Multistep photovoltaic power forecasting based on multi-timescale fluctuation aggregation attention mechanism and contrastive learning

Yuan, Liang; Wang, Xiangting; Sun, Yao; Liu, Xubin; Dong, Zhao Yang

Published in:
International Journal of Electrical Power and Energy Systems

Published: 01/03/2025

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

Published version (DOI):
10.1016/j.ijepes.2024.110389

Publication details:
Yuan, L., Wang, X., Sun, Y., Liu, X., & Dong, Z. Y. (2025). Multistep photovoltaic power forecasting based on multi-timescale fluctuation aggregation attention mechanism and contrastive learning. *International Journal of Electrical Power and Energy Systems*, *164*, Article 110389. https://doi.org/10.1016/j.ijepes.2024.110389

Electrical Power and Energy Systems 164 (2025) 110389

Contents lists available at ScienceDirect

International Journal of Electrical Power and Energy Systems

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

Multistep photovoltaic power forecasting based on multi-timescale fluctuation aggregation attention mechanism and contrastive learning

Liang Yuan [a], Xiangting Wang [a], Yao Sun [a], Xubin Liu [a,*], Zhao Yang Dong [b]

[a] School of Automation, Central South University, Changsha 410083, China
[b] Department of Electrical Engineering, City University of Hong Kong, Hong Kong

ARTICLE INFO

Keywords:
Photovoltaic power forecasting
Self-attention mechanism
Similar day selection
Contrastive learning
Transformer

ABSTRACT

The integration of photovoltaic (PV) power into electrical grids introduces significant uncertainty due to the inherent volatility and intermittency of solar energy, underscoring the need for precise short and medium-term PV power forecasting. Despite the superior performance of Transformer-based time series methods, their application to PV power prediction remains suboptimal. In response to this deficiency, this paper proposes a novel attention mechanism that aggregates fluctuations across multiple time scales. This mechanism enhances the segmentation and extraction of nonlinear correlations between PV power outputs and meteorological factors, assigning variable weights to patterns of change across different time scales. Furthermore, a novel approach for selecting similar days is also developed based on contrastive learning, which enables self-supervised identification of similarities among PV power samples and enhances the model's attention to local dynamic variations. Comparative analysis with eight state-of-the-art benchmark methods shows that the proposed MFA-attention model achieves lower prediction errors and improved effectiveness.

1. Introduction

Electricity is the cornerstone of modern social, economic and technological development [1]. As the continuous consumption of non-renewable resources like fossil fuels and coal exacerbates environmental pollution, governments are increasingly focusing on developing renewable energy technologies [2,3]. Compared to fossil fuels, solar energy exhibits outstanding competitiveness in clean, low-carbon and sustainable. The International Energy Agency's "Renewable Energy Market 2023″ report forecasts that the global cumulative capacity of renewable energy generation will surpass 4500GW by the end of 2024 [4]. However, solar power generation exhibits significant randomness, fluctuations, and intermittency due to meteorological conditions and geographical factors[5]. Grid operators are faced with the challenge of random supply and demand balance, which makes the safe and stable operation of the grid more difficult [6–8]. Therefore, accurate PV forecasting is critical to ensuring the stable and reliable operation of power systems. PV forecasting is categorized by duration: ultra-short-term, short-term, medium-term, and long-term [9]. Ultra-short-term forecasting involves real-time monitoring of PV power generation within 1 min to 4 h, which is primarily used for immediate power regulation to prevent transient power system issues [10,11]. Short-term forecasts cover 4 to 12 h, which provide essential data for formulating daily power generation plans [12] and energy management. Medium-term and long-term forecasting usually refers to the period from 12 h to 336 h [13], which belongs to the range of long sequence time-series forecasting (LSTF). These forecasts guide maintenance for solar-integrated systems. Based on this perspective, the multistep PV forecasting discussed in this paper is a medium-to-short-term prediction. The forecasting horizon can be flexibly adjusted according to the requirements of decision-makers to facilitate grid scheduling [14].

PV forecasting methods can be broadly classified into two categories: physical methods and data-driven methods. As for the physical methods, forecasts typically rely on establishing a physical model of the photoelectric conversion relationship, integrating Numerical Weather Prediction (NWP) with geographic data [15,16]. This process involves numerous parameters, making it costly and complex. Conversely, data-driven models establish a nonlinear mapping between input and output for high-precision forecasting. These models, easier to develop and generalize, are becoming the mainstream in time series forecasting. Data-driven methods include classical models [17–19], machine learning models like Boost [20], Support Vector Machine (SVM)

* Corresponding author.
E-mail address: liuxubin@csu.edu.cn (X. Liu).

https://doi.org/10.1016/j.ijepes.2024.110389
Received 30 May 2024; Received in revised form 14 September 2024; Accepted 17 November 2024
Available online 1 December 2024

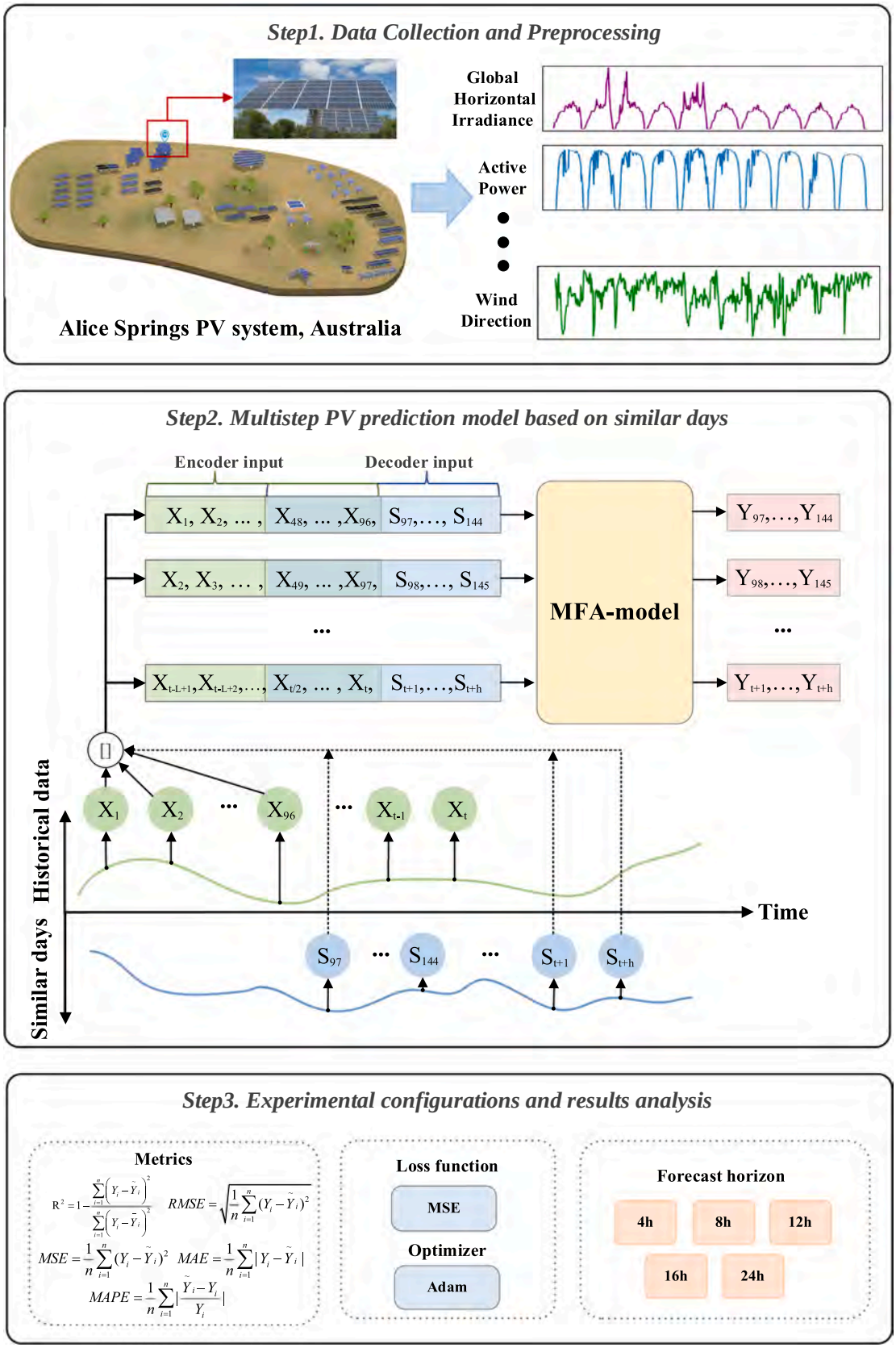

Fig. 1. Flowchart of the MFA-attention PV power forecasting.

regression [21], and Artificial Neural Networks (ANN) [22,23], and deep learning models. Classical and machine learning models are sensitive to the selection of hyperparameters and limited in extracting nonlinear features [24,25]. With increasing prediction horizons and sequence complexity, these methods often struggle to improve accuracy. The main difference between deep learning and other machine learning approaches lies in deep learning's ability to extract more complex high-dimensional features from massive datasets. In the development stage of neural networks, Recurrent Neural Networks (RNN) is undoubtedly the most suitable architecture for time series forecasting [26]. It calculates the hidden state h_t based on the current time position information and the h_{t-1} of the previous moment. However, this calculation structure becomes memory-intensive with longer prediction horizons. It struggles to fully explore repetitive patterns in LSTF and risks gradient vanishing in deep network structures. RNN variants like Long Short-Time Memory (LSTM) [27] and Gated Recurrent Unit (GRU) [28] use gating mechanisms to mitigate gradient vanishing and explosion. To enhance spatio-temporal analysis, studies have incorporated convolution operations into time series modeling. Ref. [29] replaced LSTM's matrix multiplication with convolution operations to capture spatial features in multidimensional data, which retains LSTM's temporal traits while using convolutions for spatial analysis. This innovation overcomes LSTM's limitations in handling spatiotemporal relationships. Ref. [30] combined CNN with LSTM and adds self-attention to better extract spatiotemporal correlations. However, these methods are limited by the model structures and still have the dilemma of insufficient learning ability in LSTF.

Recently, transformers have excelled in Natural Language Processing (NLP) due to their impressive ability to model long-term time series dependencies [31]. Inspired by transformers, fields like time series forecasting and computer vision have adopted transformer architectures, specifically the encoder/decoder structure and self-attention mechanism [32,33]. Researchers are addressing RNN limitations in extracting long-term dependencies and interrelationships of external factors through sequence-to-sequence models. In Ref. [34], an improved whale variational mode decomposition was employed to decompose multiple-channel sub-sequences. Causal convolutional embedding blocks were integrated to encode key sub-sequences, enhancing dynamic data representation. Ref. [35] proposed a PV forecasting method that embedded LSTM into Transformers. This approach uses LSTM-extracted temporal features in a self-attention framework to effectively capture feature correlations, enhancing the use of weather forecast data. Experimental results showed that this hybrid model outperformed standalone LSTM in both short-term and long-term solar energy forecasting accuracy.

The key to improving PV prediction performance lies in capturing the underlying patterns of complex dynamics from continuous observations and extrapolating future states. However, the original Transformer model has shortcomings that limit its effectiveness in PV power prediction. Specifically, in time series with complex temporal patterns like solar energy, there is a coupling between temporal PV generation values and spatial weather factors. Transformer models often fail to fully account for the nuances of multidimensional weather variables across different timescales and their dynamic effects on PV output, frequently missing short-term local variations. Conventional self-attention methods struggle to extract reliable correlations from the raw data's random, complex, and repetitive patterns. However, PV data exhibit distinct regularities in both short-term and long-term patterns, showing increased continuity and correlation across various segments. Therefore, a more detailed analysis of segment correlations is necessary to improve the model's ability to capture multiscale temporal dependencies. In addition to model improvements, many rely solely on historical or externally provided weather data for training. The precision and resolution of weather data crucially affect forecast accuracy. To counteract this, researchers have introduced similar day analysis (SDA) to refine solar power forecasting. In Ref. [36], the Euclidean distance between the forecast date and historical data from the preceding k days was calculated. The PV power of the past k days was aggregated based on the distance magnitude to serve as the model input. The Pearson correlation coefficient was adopted to measure the similarity between different days, and similar days were selected as inputs for the GA-ELM model in [37]. Most methods employ basic metrics like Euclidean distance or the L1 norm to assess similarity, which may overlook daily and meteorological variations, sometimes resulting in inaccurate day matching. These inaccuracies, in turn, adversely affect the predictive precision of subsequent models. Few studies integrate weather forecasts with similar day data to enhance model input relevance. Addressing these challenges, the principal contributions of this paper are summarized as follows:

(1) A multi-timescale fluctuation aggregation (MFA) attention method for PV power forecasting based on the transformer model is proposed. This method models trend and seasonal information separately, captures time dependencies across various timescales, and weights the similarity of input PV sequences of different segment lengths. This approach aims to improve prediction accuracy by addressing the dynamic impact of multi-dimensional meteorological features.

(2) A similar day selection method based on contrastive learning is proposed. Through self-supervised learning, the similarity between meteorological features in PV data is assessed. Then, based on the learned feature representations, Dynamic time warping (DTW) is employed to identify days with high similarity for prediction. This simplifies the process of similar day analysis, providing the model with more accurate information.

