Weather Foundation Model Enhanced Decentralized Photovoltaic Power Forecasting Through Spatio-temporal Knowledge Distillation

Fang He[1] , **Jiaqi Fan**[1] , **Yang Deng**[1] , **Xiaoyang Zhang**[1] , **Ka Tai Lau**[2] and **Dan Wang**[1]
[1]The Hong Kong Polytechnic University
[2]Electrical and Mechanical Services Department, HKSAR
{fanhe, marco.deng, dan.wang}@polyu.edu.hk

Abstract

The *solar photovoltaic power forecasting* (SPPF) of a PV system is vital for downstream power estimation. While approaches for recent decentralized PV systems require customized models for each PV installation, this method is labor-intensive and not scalable. Therefore, developing a general SPPF model for a decentralized PV system is essential. The primary challenge in developing such a model is accounting for regional weather variations. Recent advancements in *weather foundation models* (WFMs) offer a promising opportunity, providing accurate forecasts with reduced computational demands. However, integrating WFMs into SPPF models remains challenging due to the complexity of WFMs. This paper introduces a novel approach, *spatio-temporal knowledge distillation* (STKD), to efficiently adapt WFMs for SPPF. The proposed STKD-PV models leverage regional weather and PV power data to forecast power generation from six hours to a day ahead. Globally evaluated across six datasets, STKD-PV models demonstrate superior performance compared to state-of-the-art (SOTA) time-series models and fine-tuned WFMs, achieving significant improvements in forecasting accuracy. This study marks the first application of knowledge distillation from WFMs to SPPF, offering a scalable and cost-effective solution for decentralized PV systems.

1 Introduction

According to [Masson *et al.*, 2024], over 1.6 TW of PV systems were operational globally by 2024, producing 2,136 TWh of electricity, which accounts for 8.3% of global electricity demand. These PV systems reduced 0.92 gigatons of CO2 emissions, equivalent to 2.5% of global energy-related emissions. Accounting for more than 40% of the PV systems, the decentralized PV systems play an important role in global renewable energy infrastructure.

A *decentralized PV system* refers to a solar power setup where electricity generation is distributed across multiple locations rather than being concentrated in a single, large-scale facility [Zisos *et al.*, 2024]. Decentralized PV systems produce electricity near where it is consumed, reducing the need for long-distance transmission and minimizing energy losses associated with transporting electricity [Kim and Bae, 2017]. SPPF is one of the most fundamental applications for decentralized PV systems since SPPF performs as a prepositive task for a number of PV-oriented applications (e.g., electric grid management, solar intermittency migration, power trading, etc.) [Mansoor *et al.*, 2023]. SPPF involves predicting the power generation of PV systems on various time horizons [Antonanzas *et al.*, 2016]. Traditionally, elaborating individual SPPF models for each PV is required in decentralized PV systems, which is labor-intensive for SPPF developers and thus not scalable [Gaboitaolelwe *et al.*, 2023]. A general SPPF model is imperative for a decentralized PV system.

As highly related to weather, existing studies of SPPF place an emphasis on addressing the challenge of analyzing the impact of weather variation on PV power generation [Ahmed *et al.*, 2019]. Traditional SPPF approaches obtain the weather information from the numerical weather prediction (NWP) models, which are a set of mathematical models of the atmosphere and oceans that predict the weather based on current weather conditions [Travieso-González *et al.*, 2024]. However, using NWP models faces two limitations: 1) the requirements for extremely high computational resources, which can limit the accessibility and affordability of NWP models for SPPF developers, and 2) the errors and uncertainties of NWP models can propagate and amplify in the forecast, affecting the accuracy and reliability of SPPF. Recently, *weather foundation models* (WFMs) show potentials to break these bottlenecks since WFMs can provide highly accurate global weather forecasts up to 14 day ahead with less computation resource [Lam *et al.*, 2023].

Developed along with advanced deep learning techniques and enormous global climate datasets, the WFMs increasingly raise attentions and may be considerable for SPPF tasks [Hamann *et al.*, 2024]. Figure 1 shows a real-world decentralized PV system in California, USA. The weather variation changes significantly across this region, where the distance between the two PVs furthest apart is 930km. The WFMs generally operate at 0.25° resolution (28km x 28km at the equator) and thus are suitable for providing regional weather information for decentralized PV systems.

However, efficiently utilizing the WFMs to develop SPPF

models is still challenging. There are three possible paradigms of adapting WFMs to SPPF models: sequential modeling [Ahmed *et al.*, 2022], fine-tuning [Ding *et al.*, 2023], and knowledge distillation. The sequential modeling and fine-tuning incorporate the complete body of WFMs and output a large model with a redundant structure. SPPF models developed in these two ways have two main limitations: 1) The ceiling of these SPPF models is rather lower since, intuitively, there is non-indigenous weather knowledge irrelevant to SPPF of specific PVs reserved in the output models; and 2) Both the training and inference phases of these SPPF models induce high computational overhead which escalates the cost of PV sectors. Taking GraphCast as a reference, the training of GraphCast takes four weeks on 32 Cloud TPU v4 devices and the inference of GraphCast takes one minute on a single Cloud TPU v4 device [Lam *et al.*, 2023]. Knowledge distillation, a novel deep learning paradigm proposed by [Hinton, 2015], is feasible to address these problems.

Knowledge distillation learns a smaller refined model to mimic the performance of a large comprehensive model. The SPPF models distilled from WFMs can benefit from the following aspects: 1) the model architecture can be flexibly elaborated to fit the characteristics of SPPF; 2) the training is efficient since WFMs can provide informative weather-related knowledge; and 3) the output SPPF models are much more compact and thus affordable for the PV sectors and residential users. The main challenge of distilling WFMs to SPPF is that a model is required not only to inherit weather-related knowledge from WFMs but also to reserve generalizability for SPPF of specific PVs.

