. Appl Energy 2024;357:122502.

[64] Yu F, Huang K, Wang M, Cheng Y, Chu W, Cui L. Width & depth pruning for vision transformers. In: Proceedings of the AAAI conference on artificial intelligence. Vol. 36, 2022, p. 3143–51.

[65] Kim B, Lee SH, Kim H, Nguyen D-T, Le M-S, Chang IJ, Kwon D, Yoo JH, Choi JW, Lee H-J. PCM: precision-controlled memory system for energy efficient deep neural network training. In: 2020 design, automation & test in europe conference & exhibition. DATE, IEEE; 2020, p. 1199–204.

[66] Kim S, Park G, Yi Y. Performance evaluation of INT8 quantized inference on mobile GPUs. IEEE Access 2021;9:164245–55.

[67] Zhao Z, Lyu J, Chu Y, Liu K, Cao D, Wu C, Qin L, Qin S. Toward generalizable robot vision guidance in real-world operational manufacturing factories: A Semi-Supervised Knowledge Distillation approach. Robot Comput-Integr Manuf 2024;86:102639.

[68] Ziyabari S, Du L, Biswas S. A spatio-temporal hybrid deep learning architecture for short-term solar irradiance forecasting. In: 2020 47th IEEE photovoltaic specialists conference. PVSC, IEEE; 2020, p. 0833–8.

[69] Ahn HK, Park N. Deep RNN-based photovoltaic power short-term forecast using power IoT sensors. Energies 2021;14(2):436.

[70] Dauphin YN, Fan A, Auli M, Grangier D. Language modeling with gated convolutional networks. In: International conference on machine learning. PMLR; 2017, p. 933–41.

[71] Gal Y, Ghahramani Z. A theoretically grounded application of dropout in recurrent neural networks. Adv Neural Inf Process Syst 2016;29:1019–27.

[72] Lauret P, David M, Pedro HTC. Probabilistic solar forecasting using quantile regression models. Energies 2017;10(10):1591.

[73] Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. John Wiley & Sons; 2021.

[74] Rocha PAC, Fernandes JL, Modolo AB, Lima RJP, da Silva MEV, Bezerra CAD. Estimation of daily, weekly and monthly global solar radiation using ANNs and a long data set: a case study of Fortaleza, in Brazilian Northeast region. Int J Energy Environ Eng 2019;10:319–34.

[75] Liu J, Shao M, Sun M. The forecast of power consumption and freshwater generation in a solar-assisted seawater greenhouse system using a multi-layer perceptron neural network. Expert Syst Appl 2023;213:119289.

[76] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.

[77] Ma X, Tao Z, Wang Y, Yu H, Wang Y. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transp Res C 2015;54:187–97.

[78] Wang Y, Liao W, Chang Y. Gated recurrent unit network-based short-term photovoltaic forecasting. Energies 2018;11(8):2163.

[79] Li X, Ma L, Chen P, Xu H, Xing Q, Yan J, Lu S, Fan H, Yang L, Cheng Y. Probabilistic solar irradiance forecasting based on XGBoost. Energy Rep 2022;8:1087–95.

[80] Persson C, Bacher P, Shiga T, Madsen H. Multi-site solar power forecasting using gradient boosted regression trees. Sol Energy 2017;150:423–36.

[81] Torres-Barrán A, Alonso Á, Dorronsoro JR. Regression tree ensembles for wind energy and solar radiation prediction. Neurocomputing 2019;326:151–60.

[82] Wimalaratne S, Haputhanthri D, Kahawala S, Gamage G, Alahakoon D, Jennings A. Unisolar: An open dataset of photovoltaic solar energy generation in a large multi-campus university setting. In: 2022 15th international conference on human system interaction. HSI, IEEE; 2022, p. 1–5.