(3) A multi-step prediction method is proposed to ensure prediction accuracy and generate the prediction sequence in one step to avoid the accumulation of prediction errors. Extensive experiments are conducted to validate the performance of the proposed model with other deep learning models. The method's effectiveness is further validated through ablation study and interpretable analysis.

The rest of the paper is organized as follows: Section 2 outlines the proposed method's structure, including similar day selection and the MFA model. Section 3 details the dataset, preprocessing, experiment procedures, and model parameter settings. Section 4 analyzes and discusses the experimental results. Finally, in Section 5, the key findings of this paper are summarized.

2. Methodology

This section focuses on the proposed PV forecasting method. Firstly, the problem definition of the PV series forecasting task is given, followed by the similar day selection method based on contrastive learning. Then, the multi-timescale fluctuation aggregation attention method is introduced. Finally, the general framework of the MFA-attention model is presented. Fig. 1 shows the PV prediction flowchart, and the pseudocode is provided in Appendix A.

2.1. Problem definition

PV forecasting tasks can be described as: Given the historical PV series of length L, $X_{t-L+1:t} = \{x_{t-L+1}, x_{t-L+2}, \ldots, x_t\}$, where $X_{t-L+1:t} \in \mathbb{R}^{L\times d}$ and d denotes the characteristic dimension of the variates, usually including weather factors such as irradiance, temperature and wind speed. $S_{t+1:t+h} \in \mathbb{R}^{h\times d}$ is the concatenation value of the selected similar day's PV power and weather forecast data. We aim to predict the PV power $\widetilde{Y}_{t+1:t+h} \in \mathbb{R}^{h\times d}$ of h forecasting horizon at the time step t in future, as given in (1):

$$\widetilde{Y}_{t+1:t+h} = F(X_{t-L+1:t}, S_{t+1:t+h}; \theta) \tag{1}$$

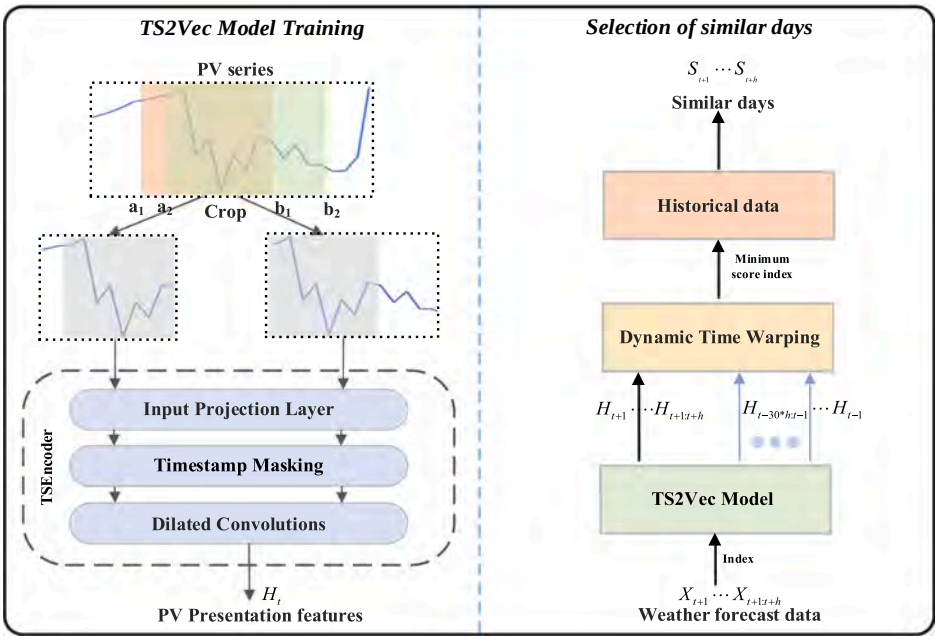

Fig. 2. The similar day selection steps based on contrastive learning.

2.2. Selection of similar days

Few studies train a deep learning model for similar days, as determining the labels is a complex and challenging process. Contrastive learning is a self-supervised learning method. Its core idea is that the model can automatically learn similar features from massive data by constructing positive and negative pairs of input data without any labels. In this paper, a general framework designed by TS2Vec [38] for time series is used to train the inherent similarity between PV data samples and other samples within 30 days. TS2Vec can generate high-dimensional features for PV sequences at any temporal granularity and randomly masks features at certain timestamps within subsequences, enhancing the model's ability to capture sequence similarities. The contrastive learning loss function ensures the minimization of cross-entropy loss between positive samples, while maintaining temporal consistency in sample selection. Dynamic time warping (DTW) is employed to measure subsequent similar samples. This aids the model in acquiring similar PV power data for the prediction period, facilitating a more reasonable inference of output power by the decoder at the predicted time.

The detailed principles of the TS2Vec model are provided in Appendix B. For two similar PV series, DTW optimally aligns their elements at different time points, creating an optimal warping path that better measures the similarity between the time series [39]. To search for the historical days with the most similar weather conditions, the steps are as follows: The PV samples used for TS2Vec model training are $X \in R^{L\times d}$. To prevent information leakage, the input features do not include PV, which is $X \in R^{L\times(d-1)}$. The feature vector output of the TS2Vec model is the learned high-dimensional representation feature of the PV series. The trained TS2Vec model is saved. During the training process, DTW is used to calculate the similarity of high-dimensional representations of the historical weather data 30 days before the forecast date, and the subset with the smallest DTW value in the set is found, the similar day power data is used as input for the decoder are automatically selected within the model, eliminating the need for complex preprocessing steps, as shown in Fig. 2.

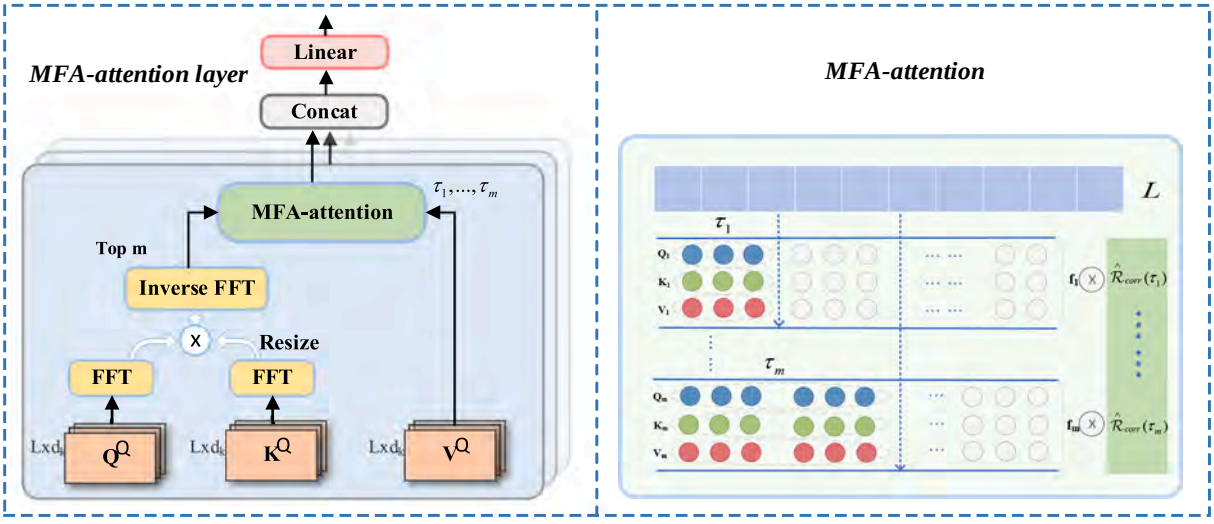

Fig. 3. The calculation method of MFA-attention.

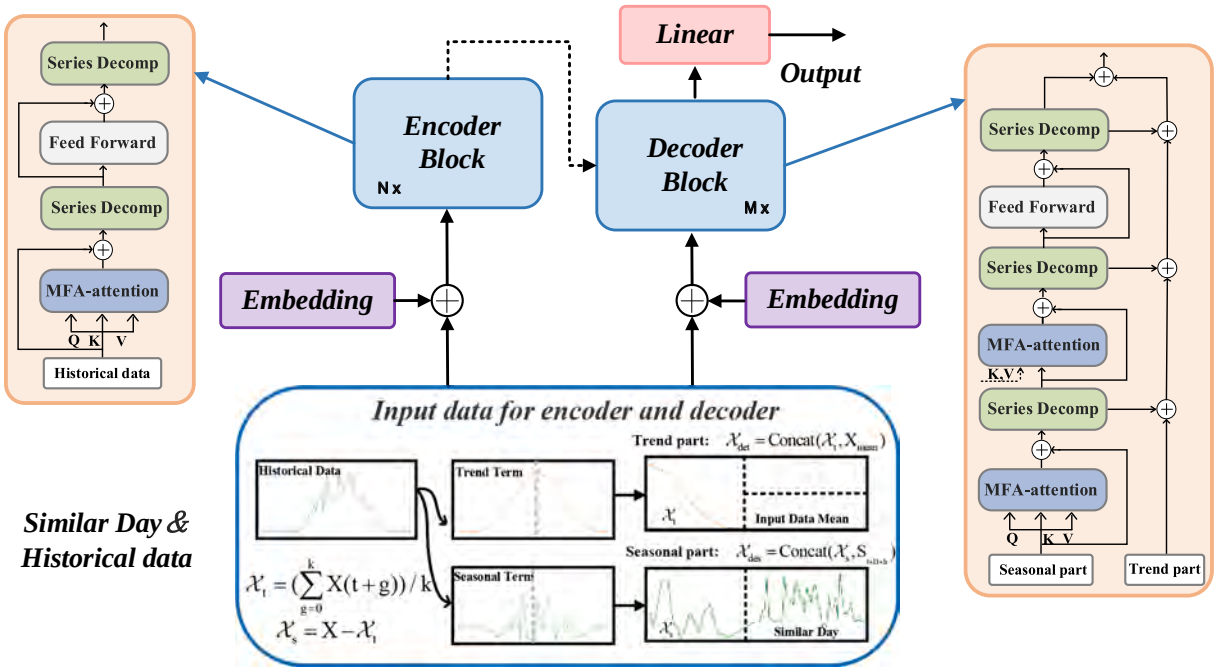

Fig. 4. Encoder and decoder block for MFA-model.

2.3. Multiscale fluctuate aggregation

The variation in the PV series is influenced by prior data segments, indicating strong continuity and local internal dynamics within a segment. The traditional scaled dot-product attention mechanism struggles to aggregate information within continuous segments. Additionally, not all queries share the same dominant characteristics as the time series. Therefore, a multi-timescale fluctuation aggregation attention mechanism is proposed. By calculating the similarity between segments of different lengths and weighting them based on their importance, it effectively captures the continuous correlations in PV series changes across various time scales. The overall calculation method for MFA-attention is illustrated in Fig. 3.

In the calculation of MFA attention, the input Q, K, and V are divided into i segments, where each segment consists of a subsequence of the input data, denoted as Q_i, K_i, V_i , $\tau_i \in \{\tau_1, ..., \tau_m\}$ represent different subsegment lengths. The $Q_i \in \mathbb{R}^{\tau_i \times d_k}$、$K_i \in \mathbb{R}^{\tau_i \times d_k}$、$V_i \in \mathbb{R}^{\tau_i \times d_k}$ are still obtained through linear project matrices, In cross-attention, K and V come from the encoder, while Q comes from the input of the decoder. MFA attention is computed by the Eq. (2):

$$f_i = \text{Aggregation}\ (Q_i, K_i, V_i) = \text{Softmax}(\frac{Q_i K_i^{\mathrm{T}}}{d_k})V_i \tag{2}$$

The segment aggregation factor f_i is obtained by applying SoftMax normalization to the corresponding length V_i. This $Q_i, K_i, V_i \in \mathbb{R}^{\tau_i \times d_k}$ needs to be zero-padded to $Q_i, K_i, V_i \in \mathbb{R}^{L \times d_k}$ before adding segment aggregation factors of different segment lengths.

According to the calculation of the segment aggregation factor, the selection of segment length τ is crucial. Small τ can capture fine-scale fluctuations in the series over short time scales, while large τ can reflect the trend of PV series over long time scales. Different times and seasons show varying degrees of fluctuation in the PV series, such as the steepness of rise/fall and the level of fluctuation during the rise. To comprehensively integrate the fluctuation information contained in various time scales, inspired by [40], we employ the Fast Fourier Transform (FFT) to analyze each frequency component from the perspective of the frequency domain. The autocorrelation of the sequence is calculated to obtain the similarity of the same series at different frequencies. The autocorrelation can be obtained as:

$$S_{\text{corr}}(f) = \mathscr{F}(X_t)\mathscr{F}^{*}(X_t) = \int_{-\infty}^{\infty} X_t e^{-i2\pi ff}\mathrm{d}t \overline{\int_{-\infty}^{\infty} X_t e^{-i2\pi tf}\mathrm{d}t} \tag{3}$$

$$\mathscr{R}_{corr}(\tau) = \mathscr{F}^{-1}(S_{\text{corr}}(f)) = \int_{-\infty}^{\infty} S_{\text{corr}}(f) e^{i2\pi f\tau}\mathrm{d}f \tag{4}$$

where X_t is the input series, $\mathscr{F}$ denotes the FFT and $\mathscr{F}^{-1}$ is its inverse. * denotes the conjugate operation. The frequency-domain correlation $S_{\text{corr}}(f)$ is converted into time-domain correlation $\mathscr{R}_{corr}(\tau)$ using the inverse FFT operation, which reflects the time-delay similarity between the X_t and its τ lag series $X_{t-\tau}$.