To address this challenge, in this paper, we proposed *spatio-temporal knowledge distillation* (STKD) from WFMs to SPPF. It outputs STKD-PV models for SPPF, which takes the regional weather data and PV power data as input and forecasts six hour to day-head PV power generation. We evaluated the STKD-PV models over six real-world datasets distributed worldwide. The results show that STKD-PV models outperforms than a set of five SOTA time-series models and models developed by fine-tuning the WFMs. The SKTD-PV models can achieve an average MAE of 0.0497 over the six datasets. We summarized our contribution as:

1. We formulated the problem of SPPF for decentralized PV systems. We explored to utilize the WFMs to enhance the decentralized SPPF by knowledge distillation. To the best of our knowledge, this is the first study to perform knowledge distillation from WFMs for SPPF.

2. To address the challenges of developing SPPF models from WFMs, we proposed the STKD-PV, which outputs transformer-based models for SPPF by spatio-temporal knowledge distillation from WFMs and fine-tuning with PV power data of the decentralized PV system;

3. We evaluated STKD-PV on six real-world solar power datasets. The results show that STKD-PV can accurately forecast six hour to a day ahead PV power generation and outperform the SOTA time-series models.

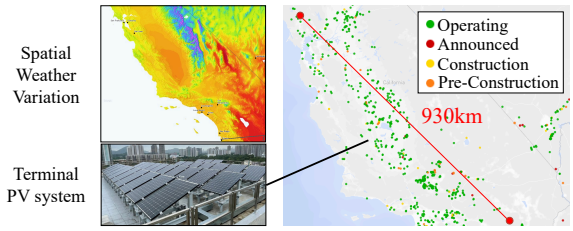

Figure 1: Real-world decentralized PV systems in California, USA

2 Related Work

2.1 SPPF Models

Traditional approaches of SPPF obtain the weather information from the physical models, e.g., the NWP models. SPPF models are developed with regional meteorological information integrated by using the forecasts of NWP models [Verbois *et al.*, 2022; Mayer *et al.*, 2023].

Recently, there has been an increasing number of deep learning models for SPPF that incorporate weather information with power-related PV system data. [Lim *et al.*, 2022] proposes a CNN-LSTM hybrid model for SPPF where the CNN classifies the weather condition and the LSTM learns power generation patterns. [López Santos *et al.*, 2022] develops a temporal fusion transformer for day-ahead SPPF. [Yan *et al.*, 2021] performs short-term SPPF by using a frequency-domain decomposition-based CNN. These studies are limited since models are developed for individual PVs and have insufficient generalizability within decentralized PV systems.

2.2 Weather Foundation Models

As a significant advancement in the use of AI for meteorology, WFMs are trained with vast global weather data and offer the potential for faster and more accurate weather predictions, which can be crucial for planning and responding to weather-related events [Chen *et al.*, 2023]. WFM learn the intricate relationships between weather and geography, capturing meteorological interactions and spatial dependencies between grids [Mukkavilli *et al.*, 2023]. There is a flourish of well-developed WFMs such as: GraphCast by Google DeepMind [Lam *et al.*, 2023], Pangu-Weather by Huawei [Bi *et al.*, 2023], FourCastNet by NVIDIA [Kurth *et al.*, 2023], and ClimaX and Aurora by Microsoft [Bodnar *et al.*, 2024; Nguyen *et al.*, 2023]. WFMs provides opportunities for SPPF models to learn the impact of meteorological variation on specific sites of PV systems.

2.3 Knowledge Distillation

Formally popularized by [Hinton, 2015], knowledge distillation is transferring knowledge from a large complex model (referred to as the "teacher") to a smaller model (known as the "student"). It is successfully applied in numerous domains such as object detection [Li *et al.*, 2022; Zhang *et al.*, 2023], large language models [Guo *et al.*, 2025], computer vision [Beyer *et al.*, 2022], etc. There also are existing efforts to transfer weather-related knowledge from a weather model by using knowledge distillation. [He *et al.*, 2024] distilling physical knowledge from a differential equation net-

work to a spatio-temporal transformer network. [Tang *et al.*, 2023] develops a spatio-temporal graph knowledge distillation to enhance regional weather prediction. Though incorporate modules to distill spatio-temporal knowledge from pre-trained weather models, these studies still show limitations: 1) the teacher models are regional weather models, pre-trained with non-global weather data, and can hardly integrate the global meteorogeographical patterns; and 2) the student models are still weather prediction models, rather than SPPF models. It is still unknown if these distilled models performs well when adapting them to SPPF tasks. In this paper, we try to address these problems by using WFMs as teacher model and develop strategies to reserve the generalizability for SPPF during knowledge distillation.

3 Problem Statement

SPPF for decentralized PV systems can be viewed as a multivariate time-series prediction task within a target region where the decentralized PV system is located. The target region can be represented by a geographical grid, which is a two-dimensional graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. $\mathcal{V}$ is a node-set that denotes the geographical grid points, and $\mathcal{E}$ is an edge set that represents the geographical proximity among different grid points. The weather data within the target region is denoted as $\mathbf{X}^w \in \mathbb{R}^{T\times N\times d^w}$, where N is the number of geographical grid points, T is the length of time-series, and d^w is the dimension of weather variables. The power data collected from the decentralized PV system is denoted as $\mathbf{X}^p \in \mathbb{R}^{T\times M\times d^p}$, where M is the number of terminal PV within the decentralized PV system and d^p is the dimension of power variables of a single terminal PV. The problem of SPPF for decentralized PV system can thus be represented as follows:

Given a target region $\mathcal{G}$, H steps of historical weather conditions $\mathbf{X}^w_{(t-H+1):t}$, and historical power generation of a decentralized PV system $\mathbf{X}^p_{(t-H+1):t}$, for a terminal PV p_m within the decentralized PV system, the SPPF aims to learn a function $F(\cdot)$ to predict Q steps of future power generation:

$$\mathbf{X}^{p_m}_{(t+1):(t+Q)} = F(\mathcal{G}, \mathbf{X}^w_{(t-H+1):t}, \mathbf{X}^p_{(t-H+1):t}), \quad (1)$$

where $\mathbf{X}^w_{(t-H+1):t} \in \mathbb{R}^{H\times N\times d^w}$, $\mathbf{X}^p_{(t-H+1):t} \in \mathbb{R}^{H\times M\times d^p}$, and $\mathbf{X}^{p_m}_{(t+1):(t+Q)} \in \mathbb{R}^{Q\times 1\times 1}$.