In order to preserve the integrity of the original signal to the maximum extent while filtering out some noise and irrelevant information, we select the top m time delay positions with the highest autocorrelation from τ, let $m = |q \times \log L/2|$,L is the length of the input sequence,q is a hyper-parameter. The Multi-scale Fluctuate Aggregation attention mechanism proposed in this paper calculates the segment aggregation factors at this position based on the time-domain relevance weighting of different segment lengths, as defined by Eq. (5)- Eq. (7).

$$\tau_1, ..., \tau_m = \underset{\tau \in \{1,...,L\}}{\mathit{argmax}}(m, \mathscr{R}_{corr}(\tau)) \tag{5}$$

$$\hat{\mathscr{R}}_{corr}(\tau_1), ..., \hat{\mathscr{R}}_{corr}(\tau_m) = \text{Softmax}[\mathscr{R}_{corr}(\tau_1), ..., \mathscr{R}_{corr}(\tau_m)] \tag{6}$$

$$\text{MFA - attention}(Q, K, V) = \sum_{i=1}^{m} f_i \hat{\mathscr{R}}_{corr}(\tau_i) \tag{7}$$

2.4. MFA-attention model framework

The model consists of an encoder-decoder structure, which mainly includes the following basic modules. Fig. 4 shows the block diagram of the encoder and decoder.

Decomposition block: Time series decomposition is a routine procedure as a pre-processing step. The better way is to add the decomposition process as an inner operation of the model, allowing it to actively participate in the parameter optimization during model

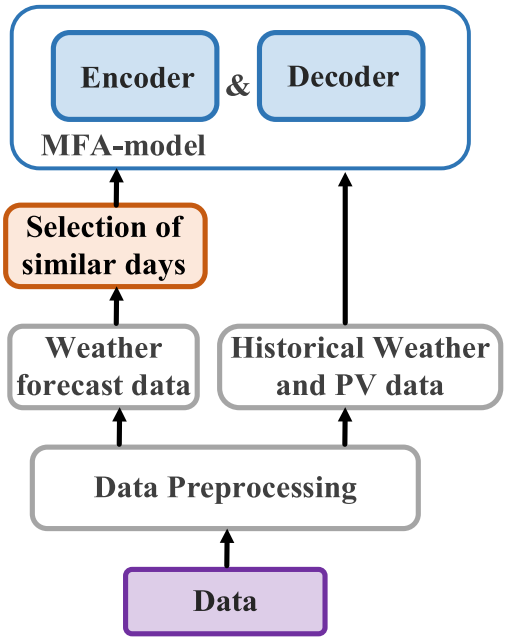

Fig. 5. The overall framework of the proposed method.

training. Decompose the historical data $X_{t-L+1:t} \in \mathbb{R}^{L\times d}$ into the seasonal part and trend part using Eq. (8) and Eq. (9).

$$\mathscr{X}_t = (\sum_{g=0}^{k} X(t+g))/k \tag{8}$$

$$\mathscr{X}_s = X - \mathscr{X}_t \tag{9}$$

The decomposition block is represented by $\mathscr{X}_s$, $\mathscr{X}_t = SeriesDecomp(X)$. k is the moving average window size.

Input data: The decoder models the seasonal part $\mathscr{X}_{des}$ and the trend part $\mathscr{X}_{det}$ separately. The obtained PV power data from similar days and the meteorological data for forecast days are combined into $S_{t+1:t+h} \in \mathbb{R}^{h\times d}$. The decoder's input $\mathscr{X}_{des}$ includes the second half length $L/2$ of the seasonal part $\mathscr{X}_s$ and $S_{t+1:t+h}$ of predicted length h is padded to $\mathscr{X}_{des} \in \mathbb{R}^{(L/2+h)\times d}$. The benefit of this approach is to furnish the decoder with contextual information, aiding the model in comprehending global patterns. The trend part $\mathscr{X}_{det}$ includes the latter half of the encoder's trend $\mathscr{X}_t$ and the mean of historical data $X_{t-L+1:t} \in \mathbb{R}^{L\times d}$ are filled by $\mathscr{X}_{\text{det}} \in \mathbb{R}^{(L/2+h)\times d}$. The encoder takes the historical data from the previous day as input, while the decoder's two-part input can be represented by the following Eq. (10)- Eq. (13):

$$\mathscr{X}_s, \mathscr{X}_t = SeriesDecomp(X_{t-L+1:t}[L/2 : L]) \tag{10}$$

$$\mathscr{X}_{\text{des}} = Concat(\mathscr{X}_s, S_{t+1:t+h}) \tag{11}$$

$$\mathscr{X}_{\text{det}} = Concat(\mathscr{X}_t, X_{mean}) \tag{12}$$

The framework of the Encoder and Decoder follows the backbone of Autoformer, as detailed in [40]. The overall framework diagram of the proposed method is shown in Fig. 5.

$$\widetilde{Y}_{t+1:t+h} = (\mathscr{T}^n_{\text{de}} + W_1 * \mathscr{X}^n_{de})[L : L+h] \tag{13}$$

3. Case study

This section details the datasets, evaluation metrics, and baseline model parameter settings to ensure reproducibility and fairness of the experiments. The experiments are based on the implementation of the PyTorch framework in Python 3.8 with Cuda 11.1, running on an NVIDIA GeForce RTX 3090 GPU and a DELL PowerEdge T640.

3.1. Datasets and preprocessing

To evaluate the MFA model's effectiveness, we used two different public datasets: Trina and Station01. The Trinadataset comes from the Desert Knowledge Australia Solar Centre (DKASC) in Alice Springs, Australia [41]. The proposed is evaluated using dataset 91-site-1A from a PV site installed in 2009 with a 10.5 kW array, recording data every 5 min. Due to substantial data gaps from breakdowns and repairs, we selected data from January 1, 2017, to December 31, 2020, for analysis. The dataset was resampled to a 15-minute resolution, resulting in 96 data points per day. A total of 139,341 data points were collected, with each daily sample consisting of 11 features. The Station01 dataset comes from PVOD [42], consisting of data from 10 PV stations located in Hebei Province, China. Given the data length and completeness, Station01 is selected as the primary dataset. This PV station01 recorded data at a 15-minute resolution from July 1, 2018, to June 12, 2019, including 7 locally measured variables, with a total of 33,313 data points. Table 1 provides a detailed description of the local weather data and PV power. The input data are analysed using Pearson correlation analysis, selecting variables with correlation > 0.5 as inputs, with results shown in Fig. 6.

The NWP data used in this study is derived from the Weather Research and Forecasting (WRF) model, which is driven by large-scale forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5. The WRF model extracts NWP variables for horizons ranging from 28 to 54 h, using a grid resolution of 4 km and a 15-minute temporal resolution. Further details regarding the NWP dataset are available in [43]. Considering that the Trina dataset includes specific measured variables rather than NWP variables, such as Current Phase Average (CPA), Radiation Global Tilted (RGT), and Radiation Diffuse Tilted (RDT), we select the common variables between historical data and NWP data for similar day search. The specific input variables are listed in Table 2.

Table 1
Statistical characteristics of each variable in the dataset.

Name	Trina Dataset					station01 Dataset				
	Unit	Max	Min	Mean	Std	Unit	Max	Min	Mean	Std
Active Power	kW	10.10	−0.03	2.009	2.60	MW	19.99	0	3.68	5.56
Current Phase Average	A	19.95	0	4.735	5.82	–	–	–	–	–
Weather Daily Rainfall	mm	36.40	0	0.243	1.79	–	–	–	–	–
Wind Direction	degree	11596.17	−4.47	108.14	136.10	degree	359.90	0	189.93	104.50
Diffuse Horizontal Radiation	W/m^2	769.86	0	51.15	84.59	W/m^2	744.00	0	81.29	118.62
Weather Temperature Celsius	∘C	45.73	−39.99	21.56	11.10	∘C	38.10	−13.10	13.05	11.76
Global Horizontal Radiation	W/m^2	1507.06	0	273.67	370.68	W/m^2	1242.30	0	183.34	271.51
Weather Relative Humidity	%	117.76	0	32.03	20.88	–	–	–	–	–
Radiation Global Tilted	°	1433.83	0.01	313.32	387.53	–	–	–	–	–
Radiation Diffuse Tilted	°	715.17	0.01	55.00	83.48	–	–	–	–	–
Wind Speed	–	–	–	–	–	m/s	11.30	0	1.039	1.22
Atmospheric Pressure	–	–	–	–	–	hPa	1049.10	993.50	1020.19	10.53

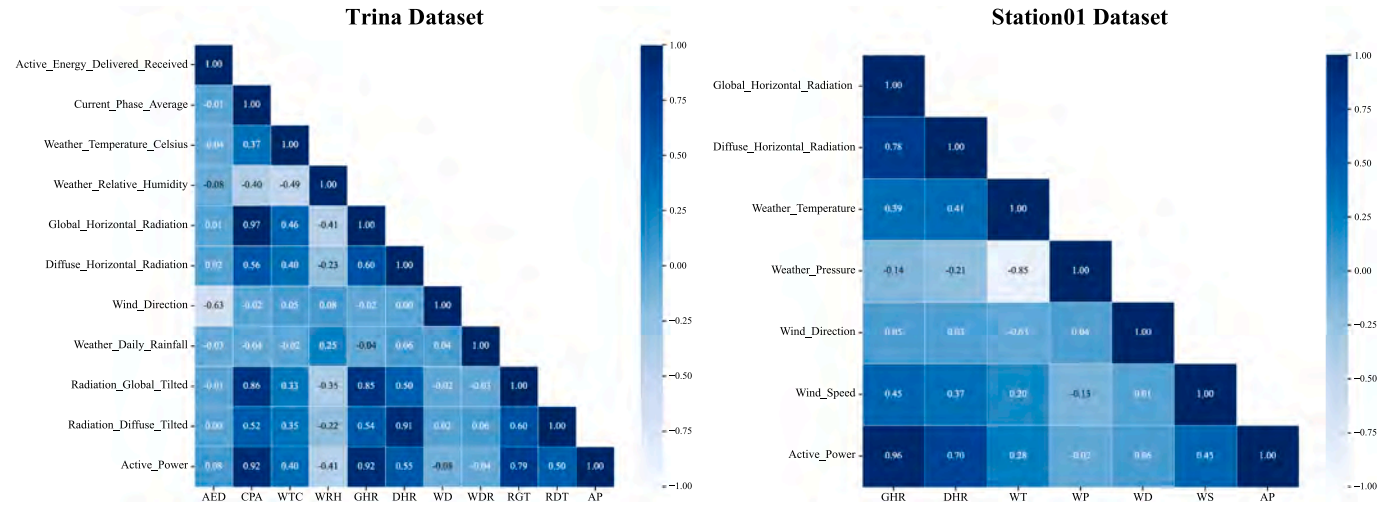

Fig. 6. Correlation analysis of PV power and weather factors.

Table 2
Detailed information on the dataset and model input variables.

Dataset ID	Capacity	Data points	Latitude (degree)	Longitude (degree)	Similar Day Variable	Input Variable
Trina	10.5 kW	139,341	–23.76	133.87	Global Horizontal Radiation	Global Horizontal Radiation
					Diffuse Horizontal Radiation	Diffuse Horizontal Radiation
					10-m Wind Direction (degree)	Radiation Global Tilted
					10-m Temperature (∘C)	Radiation Diffuse Tilted
					10-m Relative Humidity (%)	Current Phase Average
					–	Active Power
Station01	20 MW	33,313	38.18	117.46	Global Horizontal Radiation	Global Horizontal Radiation
					Diffuse Horizontal Radiation	Diffuse Horizontal Radiation
					10-m Wind Direction (degree)	Active Power
					Diffuse Horizontal Radiation	–
					10-m Temperature (∘C)	–
					Atmospheric Pressure (hPa)	–
					10-m Wind speed (m/s)	–

Firstly, the two datasets are pre-processed by using linear interpolation to fill in the missing values and eliminate extreme outliers. Finally, the dataset is divided into training, test, and validation sets in a 7:2:1 ratio. All data is Z-score standardized to scale each variable to a specific range.

$$x_i' = (x_i - \mu)/\delta \tag{14}$$

where x_i' represents the standardized data, μ and δ are the mean and standard deviations of x_i.

3.2. Evaluation metrics

To measure the difference between the true and predicted values, several metrics are used to verify the effectiveness of the proposed model, including Mean Absolute Error (MAE), Mean Squared Error (MSE), coefficient of Determination (R^2) and Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \widetilde{Y}_i| \tag{15}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widetilde{Y}_i)^2 \tag{16}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\widetilde{Y}_i - Y_i}{Y_i}\right| \tag{17}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widetilde{Y}_i)^2} \tag{18}$$

$$R^2 = 1 - \sum_{i=1}^{n}\left(Y_i - \widetilde{Y}_i\right)^2 / \sum_{i=1}^{n}(Y_i - \overline{Y}_i)^2 \tag{19}$$

In these formulas, $Y_i, \widetilde{Y}_i$ represents the true value, predicted value, and n represents the number of samples in the dataset. MSE gives more weight to larger errors, which can reflect the dispersion of errors. A lower MSE signifies improved predictive performance of the model, hence its selection as the loss function for model training.