4 Methodology

The proposed STKD-PV architecture consists of three main components: 1) **spatio-temporal transformer** as the backbone of decentralized SPPF; 2) **knowledge distillation** from WFMs to the backbone model; and 3) **fine-tuning** the distilled model with power data of decentralized PV systems to obtain SPPF models for terminal PVs.

4.1 Spatio-temporal Transformer

The decentralized SPPF requires a backbone model to: 1) incorporate the knowledge of spatio-temporal variation of regional weather from WFMs effectively; and 2) reserve generalizibility for the SPPF of specific terminal PVs within the decentralized PV system. We accordingly elaborated the STKD-PV model, which consists of two parts: 1) STKD-RW for learning regional weather and 2) an adapter head for adapting to SPPF. Here, we focus on the STKD-RW and leave the adapter head in Section 4.3.

STKD-RW is a stack of multiple transformer blocks with spatial and temporal modules starting with an embedding layer. Specifically, the input historical weather condition $\mathbf{X}^w$ is first encoded to obtain the representation $\mathbf{E} \in \mathbb{R}^{H\times N\times d}$ by an embedding layer, where d is the dimension of the feature representation. To encode the temporal information, we add the positional encoding to the output embeddings as follows:

$$\mathbf{PE}_{(i,pos)} = \begin{cases} \sin(pos/1000^{2k/d})\mathbf{1}, \text{ if } i = 2k, \\ \cos(pos/1000^{2k/d})\mathbf{1}, \text{ if } i = 2k+1, \end{cases} \quad (2)$$

$$\mathbf{Z}^{(0)} = \mathbf{PE} + \mathbf{E} \quad (3)$$

where $\mathbf{PE} \in \mathbb{R}^{H\times N\times d}$, pos is the positional index of the time step, $i = 1, ..., d$ is the index of feature dimension, and $\mathbf{1} \in \mathbb{R}^N$ is a vector of all ones. The obtained $\mathbf{Z}^{(0)}$ is fed into the spatio-temporal modules for further manipulation.

Spatial Attention Module To extract the spatial dependency of regional weather conditions among the grid points within the target region, we introduced a spatial attention module. The regional geographical grid $\mathcal{G}_{region}$ can be split from the global graph $\mathcal{G}_{global}$. A geographical adjacency matrix $\mathbf{G}_{region}$ can be derived from $\mathcal{G}_{region}$, where the connected edges are set as 1 and others as 0.

We use the multi-head attention mechanism to capture the spatial diversity of regional weather, where h heads of attention are calculated and concatenated as features of the next block. Denote the input of l-th block as $\mathbf{Z}^{(l)} \in \mathbb{R}^{H\times N\times d}$, the query, key, and value matrices of an attention head can be computed as:

$$\mathbf{Q}^{(l)} = \mathbf{Z}^{(l)}\mathbf{W}^{(l)}_Q, \mathbf{K}^{(l)} = \mathbf{Z}^{(l)}\mathbf{W}^{(l)}_K, \mathbf{V}^{(l)} = \mathbf{Z}^{(l)}\mathbf{W}^{(l)}_V, \quad (4)$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, $\mathbf{W}_V \in \mathbb{R}^{N\times d/h}$ are the linear projection matrices for query, key, and value. The spatial attention score of i-th head can be calculated as:

$$\text{head}^{(l)}_i = \text{softmax}(\frac{\mathbf{Q}^{(l)}\mathbf{K}^{T(l)}}{\sqrt{d}} \odot \mathbf{G}_{region})\mathbf{V}^{(l)}, \quad (5)$$

where $\odot$ denotes the element-wise multiplication operator. The results of multi-head attention can be concatenated accordingly as follows:

$$\text{MultiHead}^{(l)} = \text{CONCAT}(\text{head}^{(l)}_1, ..., \text{head}^{(l)}_h)\mathbf{W}^{(l)}), \quad (6)$$

$$\mathbf{Z}^{(l)}_{spatio} = \text{LayerNorm}(\mathbf{Z}^{(l)} + \text{MultiHead}^{(l)}), \quad (7)$$

where $\mathbf{W}^{(l)}$ is a projection matrix of the multi-head attention at l-th block. To stabilize the training, we adopt residual connection and layer normalization after each block of multi-head spatial attention.