3.3. Comparative models

To demonstrate the MFA model's improved prediction accuracy, we conduct a comparative analysis using the same dataset processing techniques. We compare the proposed method's performance with eight state-of-the-art models: CNN, LSTM, GRU, CNN-LSTM, MLP, Autoformer, Informer, and Transformer. LSTM and GRU, both belonging to the recurrent neural network (RNN) architecture, aim to capture long-term dependencies in sequential data by utilizing memory units and gates for information flow control. This choice allows us to contrast the predictive performance between serial and parallel architectures. CNN-LSTM is a hybrid architecture combining the strengths of CNN and LSTM. MLP consists of a multi-layer hidden structure. Autoformer, Transformer and Informer are sequence-to-sequence models based on the parallel architecture of the Transformer. This comparison aims to

Table 3
The architectures and hyperparameters of MFA model and baselines.

Model	CNN	LSTM	GRU	CNN-LSTM	MLP	Autoformer	Informer	Transformer	Proposed
Number of hidden nodes	64	64	64	128	256	512	512	512	512
Structure	2 conv layers	2 LSTM layers	2 GRU layers	2 conv layers and 2 LSTM layers	4 linear layers	2 e_layers	2 e_layers	2 e_layers 1 d_layers	2 e_layers 1 d_layers
Kernel size	2x2	–	–	2x2	–	–	–	–	–
Batch size	32	32	32	32	32	32	32	32	32
Number of heads	–	–	–	–	–	8	8	8	8
Dropout rate	0.1	0.1	0.1	0.1	–	0.05	0.05	0.05	0.05
Seq Len	–	–	–	–	–	96	96	96	96
Pred Len	–	–	–	–	–	48	48	48	48

Table 4
Comparative analysis of several PV power models with multiple prediction steps on Trina dataset.

Model	Horizon = 16			Horizon = 32			Horizon = 48			Horizon = 64			Horizon = 96		
	MSE	MAE	R^2	MSE	MAE	R^2	MSE	MAE	R^2	MSE	MAE	R^2	MSE	MAE	R^2
CNN	0.089	0.151	0.932	0.108	0.177	0.918	0.118	0.188	0.911	0.123	0.195	0.907	0.131	0.211	0.900
LSTM	0.104	0.172	0.921	0.116	0.188	0.912	0.128	0.196	0.903	0.135	0.208	0.898	0.142	0.219	0.892
GRU	0.133	0.210	0.900	0.148	0.237	0.888	0.149	0.233	0.887	0.144	0.229	0.891	0.145	0.224	0.890
CNN-LSTM	0.099	0.158	0.925	0.111	0.174	0.916	0.120	0.181	0.909	0.129	0.196	0.902	0.138	0.201	0.895
MLP	0.089	**0.146**	0.933	0.105	0.166	0.920	0.116	0.180	0.912	0.122	0.189	0.907	0.130	0.202	0.901
Autoformer	0.187	0.299	0.858	0.258	0.316	0.804	0.212	0.289	0.839	0.231	0.314	0.825	0.367	0.426	0.722
Informer	0.090	0.169	0.931	0.106	0.172	0.919	0.128	0.175	0.903	0.134	0.182	0.899	0.134	0.192	0.899
Transformer	0.093	0.149	0.929	0.110	0.191	0.916	0.137	0.171	0.896	0.141	0.192	0.893	0.154	0.184	0.883
Proposed	**0.045**	0.153	**0.974**	**0.053**	**0.145**	**0.959**	**0.055**	**0.157**	**0.958**	**0.041**	**0.122**	**0.969**	**0.038**	**0.121**	**0.971**

Table 5
Comparative analysis of several PV power models with multiple prediction steps on Station01 dataset.

Model	Horizon = 16			Horizon = 32			Horizon = 48			Horizon = 64			Horizon = 96		
	MSE	MAE	R^2	MSE	MAE	R^2	MSE	MAE	R^2	MSE	MAE	R^2	MSE	MAE	R^2
CNN	0.176	0.243	0.865	0.231	0.293	0.823	0.257	0.317	0.804	0.273	0.334	0.791	0.287	0.347	0.780
LSTM	0.209	0.271	0.839	0.309	0.360	0.763	0.363	0.381	0.726	0.358	0.381	0.756	0.339	0.377	0.740
GRU	0.214	0.271	0.836	0.332	0.357	0.746	0.345	0.370	0.736	0.364	0.387	0.721	0.362	0.392	0.722
CNN-LSTM	0.214	0.250	0.836	0.323	0.329	0.753	0.304	0.325	0.767	0.318	0.336	0.757	0.283	0.327	0.783
MLP	0.163	0.219	0.875	0.222	0.265	0.829	0.244	**0.289**	0.813	0.259	0.303	0.802	0.280	0.321	0.785
Autoformer	0.283	0.391	0.783	0.404	0.478	0.690	0.449	0.511	0.656	0.445	0.483	0.659	0.715	0.620	0.452
Informer	**0.153**	0.212	**0.883**	0.227	0.251	0.826	0.261	0.297	0.800	0.265	0.292	0.797	0.296	0.316	0.773
Transformer	0.159	**0.203**	0.878	**0.222**	0.262	0.830	0.299	0.304	0.771	0.289	0.318	0.779	0.456	0.368	0.650
Proposed	0.168	0.231	0.864	0.234	**0.248**	**0.833**	**0.238**	0.303	**0.842**	**0.231**	**0.290**	**0.816**	**0.259**	**0.301**	**0.835**

validate the effectiveness of different model architectures. The detailed experiment hyperparameters for the proposed method and baseline models are presented in Table 3.

4. Results and discussion

In this section, the experimental results comparing the proposed method with the baseline are analyzed. Secondly, similar day selection, ablation study, and interpretability analysis are carried out. Finally, the advantages and disadvantages of the proposed method are discussed.

4.1. Comparison of prediction performance of models

To demonstrate the stability of the proposed method in multi-step forecasting, we evaluate its prediction performance and compare it with baselines at different forecasting steps: 16 steps (4 h), 32 steps (8 h), 48 steps (12 h), 64 steps (16 h), and 96 steps (24 h). We take the average of 5 experimental results as the outcome, as illustrated in Table 4 (Trina dataset) and Table 5 (Station01 dataset). In short-term forecasting with a 4-hour lead time, CNN and RNN-based models effectively establish relationships between input features and forecasted values. As the horizon increases, these models show limitations in capturing long-term dependencies, leading to the accumulation of errors. The Autoformer, by learning the cyclical variation patterns of sequences, shows a notable reduction in errors at the 32 to 48-step forecast horizons on the Trina dataset.

However, it does not account for the randomness and meteorological dependency characteristic of the PV series, reducing its advantage in solar power prediction. The proposed method consistently maintains the lowest prediction error starting from the 32-step forecast, showing stable performance without error accumulation. Analyzing the performance of the MFA-model on both datasets, the overall prediction error is smaller on the Trina dataset. This is because Station01 does not benefit from the same sunlight advantages due to its geographical location compared to Trina. Station01 has higher mean and standard deviation values for active power and diffuse irradiance, leading to lower prediction accuracy for both the MFA model and baselines. Nevertheless, the proposed method surpasses all other baseline models across various metrics. On the Station01 dataset, MAE, MSE, and MAPE are reduced by 7.5 % (0.280 → 0.259), 4.7 % (0.316 → 0.301), and 39 % (2.367 → 2.189), while R^2 improves by 6.3 % (0.785 → 0.835), achieving 10 optimal performance metrics. This highlights its robustness and generalization capacity in long-sequence time-series forecasting.

Figs. 7-8 illustrate the prediction error variations of different models at various forecast horizons on both datasets. We further analyze the relative error and scatter plots for the top three models with the lowest

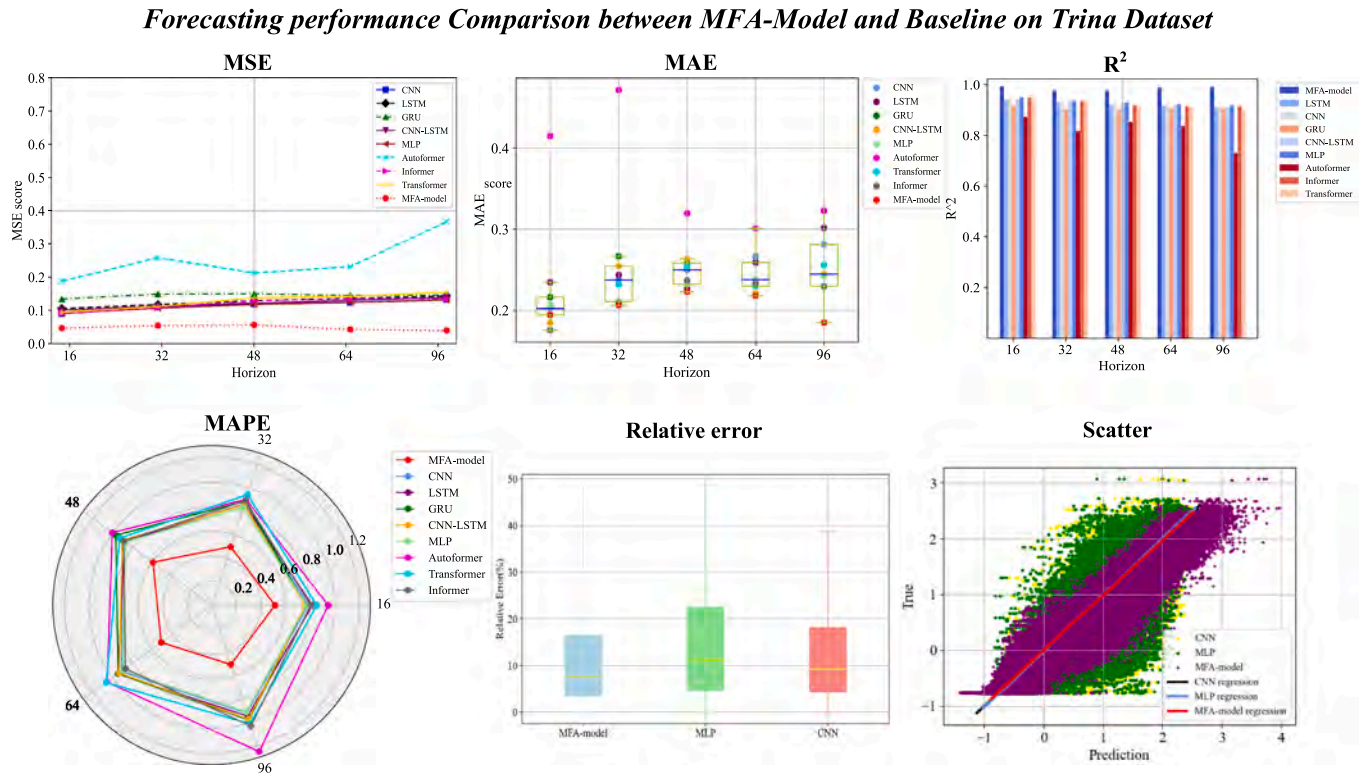

Fig. 7. Comparison of prediction performance of different models on Trina dataset.

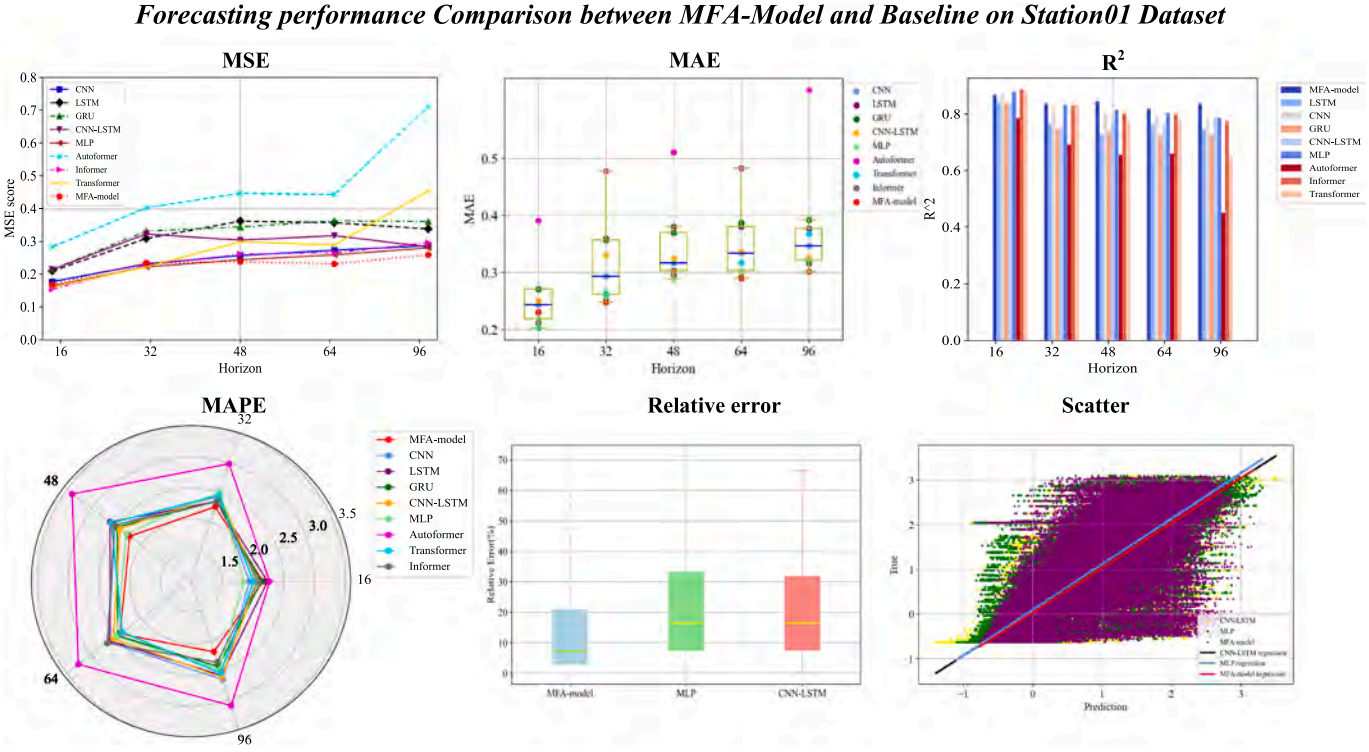

Fig. 8. Comparison of prediction performance of different models on station01 dataset.