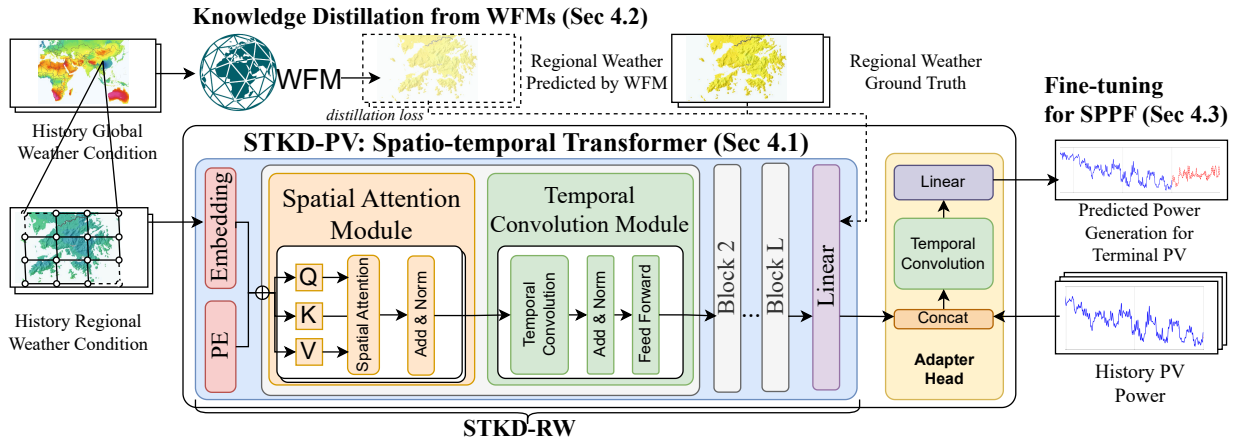

Figure 2: The proposed STKD-PV architecture

Temporal Convolution Module Given the output of the l-th spatial attention, for capturing the temporal dynamics of the regional weather conditions, we use a standard 1D convolution layer with the length of K to force on processing along the temporal dimension, as expressed in the following:

$$\mathbf{Z}^{(l)}_{temporal} = \text{LayerNorm}(\mathbf{Z}^{(l)}_{spatio} + \text{Conv}^{(l)}(\mathbf{Z}^{(l)}_{spatio})) \quad (8)$$

Following the regular operation in the Transformer structure, we also apply layer normalization after residual connections at the end of the temporal convolution layer. The output is finally fed to a feed-forward layer to introduce non-linearity and further transform the learned representation:

$$\mathbf{Z}^{(l+1)} = \text{FFN}^{(l)}(\mathbf{Z}^{(l)}_{temporal}) \quad (9)$$

4.2 Knowledge Distillation from WFMs

In the knowledge distillation, the WFMs perform as teacher networks, and the STKD-RW models perform as student networks. The output of WFMs is firstly *pre-processed* to align the forecasting horizon with SPPF tasks. We use *hybrid distillation strategy* to train the STKD-RW models.

Pre-processing of WFM output The inputs of a WFM is the latest previous observed global weather conditions $\mathbf{X}^w_{(t-1):t}$, and the outputs is the predicted global weather condition of the next time step $\mathbf{X}^w_{t:(t+1)}$. There is a mismatch of prediction horizons between the WFMs and their student networks. To handle this, we introduced a *weather data composition* by taking auto-regressive prediction with WFMs and concatenating the continuous time steps of predicted weather conditions:

$$\hat{\mathbf{X}}^w_{(t+1):(t+Q)} = \text{CONCAT}(\hat{\mathbf{X}}^w_{t+1}, \hat{\mathbf{X}}^w_{t+2}, ..., \hat{\mathbf{X}}^w_{t+Q}), \quad (10)$$

where $\hat{\mathbf{X}}^w_{t+i}$ is the i-th time step of future weather conditions predicted by the WFM and $\hat{\mathbf{X}}^w_{(t+1):(t+Q)}$ is the composited future weather conditions over Q time steps. Note that the the output of WFMs is still a global weather condition, and the regional weather condition for knowledge distillation can be easily split from the global weather condition.

Hybrid Distillation Strategy To retain the generalizability of STKD-RW models during the knowledge distillation for the downstream SPPF tasks, we introduced the *masked self-supervised distillation*, and *data-free and data-driven distillation*. These distillation strategies can be flexibly combined and conducted together.

Masked Self-supervised Distillation We follow the masked self-supervised learning to enhance the generalizability of our model in the spatial dimension [Hou *et al.*, 2022]. During the knowledge distillation training, the geographical adjacency matrix $\mathbf{G}_{region}$ is randomly masked with a pre-defined masking threshold. With such masking, weather information of partial grid points is missing, so the trained STKD-RW models are enforced to infer the weather condition of a certain grid point by gathering information from its adjacent points as far as possible. This enhances the generalizability of STKD-RW models over the target region.

Data-free and Data-driven Distillation To enhance the generalizability in the temporal dimension, we combined data-free and data-driven distillation. Denoting the prediction results of the student networks as $\hat{\mathbf{Y}}_{student} \in \mathbb{R}^{Q\times N\times d}$, the distillation loss is defined by calculating the distance between $\hat{\mathbf{Y}}_{student}$ and the output of regional weather conditions $\mathbf{Y}_{teacher}$ from the WFMs:

$$\mathcal{L}_d = \|\mathbf{Y}_{teacher} - \hat{\mathbf{Y}}_{student}\|_2. \quad (11)$$

For *data-free distillation*, the input data of WFMs is randomly generated from a pre-defined Gaussian distribution. The WFM takes the input and outputs $\mathbf{Y}_{teacher}$ as labels of future weather conditions. The student network also takes the input and outputs predicted weather condition $\hat{\mathbf{Y}}_{student}$ and is trained by back propagation of the distillation loss $\mathcal{L}_d$.

For *data-driven distillation*, the training data is the true data, where the ground truth future weather condition $\mathbf{Y}$ is used as labels. The loss in this process is comprised of two parts: the distillation loss $\mathcal{L}_d$ and prediction loss $\mathcal{L}_p$:

$$\mathcal{L}_s = \alpha\mathcal{L}_d + (1-\alpha)\mathcal{L}_p,\ \mathcal{L}_p = \|\mathbf{Y} - \hat{\mathbf{Y}}_{student}\|_2, \quad (12)$$

where α is the weight to balance the distillation loss and prediction loss. By default, we conduct data-free and data-driven distillation sequentially.

4.3 Fine-tuning for SPPF

For a single terminal PV within a decentralized PV system, the power data of the target decentralized PV system $\mathbf{X}^p \in \mathbb{R}^{T\times M\times d^p}$ is used to fine-tune the distilled STKD-RW models into STKD-PV models. $\mathbf{X}^p$ is combined with the outputs of the STKD-RW models and passed through an additional adapter head. The adapter head is built with a similar structure to the temporal convolution module for further capturing the time-series dynamics of PV power data:

$$\mathbf{Z}^p = \text{CONCAT}(\mathbf{X}^p, \text{Output}_{student}) \tag{13}$$

$$\mathbf{Y}^p = \text{Linear}(\text{Conv}(\mathbf{Z}^p)) \tag{14}$$

During the fine-tuning process, the parameters of STKD-RW are frozen, and only the parameters of the adapter head are updated with the mean squared error loss.

5 Evaluation

5.1 Evaluation Methodology

Datasets Two parts of time-series data are required in our evaluation: weather data and PV power data. For the weather data, we follow the WFMs and adopt ERA5 [Hersbach *et al.*, 2020], which is a climate reanalysis dataset covering the period 1950 to the present. ERA5 is being developed through the Copernicus Climate Change Service (C3S) [1]. ERA5 covers the Earth on a 0.25 degree latitude-longitude grid and provides hourly estimates of a large number of atmospheric, land and oceanic climate variables. For the PV power data, we select five public datasets[2] and collect a private dataset for decentralized PV systems, the specifications of which are shown in Table 1. The regions of these datasets cover the world, which can verify the performance of the proposed STKD-PV on a global scale and thus ensure the scalability of STKD-PV.

Weather Foundation Models We select four impactive WFMs as the knowledge source of weather conditions for SPPF in the experiments: GraphCast by Google DeepMind, Pangu-Weather by Huawei, FourCastNet by NVIDIA, and Aurora by Microsoft. All these four WFMs can provide high-resolution, accurate global weather forecasts. We conduct the knowledge distillation for STKD-PV by using the four WFMs as teacher models individually.

Baseline models To conduct a comparison study for the proposed STKD-PV, considering that the solar power data and weather condition data are both time-series data in essential, we select five SOTA time-series models that are widely practiced in real-world applications as our baseline models:

- **TimesNet** [Wu *et al.*, 2022]: It is a convolution-based network that focuses on discovering the periodicity within the input time-series data. It utilizes a Fourier Transform to transform the input 1D time-series data into 2D tensors so that the convolutional kernels can detect the inter- and intra-period variation of the input.
- **Autoformer** [Wu *et al.*, 2021]: It renovates the Transformer as a deep decomposition architecture, which can progressively decompose the trend and seasonal components during the time-series analysis process.
- **STGCN** [Yu *et al.*, 2017]: It is a spatio-temporal graph convolutional network for time-series prediction tasks by using complete convolutional structures.
- **TCGAN** [Huang and Deng, 2023]: It is a convolutional GAN for time-series data, which learns by playing an adversarial game between two CNNs (i.e., a generator and a discriminator) in the absence of label information.
- **TimeGAN** [Yoon *et al.*, 2019]: It is a GAN for generating realistic time-series data that combines the flexibility of the unsupervised paradigm with the control afforded by supervised training.

All these models are trained with regional weather data and PV power data. Besides, we also consider fine-tuning the WFMs with PV power data as our baselines. Specifically, an adapter head is added after a WFM to integrate the PV power data and further training for SPPF. During the fine-tuning, the parameters of the WFMs are frozen, and the adapter head's parameters are updated. We extend the four aforementioned WFMs with adapter heads as the backbone for fine-tuning, which are denoted as GraphCast-PV, Pangu-PV, FourCastNet-PV, and Aurora-PV, respectively.

Metrics To measure the performance of developed SPPF models, we use the mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of variation of root mean squared error (CV-RMSE) as our metrics (lower is better), which are all widely adopted metrics in the energy domain [Granderson *et al.*, 2016].

Configuration We split each dataset into non-overlapping subsets by 7:1:2 partitions along with the time dimension as our train/validate/test set. Each data set is normalized prior to training. The frequency of time-series is synchronized to 15 min. By default, the input sequence length and output sequence length of models are set to 24 time step (6 hours). For our method, STKD-PV is trained with an Adam optimizer with maximum 200 epochs. The number of spatio-temporal blocks of STKD-PV models is set to three. The balance weight between distillation loss and prediction loss is set to 0.4. The masking rate of masked self-supervised distillation is set to 0.15. The dimension of the embedding layer is set to 256. The results are concluded with 10 trials using different random seeds.

5.2 Results

Overall performance Table 2 shows that STKD-PV models outperform the baseline models on all six datasets. STKD-PV can achieve best performance on all three metrics reported. This indicates that STKD-PV can accurately perform SPPF worldwide. The MAEs of STKD-PV are 0.0584, 0.0336, 0.0406, 0.0578, 0.0453, and 0.0607 on the six dataset,

[1] https://climate.copernicus.eu/

[2] https://www.tudelft.nl/ewi/over-de-faculteit/afdelingen/electrical-sustainable-energy/photovoltaic-materials-and-devices/dutch-pv-portal/pv-power-databases

Dataset	Region	Duration	Frequency	Num. of PV	Num. of Variables
HKPV (Private)	Hong Kong	Apr 2022 - Dec 2023	15 min	17	11
Solar Centre DKA	Australia	Jan 2009 - Dec 2024	5min	40	12
Solar2-Irece	Brazil	Jul 2018 - Jul 2019	15min	96	25
PVDAQ NREL	USA	Dec 2018 - Nov 2023	15min	157	9
Open Power System Data	EU	Jan 2015- Sep 2020	15,30,60min	32	6
PVSPEG	UK	Jul 2013 - Nov 2014	1min	20	30

Table 1: Specifications of PV datasets

respectively. We emphasis the optimal values of MAE, RMSE, and CV-RMSE with bold font in Table 2.