MSE loss at the 96-step forecast horizon. On the Trina dataset, the top three models exhibit results close to the line y = x. On the Station01 dataset, the models with the lowest MSE are the MFA-model, CNN-LSTM, and MLP. On both datasets, the MLP model tends to underestimate future PV power output, with its regression line tilting to the left. The scatter points for the proposed method are more concentrated around the actual and predicted values, indicating smaller relative error and less fluctuation, consistent with its relative error performance. The linear regression results are summarized in Table 6.

Figs. 9-10 illustrate the predictive performance between baselines

Table 6
Statistical parameters of the three top models.

Dataset	Model	Const	Coeff	P value	Standard Error
Trina	MFA	−0.0005	0.9896	0.000	1.035e-04
	MLP	−0.0137	1.0173	0.000	1.037e-04
	CNN	−0.0199	0.9753	0.000	1.089e-04
Station01	MFA	0.0357	1.0113	0.000	1.827e-04
	MLP	0.0969	1.0285	0.000	1.748e-04
	CNN-LSTM	0.0779	1.0257	0.000	1.731 e-04

and the proposed model across forecast horizons of 32, 48, 64, and 96 steps. In Fig. 9, for the 32-step forecast, the proposed model exhibits significant errors compared to the actual values. The gap between the predicted and true values in the mid-range forecasts indicates a weakness in the MFA-model's attention to global information. This issue can be traced back to the non-periodic length of the forecast steps, hindering the ability of MFA-attention to recognize periodic and seasonal trends. This limitation increases the likelihood of repeated selection of similar days. However, as the forecast length increases, the model gradually improves its ability to capture global information, approaching the ground truth curve. In contrast, the baseline models merely repeat previous rising and falling patterns. The Trina dataset, with more turning points on clear days, makes it easier to find effective similar days. In contrast, Station01 dataset's more complex fluctuations limit the MFA-model's ability to capture reliable volatility. As seen in Fig. 10, although transformer-based baseline models appear closer to the actual curve, they fail to capture local fluctuations in PV power, resulting in predictions that average out upward and downward fluctuations to reduce error scores. In contrast, the proposed method effectively captures short-term ramp events, highlighted by red circles. This underscores the effectiveness of MFA-attention in detecting volatility and trend shifts across different segment lengths, thereby improving forecast accuracy.

4.2. Similar days analysis

We compare the predictive performance at 96 steps using similar days as input in the Decoders of parallel models such as Transformer, Informer, and Autoformer. The model that integrates similar days based on contrastive learning is denoted as "Ts2vec-s," while the model utilizing DTW is denoted as "dtw-s." The results, presented in Table 7, show that adding similar day information improves the prediction accuracy of various Transformer-based models. Notably, the "Ts2vec-s" model demonstrates varying degrees of improvement in MSE and MAE metrics compared to "dtw-s." Among these models, Autoformer achieves the most significant enhancement on the Station01 dataset, with a 5.2 % reduction in MSE. However, on the Trina dataset, where the PV power curves are smoother and exhibit fewer fluctuations, the improvement from adding similar day information is relatively limited.

We further analyze the prediction results on the Station01 dataset. Fig. 11 compares the predictions of the MFA model with the baseline, both using the same NWP data as input. All models utilize the same similar day selection based on contrastive learning, driven by NWP variables. However, within the models, GHI and DNI from the NWP data are the primary factors influencing prediction accuracy. Compared to traditional search methods, the similar day selection via contrastive learning aligns more closely with key time points of model fluctuations, such as at timesteps 48 and 62. Due to lower GHI values, all models tend

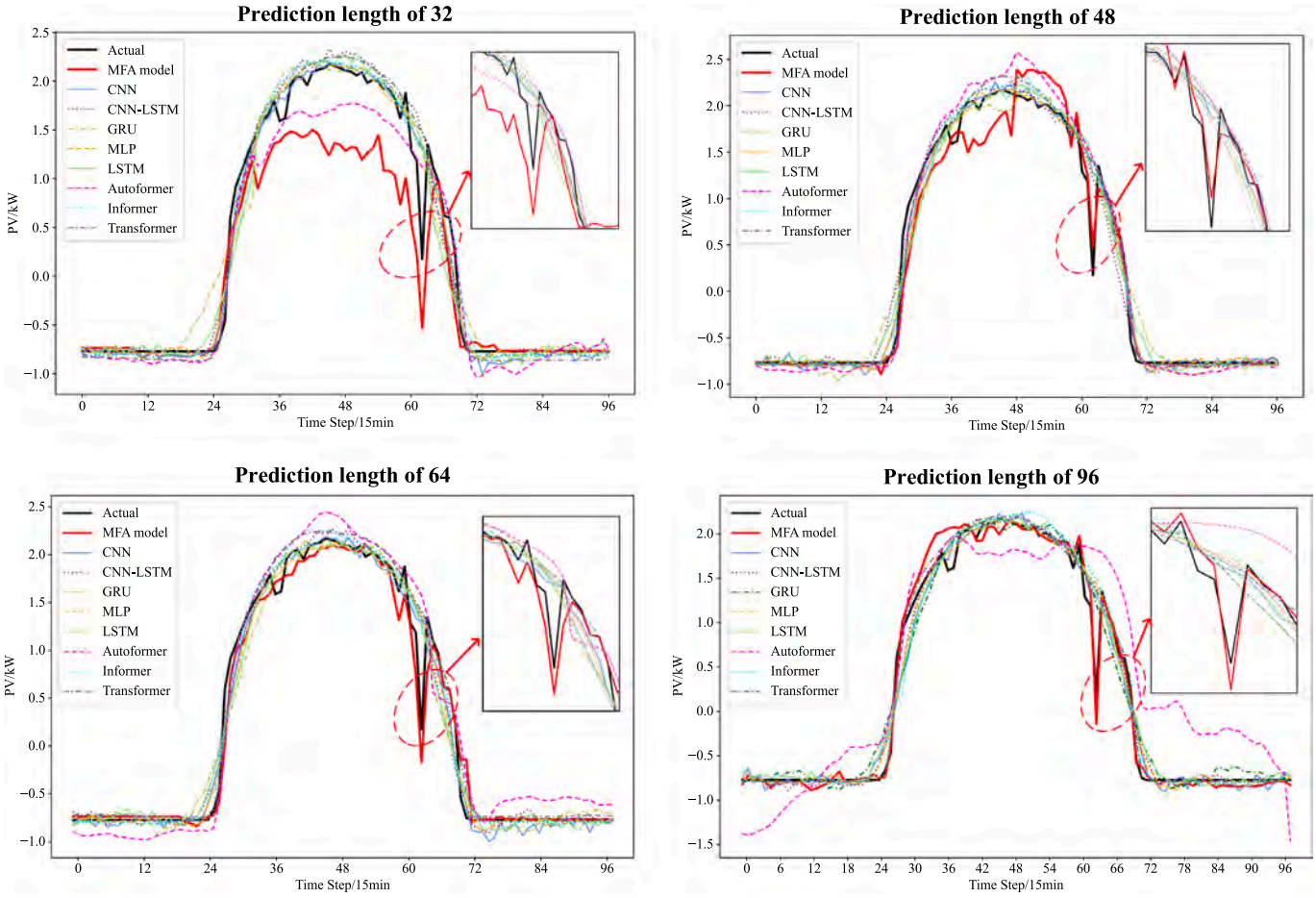

Fig. 9. Prediction curves for different forecast horizons on the Trina dataset.

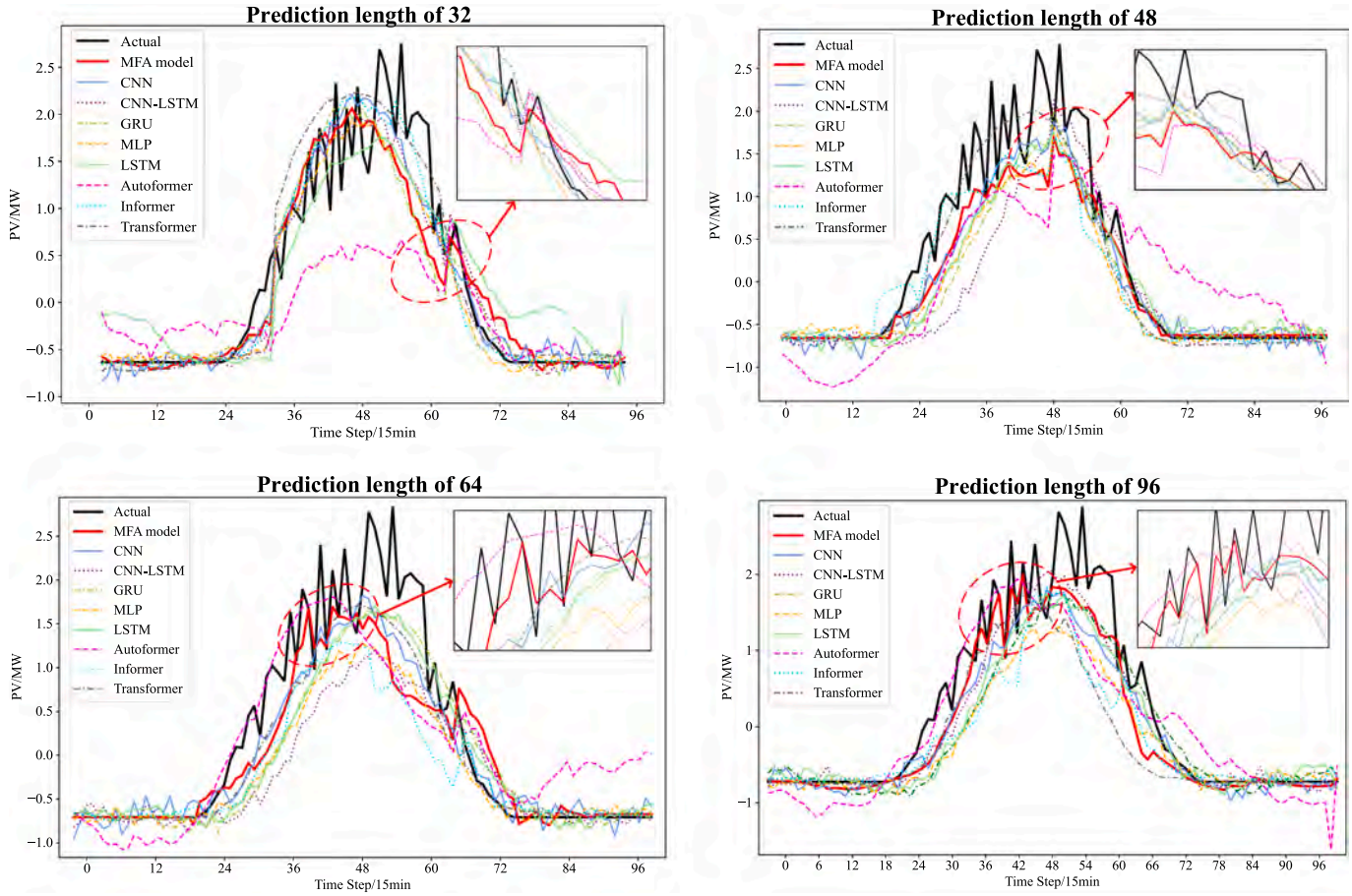

Fig. 10. Prediction curves for different forecast horizons on the Station01 dataset.

Table 7
Model Performance on Similar Days Selected by Contrastive Learning and DTW.

Model	Metric	Trina (dtw-s)	Trina (Ts2vec-s)	Station01 (dtw-s)	Station01 (Ts2vec-s)
Autoformer	MSE	0.273	**0.205**	0.652	**0.618**
	MAE	0.353	**0.301**	0.623	**0.588**
Informer	MSE	0.051	**0.047**	0.308	**0.281**
	MAE	0.138	**0.129**	0.337	**0.302**
Transformer	MSE	0.057	**0.046**	0.423	**0.377**
	MAE	0.148	**0.127**	0.364	**0.346**
Proposed	MSE	0.048	**0.038**	0.324	**0.259**
	MAE	0.133	**0.121**	0.345	**0.301**

to underpredict compared to the actual values, but the MFA model better captures these fluctuations and effectively leverages both NWP data and historical information. Conversely, Autoformer performs less effectively on both datasets, potentially due to its insufficient focus on the seasonal component, which may result in the loss of critical information. As PV power generation is particularly sensitive to weather factors, this remains an area warranting further investigation.