We have observations as follows: 1) Compared to SOTA time-series models, the STKD-PV models perform better since they not only consider the spatio-temporal features of regional weather but also learn high-level global weather knowledge from WFMs. Specifically, STKD-PV models improve 65.04% of MAE, 57.46% of RMSE, and 44.39% of CV-RMSE, respectively, as compared to the SOTA time-series models; 2) Compared to models directly fine-tuned from WFMs which redundantly reserve information irrelevant to weather of the target regions, the STKD-PV models refine the regional weather information and remove this irrelevant information. Specifically, STKD-PV models improves 20.98% of MAE, 20.77% of RMSE, 25.10% of CV-RMSE respectively as compared to the models by fine-tuning WFMs; 3) Within the STKD-PV models, their performance differ from each other. The models distilled from Aurora outperform models distilled from other WFMs. This is because compared to GraphCast, Pangu, and FourCastNet, Aurora is a more general system that can learn from many diverse datasets and adapt to various prediction tasks. Specifically, STKD-PV distilled from Aurora improves 29.86% of MAE, 39.08% of RMSE, 18.41% of CV-RMSE as compared to those distilled from other WFMs.

Internal factors We analyze three internal factors which may influence the performance of STKD-PV models: 1) masked self-supervised distillation, 2) data-free and data-drive distillation, and 3) spatio-temporal modules. Since STKD-PV distilled from Aurora performs best, we use it as the default setting in the following evaluation.

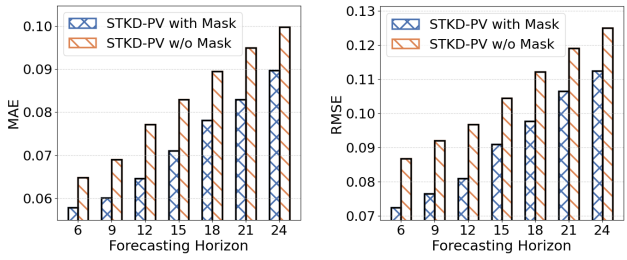

Figure 3: Comparison of distillation with and without mask

Masked Self-supervised Distillation Figure 3 shows the results of STKD-PV with and without masked self-supervised distillation. Compared to STKD-PV without masked self-supervised distillation, STKD-PV performs better on both MAE and RMSE over a six to 24-hour forecasting horizon. Specifically, STKD-PV models have an average improvement of 13.76% on MAE and 23.24% on RMSE. This reveals that masked self-supervised distillation can successfully improve the generalizability of STKD-PV models and thus achieve better performance.

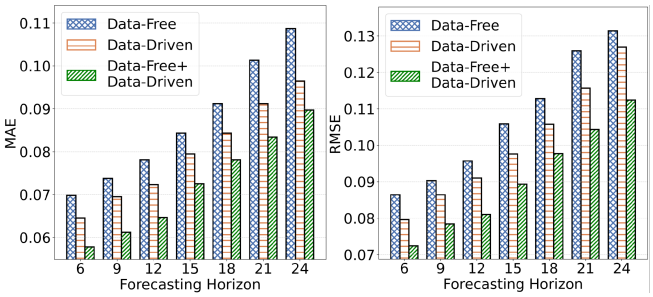

Figure 4: Comparison of data-free and data-driven distillation

Data-free and Data-Driven Distillation To evaluate the strategy of data-free and data-driven distillation, we developed STKD-PV models with three strategies: data-free distillation only, data-driven distillation only, and combining data-free and data-driven distillation together. The size of the training data of data-free distillation stays the same as that of the ground truth training data. Figure 4 shows the comparison results. The combined strategy outperforms the individual data-free and data-driven strategy. Compared to data-free distillation, combined strategy can have an average improvement of 19.84% on MAE and 17.27% on RMSE for STKD-PV models. Compared to data-driven distillation, it also improves 9.44% on MAE and 10.94% on RMSE averagely. This indicates that both data-free and data-driven distillation strategy are effective for enhancing the STKD-PV models.

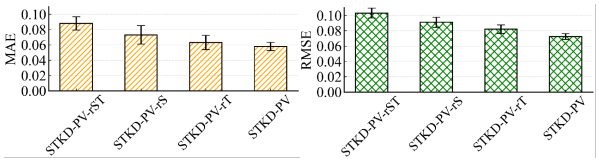

Figure 5: Comparison of STKD-PV and its three varieties

Spatio-temporal Modules To evaluate effectiveness of designed spatio-temporal modules, we developed three variants of STKD-PV: **STKD-PV-rS**: STKD-PV with spatial attention module replaced by a dense layer; **STKD-PV-rT**:

Model	HKPV			Solar Centre DKA			Solar2-Irece		
	MAE	RMSE	CV-RMSE	MAE	RMSE	CV-RMSE	MAE	RMSE	CV-RMSE
TimesNet	0.2771	0.3453	0.6732	0.0875	0.1433	0.4747	0.1950	0.2518	0.4879
AutoFormer	0.3058	0.3707	0.7303	0.1751	0.2366	0.5095	0.3530	0.4059	0.6759
STGCN	0.3089	0.3705	0.7298	0.1727	0.2408	0.5384	0.3530	0.4051	0.6741
TCGAN	0.2114	0.2458	0.6307	0.1987	0.2552	0.5320	0.2077	0.2452	0.4847
TimeGAN	0.2073	0.2450	0.6329	0.2272	0.2982	0.5985	0.2906	0.3591	0.5261
GraphCast-PV	0.0843	0.1465	0.4804	0.0729	0.1273	0.4718	0.0861	0.1409	0.4062
Pangu-PV	0.0814	0.1394	0.4848	0.0843	0.1456	0.4944	0.0936	0.1676	0.4699
FourCastNet-PV	0.1156	0.1879	0.4594	0.1216	0.2056	0.5147	0.1181	0.1961	0.4346
Aurora-PV	0.0712	0.1005	0.4164	0.0637	0.0619	0.4204	0.0528	0.0690	0.3331
STKD-PV(GraphCast)	0.0745	0.1275	0.3567	0.0583	0.1018	0.2961	0.0689	0.1127	0.3122
STKD-PV(Pangu)	0.0634	0.0935	0.3376	0.0674	0.1165	0.3136	0.0749	0.1341	0.3523
STKD-PV(FourCastNet)	0.0894	0.1476	0.3604	0.0973	0.1645	0.3564	0.0945	0.1569	0.3919
STKD-PV(Aurora)	**0.0584**	**0.0721**	**0.3113**	**0.0336**	**0.0476**	**0.2435**	**0.0406**	**0.0531**	**0.2753**
Model	PVDAQ NREL			Open Power System Data			PVSPEG		
	MAE	RMSE	CV-RMSE	MAE	RMSE	CV-RMSE	MAE	RMSE	CV-RMSE
TimesNet	0.0963	0.1590	0.3173	0.1341	0.1646	0.3335	0.1592	0.2337	0.3090
AutoFormer	0.1099	0.1750	0.3492	0.1313	0.1581	0.3018	0.1463	0.2064	0.2766
STGCN	0.1047	0.1683	0.3359	0.1314	0.1381	0.3023	0.1472	0.2044	0.2739
TCGAN	0.1404	0.1850	0.3268	0.1220	0.1287	0.2819	0.1452	0.2005	0.2602
TimeGAN	0.1350	0.1839	0.3245	0.1202	0.1270	0.2701	0.2272	0.2524	0.3293
GraphCast-PV	0.0904	0.1366	0.2945	0.0703	0.0869	0.2013	0.0798	0.1230	0.1794
Pangu-PV	0.0871	0.1283	0.2886	0.0736	0.0921	0.2214	0.0759	0.1140	0.1543
FourCastNet-PV	0.0991	0.1579	0.3074	0.0741	0.1015	0.2263	0.0824	0.1286	0.1803
Aurora-PV	0.0783	0.1084	0.2512	0.0566	0.0689	0.1599	0.0779	0.1170	0.1649
STKD-PV(GraphCast)	0.0723	0.1093	0.2415	0.0562	0.0695	0.1436	0.0638	0.0984	0.1491
STKD-PV(Pangu)	0.0697	0.1026	0.2178	0.0589	0.0737	0.1613	0.0623	0.0936	**0.1436**
STKD-PV(FourCastNet)	0.0793	0.1263	0.2651	0.0593	0.0812	0.1639	0.0659	0.1029	0.1536
STKD-PV(Aurora)	**0.0578**	**0.0867**	**0.1781**	**0.0453**	**0.0551**	**0.1279**	**0.0607**	**0.0912**	0.1473

Table 2: Overall Performance across six PV datasets and ERA5 data

STKD-PV with temporal convolution module replaced by a dense layer; and **STKD-PV-rST**: STKD-PV with both spatio-temporal modules replaced by dense layers. Figure 5 shows that STKD-PV outperforms the three developed varieties. Specifically, STKD-PV lower 20.82%,8.25% and 34.31% of MAE as compared to STKD-PV-rS, STKD-PV-rT, and STKD-PV-rST. STKD-PV also lower 20.43%, and 11.70%,29.70% of RMSE correspondingly. This demonstrates that both the elaborated spatial attention module and temporal convolution module are crucial for learning the spatio-temporal variation of the regional weather and thus can improve the performance of STKD models.

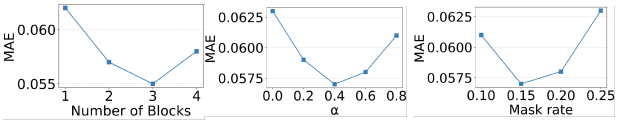

Figure 6: Performance of STKD-PV on different hyper-parameters

Hyper-parameter Analysis we conduct a hyper-parameter analysis on three vital hyper-parameters of STKD-PV: 1) the number of blocks of spatio-temporal modules; 2) the balance weight α between the distillation loss and prediction loss; and 3) the mask rate on the geographical grid during the masked self-supervised distillation.

Figure 6 shows the variation of MAE to the hyper-parameters. For the number of spatio-temporal blocks, we can observe that models with three spatio-temporal blocks perform best. We can also see that a smaller number of blocks may lead to insufficient distillation, and a larger number of blocks may lead to a redundant model structure. For the balance weight α, we can observe that the optimal value of α is 0.4. Larger α leads to overdependent on distilling from the WFM and smaller α results in insufficient distillation. For the mask rate during distillation, we can observe that the optimal value is 0.15. A larger mask rate will make the distillation tough, and a smaller mask rate may lower the generalizability of STKD-PV models.

6 Conclusion

In this paper, to adapt the knowledge of WFMs for SPPF, we proposed the STKD-PV model which consists of spatio-temporal transformers. STKD-PV model is first trained by distilling knowledge from WFMs and then fine-tuned with PV power data. To enhance the generalizability of the STKD-PV model, we introduce a hybrid knowledge distillation strategy that includes masked self-supervised distillation, as well as data-free and data-driven distillation approaches. Experimental results demonstrate that the STKD-PV model achieves significant performance improvements, surpassing baseline models.

Acknowledgements

The authors greatly thank assistance of Qicong Fu and Kerui Wang. Dan Wang's work is supported in part by RGC GRF 15200321, 15201322, 15230624, ITC ITF-ITS/056/22MX, ITS/052/23MX, and PolyU 1-CDKK, G-SAC8. What's more, the authors are indebted to the anonymous reviewers for their constructive comments

References

[Ahmed *et al.*, 2019] Arif Ahmed, Fiona J Stevens McFadden, and Ramesh Rayudu. Weather-dependent power flow algorithm for accurate power system analysis under variable weather conditions. *IEEE Transactions on power systems*, 34(4):2719–2729, 2019.