4.3. Ablation study

The performance of MFA attention: We add the MFA module to several baseline models of CNN, LSTM, GRU and CNN-LSTM respectively. We conduct experiments under the input-96-predict-96 setting, and the parameters are set to (epochs = 5, heads = 8, dim = 64). All experiments are repeated five times, and the average of the results of 5 experiments is taken as the result.

As shown in Table 8, "+" indicates that the MFA-attention module is added. The proposed MFA-attention achieves the best performance for most baseline models, The prediction accuracy of baseline models is improved. Serial models like LSTM and GRU show the most significant improvements, with the GRU model's MSE reduced by 3.87 %, MAE by 12.76 %, and RMSE by 2.16 %. MFA-attention enhances the serial models' ability to capture local features, especially on the highly fluctuating Station01 dataset. However, for CNN, which excels at capturing short-term dependencies, MFA-attention may disrupt the rhythm of its receptive field, leading to limited improvement. These advancements are inherently limited by the structural characteristics of each model, which define their capability to process complex, nonlinear, long-term dependencies. Consequently, there is an upper limit to how much the proposed method can optimize the prediction accuracy of existing models.

4.4. Interpretability analysis

MFA-model stacks two layers of the Encoder, and the attention matrix obtained by averaging the attention across eight heads is shown in Fig. 12. In Fig. 12 (a), the first-layer matrix displays more significant weights on the right edge, indicating the model's enhanced focus on the sequence's final part. This enables the capturing of local features and initial change patterns. The second-layer MFA-attention's score matrix, as shown in Fig. 12 (b), emphasizes the diagonal in its score matrix, reflecting self-attention's autocorrelation trait. Here, both Query and Key are derived from a linear transformation of the input sequence, essentially comparing different representations of the same sequence. The pronounced diagonal weights suggest a heightened focus on each timestep's information and its immediate neighbors in this layer, indicating that the second-layer MFA-attention has learned more complex

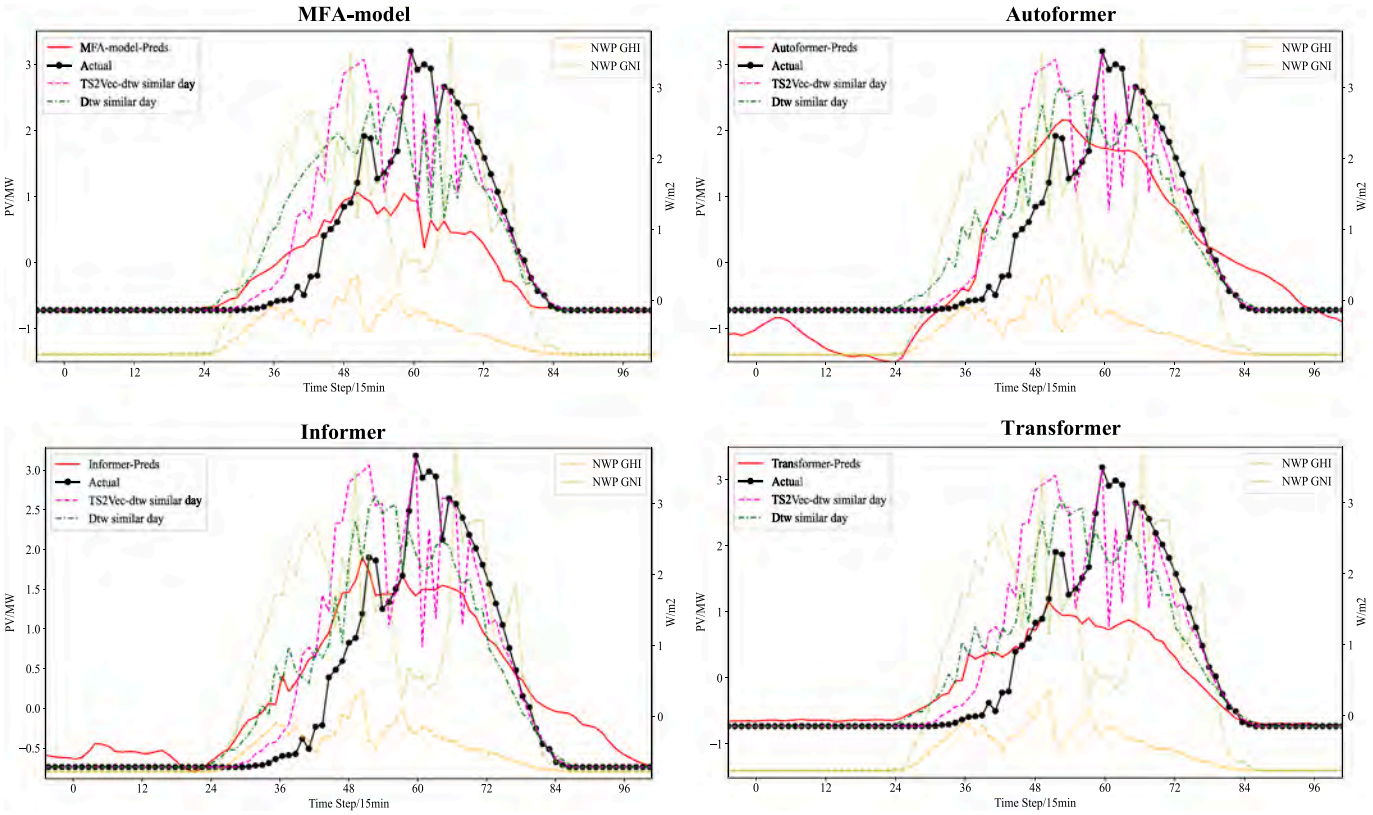

Fig. 11. Similar day search comparison based on contrast learning.

Table 8
Ablation study of the MFA-attention.

Model	Trina				Station01			
	MSE	MAE	RMSE	MAPE	MSE	MAE	RMSE	MAPE
GRU	0.144	0.224	0.381	0.979	0.362	0.392	0.602	2.441
GRU^{+}	**0.135**	**0.191**	**0.367**	**0.971**	**0.348**	**0.342**	**0.589**	**2.410**
LSTM	0.142	0.219	0.377	**0.946**	0.339	0.377	0.583	2.586
LSTM^{+}	**0.139**	**0.203**	**0.374**	0.951	**0.326**	**0.336**	**0.572**	**2.501**
CNN	0.131	0.211	0.362	0.919	0.287	**0.347**	**0.535**	2.658
CNN^{+}	**0.130**	**0.198**	**0.354**	**0.896**	**0.284**	0.359	0.553	**2.621**
CNN-LSTM	0.138	**0.201**	0.372	1.079	0.283	0.327	**0.532**	**2.620**
CNN-LSTM^{+}	**0.136**	0.203	**0.369**	**0.971**	**0.281**	**0.306**	0.551	2.631

self-dependency patterns within the sequence compared to the first layer.

Extracting all time-delay positions calculated in the second layer of MFA-attention in the Decoder, as illustrated in Fig. 13, reveals significant insights. Given the Decoder's input vector $\mathscr{X}_{des} \in R^{(L/2+h)\times d}$ includes half a cycle's adjacent data, the MFA-model learns lags encompassing semi-diurnal and diurnal periods. This demonstrates the MFA-attention's effectiveness in aggregating the periodic and seasonal fluctuations of PV sequences, thereby enhancing the model's interpretability by reflecting these essential characteristics.

4.5. Decomposition visualization

A detailed visualization analysis of the decomposition modules is conducted, as shown in Fig. 14. With the model input set to 96 and the prediction horizon set to 384, the impact of the trend and seasonal components on long-term forecasting is observed. By gradually adding decomposition blocks from 0 to 3 in the decoder, it is evident that the trend part governs the magnitude and general direction of the forecast, while the seasonal part captures more specific fluctuations. A single decomposition block is insufficient to fully capture the exact time intervals of seasonal variations. However, as the number of decomposition blocks increases, the forecasted values gradually converge to the true range and magnitude. This demonstrates the effectiveness of the decomposition blocks in capturing temporal information accuracy.

4.6. Discussion

In this subsection, we analyze the benefits and limitations of the proposed approach by examining the internal structure of the baseline model. The serial structure of RNN models limits their capacity to capture long-term dependencies and disregards relationships between variables. As the prediction horizon increases, prediction errors also increase. In contrast, Transformer models, which lack recursion, output predictions for a given time window simultaneously. However, their dot-attention mechanism does not consider the temporal characteristics between input sequences, leading to inferior performance metrics compared to RNN models in PV forecasting tasks. The MFA model makes use of the autoregressive property of time series to refine attention calculations at different time granularities. It considers both the

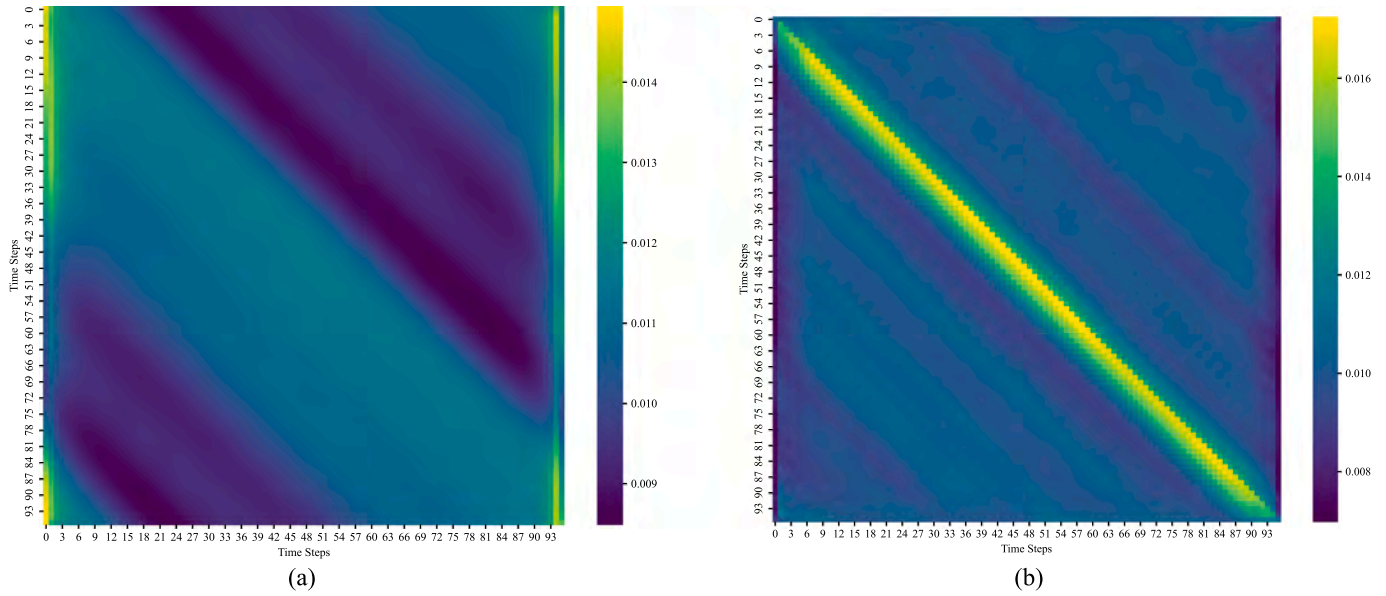

Fig. 12. The attention matrix obtained by averaging the attention across eight heads (a)Encoder stack 1; (b)Encoder stack 2.

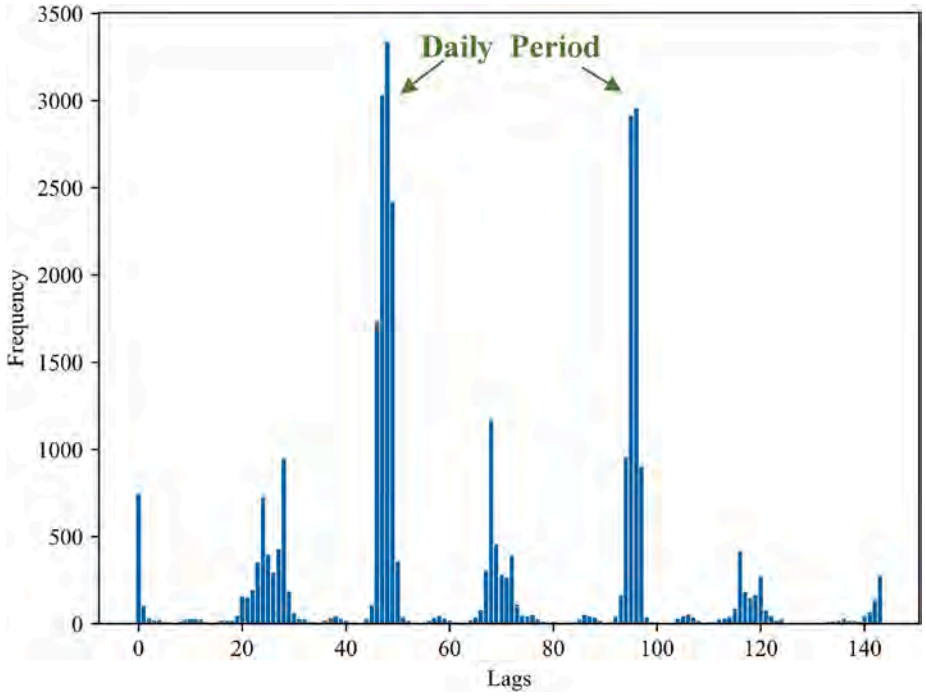

Fig. 13. Lag statistics in the Decoder's second MFA-Attention layer.

temporal property and the spatial characteristics between variables, thereby improving the model's sensitivity to local fluctuations.

Furthermore, in short to medium-term PV forecasting, models primarily rely on historical data patterns to infer future predictions, making it challenging to account for abrupt changes in PV power generation during the prediction process. The proposed similar day data has been proven effective, as it can provide crucial information to help the model effectively identify peak changes. However, the model may overly rely on similar day information, leading to inaccuracies when historical data is insufficient or extreme weather conditions frequently occur at PV sites.

5. Conclusion

Accurate identification of seasonal and trend fluctuations is pivotal for short-term PV power forecasting. A novel PV power forecasting method that aggregates multi-timescale fluctuation correlations is proposed in this paper. The development of the MFA-attention mechanism maps the interplay between solar power variations and meteorological factors over diverse timescales. The proposed self-supervised similar day selecting technique provides the model with effective input and better detects local dynamic changes. The proposed multi-step forecasting approach achieves higher accuracy in predictions across different forecast horizons, effectively mitigating error accumulation seen in conventional models. The proposed method is benchmarked against eight

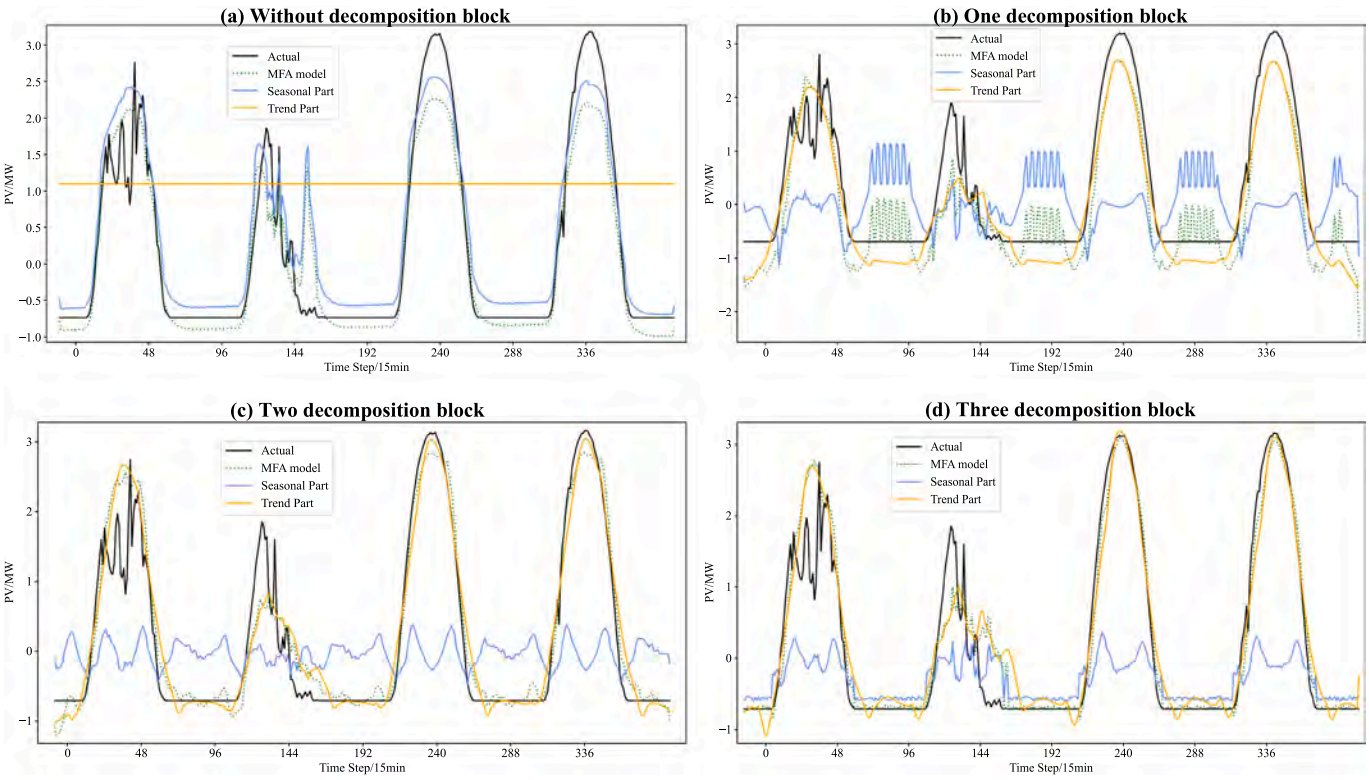

Fig. 14. Decomposition blocks visualization of the final layer decoder of the station01 dataset.

renowned models through comprehensive case studies. The results highlight the MFA model's superior predictive performance, notably achieving a 39 % increase in forecasting accuracy for 96-step predictions. Overall, the proposed model excels in handling complex meteorological scenarios, presenting a robust framework for multistep solar power forecasting.

CRediT authorship contribution statement

Xiangting Wang: Writing – original draft, Data curation. **Liang Yuan:** Writing – review & editing. **Yao Sun:** Methodology, Conceptualization. **Xubin Liu:** Visualization. **Zhao Yang Dong:** Supervision, Project administration.

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.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62125308, 52307156 and 52337008.

Appendix A

A.1. Algorithm of MFA-model procedure

.

Algorithm 1 Overall MFA-model Procedure

Input: Historical PV time series $X_{t-L+1:t}$; Input Length L; Predict length h; Data dimension d; Hidden state channel d_k; Encoder layers number $N = 2$; Decoder layers number $M = 1$; Moving average window size k. We set d_k as 512, k as 25.

$\mathscr{X}_s, \mathscr{X}_t = SeriesDecomp(X_{\frac{t-L+1}{2}:t}) \triangleright X \in \mathbb{R}^{L\times d}, \mathscr{X}_s, \mathscr{X}_t \in \mathbb{R}^{\frac{L}{2}\times d}$

$S_{t+1:t+h}, X_{mean} = TS2Vec - DTW(X_{t+1:t+h}), \text{Repeat}(Mean(X_{\frac{t-L+1}{2}:t}, \text{dim} = 0), \text{dim} = 0) \triangleright S_{t+1:t+h}, X_{mean} \in \mathbb{R}^{h\times d}$

$\mathscr{X}_{des}, \mathscr{X}_{\text{det}} = Concat(\mathscr{X}_s, S_{t+1:t+h}), Concat(\mathscr{X}_t, X_{mean}) \triangleright \mathscr{X}_{des}, \mathscr{X}_{\text{det}} \in \mathbb{R}^{(\frac{L}{2}+h)\times d}$

$\mathscr{X}^0_{\text{en}} = Embed(X_{t-L+1:t}) \triangleright \mathscr{X}^0_{\text{en}} \in \mathbb{R}^{L\times d_k}$

for n **in** {1,…,N}: ▷ MFA-model Encoder

$\mathscr{S}^{n,1}_{\text{en}}, \_ = SeriesDecomp(\text{MF - } Aggregation(\mathscr{X}^{n-1}_{\text{en}}) + \mathscr{X}^{n-1}_{\text{en}}) \triangleright \mathscr{S}^{n,1}_{\text{en}} \in \mathbb{R}^{L\times d_k}$

$\mathscr{S}^{n,2}_{\text{en}}, \_ = SeriesDecomp(\text{FeedForward}(\mathscr{S}^{n,1}_{\text{en}}) + \mathscr{S}^{n,1}_{\text{en}}) \triangleright \mathscr{S}^{n,2}_{\text{en}} \in \mathbb{R}^{L\times d_k}$

$\mathscr{X}^n_{\text{en}} = \mathscr{S}^{n,2}_{\text{en}} \triangleright \mathscr{X}^n_{\text{en}} \in \mathbb{R}^{L\times d_k}$

end for

(*continued on next page*)

(*continued*)

Algorithm 1 Overall MFA-model Procedure

$\mathscr{X}_{de}^{0} = Embed(\mathscr{X}_{des}), \mathscr{T}_{\text{de}}^{0} = \mathscr{X}_{\text{det}} \triangleright \mathscr{X}_{de}^{0} \in \mathbb{R}^{(\frac{L}{2}+h)\times d_k}, \mathscr{T}_{\text{de}}^{0} \in \mathbb{R}^{(\frac{L}{2}+h)\times d}$

for n **in** {1,…,*M*}**:** ▷ MFA-model Decoder

$\mathscr{S}_{\text{de}}^{n,1}, \mathscr{T}_{\text{de}}^{n,1} = SeriesDecomp\left(\text{MF - } Aggregation\left(\mathscr{X}_{de}^{n-1}\right) + \mathscr{X}_{de}^{n-1}\right)$

$\mathscr{S}_{\text{de}}^{n,2}, \mathscr{T}_{\text{de}}^{n,2} = SeriesDecomp\left(\text{MF - } Aggregation\left(\mathscr{S}_{\text{de}}^{n,l}, \mathscr{X}_{\text{en}}^{N}\right) + \mathscr{S}_{\text{de}}^{n,l}\right)$

$\mathscr{S}_{\text{de}}^{n,3}, \mathscr{T}_{\text{de}}^{n,3} = SeriesDecomp\left(\text{FeedForward}\left(\mathscr{S}_{\text{de}}^{n,2}\right) + \mathscr{S}_{\text{de}}^{n,2}\right) \triangleright \mathscr{S}_{\text{de}}^{n,\cdot}, \mathscr{T}_{\text{de}}^{n,\cdot} \in \mathbb{R}^{(\frac{L}{2}+h)\times d_k}$

$\mathscr{T}_{\text{de}}^{n} = \mathscr{T}_{\text{de}}^{n-1} + MLP(\mathscr{T}_{\text{de}}^{n,1}) + MLP(\mathscr{T}_{\text{de}}^{n,2}) + MLP(\mathscr{T}_{\text{de}}^{n,3}) \triangleright \mathscr{T}_{\text{de}}^{n} \in \mathbb{R}^{(\frac{L}{2}+h)\times d}$

$\mathscr{X}_{de}^{n} = \mathscr{S}_{\text{de}}^{n,3} \triangleright \mathscr{X}_{de}^{n} \in \mathbb{R}^{(\frac{L}{2}+h)\times d_k}$

end for

$\widetilde{Y} = \mathscr{T}_{\text{de}}^{n} + MLP(\mathscr{X}_{de}^{n}) \triangleright \widetilde{Y} \in \mathbb{R}^{(\frac{L}{2}+h)\times d}$

Return $\widetilde{Y}_{t+1:t+h}$

A.2. Algorithm of the multiscale fluctuate aggregation

.

Algorithm 2 Multiscale Fluctuate Aggregation

Input: Input data series X_t, Input Length L; Predict length h;Data dimension d_k;Hyper-parameter q; We set d_k as 512, q as 3.

$Q, K, V = \text{Re}shape(MLP(X_t)) \triangleright Q, K, V \in \mathbb{R}^{B\times L\times h\times \frac{d_k}{h}}$

$Q = FFT(Q, \dim = 0), K = FFT(K, \dim = 0) \triangleright Q, K \in \mathbb{C}^{B\times L\times h\times \frac{d_k}{h}}$

$\mathscr{R}_{corr} = IFFT(Q \times conj(K), \dim = 0) \triangleright \mathscr{R}_{corr} \in \mathbb{R}^{B\times L\times h\times \frac{d_k}{h}}$

$\mathscr{R}_{corr}(\tau), \tau = Topm([q \times \frac{\log L}{2}], \mathscr{R}_{corr}, \dim = 0) \triangleright \mathscr{R}_{corr}(\tau) \in \mathbb{R}^{B\times \tau_m}$

$\overset{\wedge}{\mathscr{R}}_{corr}(\tau_1), \ldots, \overset{\wedge}{\mathscr{R}}_{corr}(\tau_m) = \text{Softmax}[\mathscr{R}_{corr}(\tau_1), \ldots, \mathscr{R}_{corr}(\tau_m)]$

for i **in** range $[q \times \frac{\log L}{2}]$**:** ▷ MFA- Aggregation

$f_i = \text{Softmax}(\frac{Q_i K_i^T}{d_k}) V_i \triangleright f_i \in \mathbb{R}^{B\times \tau_m\times h\times \frac{d_k}{h}}$

$f_i = Padding(f_i, \dim = 1) \triangleright f_i \in \mathbb{R}^{B\times L\times h\times \frac{d_k}{h}}$

end for

$\overset{\wedge}{\mathscr{R}}_{corr}(\tau_i) = \text{Re}peat(\overset{\wedge}{\mathscr{R}}_{corr}(\tau_i)) \triangleright \overset{\wedge}{\mathscr{R}}_{corr}(\tau_i) \in \mathbb{R}^{B\times h\times \frac{d_k}{h}\times L}$

$R = \sum_{i=1}^{m} f_i \overset{\wedge}{\mathscr{R}}_{corr}(\tau_i) \triangleright R \in \mathbb{R}^{B\times L\times h\times \frac{d_k}{h}}$

Return R

A.3. Algorithm of the selection of similar days

.

Algorithm 3 Selection of similar days

Input: Historical PV time series $X_{t-L+1:t}$; Input Length L; All Input Length $L_{\max}$;Data dimension d; Hidden state channel d_s; Time feature d_{time};We set d_s as 25, d_{time} as 7.

Procedure TS2vec model

$X_{dt} = TimeEmbed(X_{t-L+1:t}) \triangleright X_{dt} \in \mathbb{R}^{B\times d_{time}}$

$X_{dt} = concat(X_{t-L+1:t}, X_{dt}) \triangleright X_{dt} \in \mathbb{R}^{B\times (d_{time}+d)}$

$r_{i,t}, r'_{i,t} = TSEncoder(X_{dt}[a_1, b_1], X_{dt}[a_2, b_2]) \triangleright r_{i,t}, r'_{i,t} \in \mathbb{R}^{B\times d_s}$

Procedure Contrastive Loss $\left(r_{i,t}, r'_{i,t}\right)$

$\mathscr{L}_{\text{hier}} = \mathscr{L}_{\text{dual}}\left(r_{i,t}, r'_{i,t}\right)$

$d = 1$

While*time_length*$(h_1) > 1$ **do**

$r_{i,t} = \text{max}pool1d(r_{i,t}, \text{ker}nel\_size = 2)$

$r'_{i,t} = \text{max}pool1d(r'_{i,t}, \text{ker}nel\_size = 2)$

$\mathscr{L}_{\text{hier}} = \mathscr{L}_{\text{hier}} + \mathscr{L}_{\text{dual}}\left(r_{i,t}, r'_{i,t}\right)$

$d = d + 1$

end while

$\mathscr{L}_{\text{hier}} = \mathscr{L}_{\text{hier}}/d$

return $\mathscr{L}_{\text{hier}}$

Return save TS2vecmodel

(*continued on next page*)

(*continued*)

Algorithm 3 Selection of similar days

16: **For** i **in** range[index(X_{t+1})-30*h]:
17: $repr = TS2vecmodel(X) \triangleright repr \in \mathbb{R}^{L_{max} \times d_s}$
18: $H_{t+1:t+h} = repr[index : index + h] \triangleright H_{t+1:t+h} \in \mathbb{R}^{h \times d_s}$
19: $seq = repr[i : i + h] \triangleright seq \in \mathbb{R}^{h \times d_s}$
20: $score = fastdtw(H_{t+1:t+h}, seq)$
21: **end for**
22: $pre\_index = \min(score)$
23: **Return** $S_{t+1:t+h} = concat(X[pre\_index, pre\_index + h], X'_{t+1:t+h}) \triangleright X'_{t+1:t+h} \in \mathbb{R}^{h \times (d-1)}, S_{t+1:t+h} \in \mathbb{R}^{h \times d}$

Appendix B

This section describes the contrastive loss principle of the TS2Vec model.

TS2Vec contains two loss functions in the instances and time dimensions. It can learn to distinguish the multi-scale contextual information with different granularities. In instance-wise, Random cropping is used to turn the input time series into subseries $[a_1, b_1]$ and subseries $[a_2, b_2]$such that $0 < a_1 \leq a_2 \leq b_1 \leq b_2 \leq T$ as positive pairs, and to ensure that the two segments have random length overlapped segments $[a_2, b_1]$, while other time series become negative samples. The instance-wise contrastive loss can be formulated as:

$$\ell_{inst}^{(i,t)} = -\log \frac{\exp\left(r_{i,t} \cdot r'_{i,t}\right)}{\sum_{j=1}^{B}\left(\exp\left(r_{i,t} \cdot r'_{j,t}\right) + \mathrm{I}_{[i \neq j]} \exp\left(r_{i,t} \cdot r_{j,t}\right)\right)} \tag{B.1}$$

where B is the batch size, (i, t) donates the i-th time series at timestamp t. $r_{i,t}$ and $r'_{i,t}$ represents the positive pairs.

In terms of time, TS2Vec takes representations of two positive samples of the input time series at the same timestamp as positives, and representations at different timestamps as negatives. the temporal contrastive loss is given in (2).

$$\ell_{\text{temp}}^{(i,t)} = -\log \frac{\exp\left(r_{i,t} \cdot r'_{i,t}\right)}{\sum_{t' \in \Omega}\left(\exp\left(r_{i,t} \cdot r'_{i,t'}\right) + \mathrm{I}_{[t \neq t']} \exp\left(r_{i,t} \cdot r_{i,t'}\right)\right)} \tag{B.2}$$

where Ω is the set of timestamps within the overlapped segment. The overall loss function is the sum of the two contrast losses. The overall loss is defined as:

$$\mathscr{L}_{\text{dual}} = \frac{1}{NT} \sum_{i} \sum_{t} \left(\ell_{\text{temp}}^{(i,t)} + \ell_{\text{inst}}^{(i,t)}\right) \tag{B.3}$$

The contextual features within the overlapped Segment should be consistent. In this way, the model can minimize loss by learning the contextual information, rather than the absolute position. For any subseries, the overall representation can be obtained by maximizing the pooling of features on the timestamp dimension.

Data availability

Data will be made available on request.

References

[1] Das UK, et al. Forecasting of photovoltaic power generation and model optimization: A review. *Renew Sustain Energy Rev* 2018;81:912–28.
[2] Yu SW, Han RL, Zhang J. Reassessment of the potential for centralized and distributed photovoltaic power generation in China: on a prefecture-level city scale. *Energy* 2023;262:125436.
[3] Song C, Guo Z, Liu Z. Application of photovoltaics on different types of land in China: Opportunities, status and challenges. *Renew Sustain Energy Rev* 2024;191:114146.
[4] IEA. (2023). " Renewable Energy Market Update - June 2023." *International Energy Agency*. Available: https://www.iea.org/reports/renewable-energy-market-update-june-2023.
[5] Mayer MJ. Benefits of physical and machine learning hybridization for photovoltaic power forecasting. *Renew Sustain Energy Rev* 2022;168:112772.
[6] Comello S, Reichelstein S, Sahoo A. The road ahead for solar PV power. Renew *Sustain Energy Rev* 2018;92:744–56.
[7] Sarmas E, Spiliotis E, Stamatopoulos E. Short-term photovoltaic power forecasting using meta-learning and numerical weather prediction independent Long Short-Term Memory models. *Renew Energy* 2023;216:118997.
[8] Khan ZA, Hussain T, Baik SW. Dual stream network with attention mechanism for photovoltaic power forecasting. *Appl Energy* 2023;338:120916.
[9] Ahmed R, Sreeram V, Mishra Y, Arif MD. A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. *Renewable Sustain Energy Rev* 2020;124:109792.
[10] Huang C, Yang M. Memory long and short term time series network for ultra-short-term photovoltaic power forecasting. *Energy* 2023;279:127961.
[11] Tariq L, Reda Y, Khalid B, et al. Accurate one step and multistep forecasting of very short-term PV power using LSTM-TCN model. *Renew Energy* 2023;205:1010–24.
[12] Sobri S, Sam KK, Rahim NA. Solar photovoltaic generation forecasting methods: a review. *Energy Convers Manag* 2018;156:459–97.
[13] Cao Y, Liu G, Luo D, Bavirisetti DP. Multi-timescale photovoltaic power forecasting using an improved Stacking ensemble algorithm based LSTM-Informer model. *Energy* 2023;283:128669.
[14] Zhang Y, Qin C, Srivastava AK, Jin C. Data-Driven Day-Ahead PV Estimation Using Autoencoder-LSTM and Persistence Model. *IEEE Trans Ind Appl* 2020;56(6):7185–92.
[15] Mayer MJ, Gróf G. Extensive comparison of physical models for photovoltaic power forecasting. *Appl Energy* 2021;283:116239.
[16] Book H, Lindfors AV. Site-specific adjustment of a NWP-based photovoltaic production forecast. *Sol Energy* 2020;211:779–88.
[17] Contreras J, Espinola R, Nogales FJ, Conejo AJ. ARIMA models to predict next-day electricity prices. *IEEE Trans Power Syst* 2003;18(3):1014–20.
[18] Sheng H, Xiao J, Cheng Y, Ni Q, Wang S. Short-Term Solar Power Forecasting Based on Weighted Gaussian Process Regression. *IEEE Trans Ind Electron* 2018;65:300–8.
[19] Zhang Z, Wang C, Peng X, et al. Solar Radiation Intensity Probabilistic Forecasting Based on K-Means Time Series Clustering and Gaussian Process Regression. *IEEE Access* 2021;9:89079–92.
[20] Zhu J, Li M, Luo L. Short-term PV power forecast methodology based on multi-scale fluctuation characteristics extraction. *Renew Energy* 2023;208:141–51.

[21] Ahmad MW, Mourshed M, Rezgui Y. Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression. *Energy* 2018;164:465–74.
[22] Cervone G, Clemente Harding L, Alessandrini S, Delle ML. Short-term photovoltaic power forecasting using artificial neural networks and an analog ensemble. *Renew Energy* 2017;108:274–86.
[23] Jebli I, Belouadha F-Z, Kabbaj MI, Tilioua A. Prediction of solar energy guided by pearson correlation using machine learning. *Energy* 2021;224:120109.
[24] VanDeventer W, Jamei E, Thirunavukkarasu GS, Seyedmahmoudian M. Short-term PV power forecasting using hybrid GASVM technique. *Renew Energy* 2019;140:367–79.
[25] Akhter MN, Mekhilef S, Mokhlis H, Mohamed SaN. Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques. *IET Renew Power Gen* 2019;13:1009–23.
[26] Akhter MN, Mekhilef S, Mokhlis H, et al. A hybrid deep learning method for an hour ahead power output forecasting of three different photovoltaic systems. *Appl Energy* 2022;307:118185.
[27] Wang L, Mao MX, Xie J, Liao Z, et al. Accurate solar PV power prediction interval method based on frequency-domain decomposition and LSTM model. *Energy* 2023;262:125592.
[28] Wang Y, Liao W, Chang Y. Gated Recurrent Unit Network-Based Short-Term Photovoltaic Forecasting. *Energies* 2018;11:2163.
[29] Shi X, Chen Z, Wang H, Yeung D-Y, Wong W-k. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Proceedings of the 28th International Conference on Neural Information Processing Systems,* Montreal, Canada, 2015.
[30] Qu J, Qian Z, Pei Y. Day-ahead hourly photovoltaic power forecasting using attention-based CNN-LSTM neural network embedded with multiple relevant and target variables prediction pattern. *Energy* 2021;232:120996.
[31] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez N. Attention is all you need, *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, California, USA, 2017.
[32] Zhang L, Wilson R, Sumner M, Wu Y. Advanced multimodal fusion method for very short-term solar irradiance forecasting using sky images and meteorological data: A gate and transformer mechanism approach. *Renew Energy* 2023;216:118952.
[33] Liu W, Mao Z. Short-term photovoltaic power forecasting with feature extraction and attention mechanisms. *Renew Energy* 2024;226:120437.
[34] Wang X, Ma W. A hybrid deep learning model with an optimal strategy based on improved VMD and transformer for short-term photovoltaic power forecasting. *Energy* 2024;295:131071.
[35] Hu Z, Gao Y, Ji S, Mae M. Improved multistep ahead photovoltaic power prediction model based on LSTM and self-attention with weather forecast data. *Appl Energy* 2024;359:122709.
[36] Sangrody H, Zhou N, Zhang Z. Similarity-Based Models for Day-Ahead Solar PV Generation Forecasting. *IEEE Access* 2020;8:104469–78.
[37] Zhou Y, Zhou N, Gong L, Jiang M. Prediction of photovoltaic power output based on similar day analysis, genetic algorithm and extreme learning machine. *Energy* 2020;204:117894.
[38] Yue Z, Wang Y, Duan J, Yang T, Huang C, Tong Y, Xu B. TS2Vec: Towards Universal Representation of Time Series. *AAAI Conference on Artificial Intelligence*, 2021.
[39] Rakthanmanon T, Campana B, Mueen A, et al. Addressing Big Data Time Series: Mining Trillions of Time Series Subsequences Under Dynamic Time Warping. *ACM Trans Knowl Discov Data* 2013;7(3):1–31.
[40] Wu H, Xu J, Wang J, Autoformer LM. Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. *Neural Information Processing Systems* 2021.
[41] Desert Knowledge Australia Centre. Download Data: Array Trina, 10.5kW, mono-Si, Dual, 2009. *Alice Springs.* https://dkasolarcentre.com.au/download/notes-on-the-data, date accessed: 16/7/2023.
[42] Yao T, Wang J, Wu H, et al. A Photovoltaic Power Output Dataset. Science Data Bank 2021. https://cstr.cn/31253.11.sciencedb.01094.
[43] Yao T, Wang J, Wu H, et al. A photovoltaic power output dataset: multi-source photovoltaic power output dataset with Python toolkit. *Sol Energy* 2021;230:122–30.