[Ahmed *et al.*, 2022] Dozdar Mahdi Ahmed, Masoud Muhammed Hassan, and Ramadhan J Mstafa. A review on deep sequential models for forecasting time series data. *Applied Computational Intelligence and Soft Computing*, 2022(1):6596397, 2022.

[Antonanzas *et al.*, 2016] Javier Antonanzas, Natalia Osorio, Rodrigo Escobar, Ruben Urraca, Francisco J Martinez-de Pison, and Fernando Antonanzas-Torres. Review of photovoltaic power forecasting. *Solar energy*, 136:78–111, 2016.

[Beyer *et al.*, 2022] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10925–10934, 2022.

[Bi *et al.*, 2023] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

[Bodnar *et al.*, 2024] Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan Weyn, Haiyu Dong, Anna Vaughan, et al. Aurora: A foundation model of the atmosphere. *arXiv preprint arXiv:2405.13063*, 2024.

[Chen *et al.*, 2023] Shengchao Chen, Guodong Long, Jing Jiang, Dikai Liu, and Chengqi Zhang. Foundation models for weather and climate data understanding: A comprehensive survey. *arXiv preprint arXiv:2312.03014*, 2023.

[Ding *et al.*, 2023] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

[Gaboitaolelwe *et al.*, 2023] Jwaone Gaboitaolelwe, Adamu Murtala Zungeru, Abid Yahya, Caspar K Lebekwe, Dasari Naga Vinod, and Ayodeji Olalekan Salau. Machine learning based solar photovoltaic power forecasting: a review and comparison. *IEEE Access*, 11:40820–40845, 2023.

[Granderson *et al.*, 2016] Jessica Granderson, Samir Touzani, Claudine Custodio, Michael D Sohn, David Jump, and Samuel Fernandes. Accuracy of automated measurement and verification (m&v) techniques for energy savings in commercial buildings. *Applied Energy*, 173:296–308, 2016.

[Guo *et al.*, 2025] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[Hamann *et al.*, 2024] Hendrik F Hamann, Thomas Brunschwiler, Blazhe Gjorgiev, Leonardo SA Martins, Alban Puech, Anna Varbella, Jonas Weiss, Juan Bernabe-Moreno, Alexandre Blondin Massé, Seong Choi, et al. A perspective on foundation models for the electric power grid. *arXiv preprint arXiv:2407.09434*, 2024.

[He *et al.*, 2024] Jing He, Junzhong Ji, and Minglong Lei. Spatio-temporal transformer network with physical knowledge distillation for weather forecasting. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 819–828, 2024.

[Hersbach *et al.*, 2020] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

[Hinton, 2015] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Hou *et al.*, 2022] Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 594–604, 2022.

[Huang and Deng, 2023] Fanling Huang and Yangdong Deng. Tcgan: Convolutional generative adversarial network for time series classification and clustering. *Neural Networks*, 165:868–883, 2023.

[Kim and Bae, 2017] Myungchin Kim and Sungwoo Bae. Decentralized control of a scalable photovoltaic (pv)-battery hybrid power system. *Applied energy*, 188:444–455, 2017.

[Kurth *et al.*, 2023] Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the platform for advanced scientific computing conference*, pages 1–11, 2023.

[Lam *et al.*, 2023] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range

global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

[Li *et al.*, 2022] Gang Li, Xiang Li, Yujie Wang, Shanshan Zhang, Yichao Wu, and Ding Liang. Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 1306–1313, 2022.

[Lim *et al.*, 2022] Su-Chang Lim, Jun-Ho Huh, Seok-Hoon Hong, Chul-Young Park, and Jong-Chan Kim. Solar power forecasting using cnn-lstm hybrid model. *Energies*, 15(21):8233, 2022.

[López Santos *et al.*, 2022] Miguel López Santos, Xela García-Santiago, Fernando Echevarría Camarero, Gonzalo Blázquez Gil, and Pablo Carrasco Ortega. Application of temporal fusion transformer for day-ahead pv power forecasting. *Energies*, 15(14):5232, 2022.

[Mansoor *et al.*, 2023] Majad Mansoor, Adeel Feroz Mirza, Muhammad Usman, and Qiang Ling. Hybrid forecasting models for wind-pv systems in diverse geographical locations: performance and power potential analysis. *Energy Conversion and Management*, 287:117080, 2023.

[Masson *et al.*, 2024] Gaëtan Masson, Arnulf Jäger-Waldau, Izumi Kaizuka, Johan Lindahl, José Donoso, and Melodie de l'Epine. A snapshot of the global pv market. In *2024 IEEE 52nd Photovoltaic Specialist Conference (PVSC)*, pages 0566–0568. IEEE, 2024.

[Mayer *et al.*, 2023] Martin János Mayer, Dazhi Yang, and Balázs Szintai. Comparing global and regional downscaled nwp models for irradiance and photovoltaic power forecasting: Ecmwf versus arome. *Applied Energy*, 352:121958, 2023.

[Mukkavilli *et al.*, 2023] S Karthik Mukkavilli, Daniel Salles Civitarese, Johannes Schmude, Johannes Jakubik, Anne Jones, Nam Nguyen, Christopher Phillips, Sujit Roy, Shraddha Singh, Campbell Watson, et al. Ai foundation models for weather and climate: Applications, design, and implementation. *arXiv preprint arXiv:2309.10808*, 2023.

[Nguyen *et al.*, 2023] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

[Tang *et al.*, 2023] Jiabin Tang, Wei Wei, Lianghao Xia, and Chao Huang. Spatio-temporal graph knowledge distillation. 2023.

[Travieso-González *et al.*, 2024] Carlos M Travieso-González, Fidel Cabrera-Quintero, Alejandro Piñán-Roescher, and Sergio Celada-Bernal. A review and evaluation of the state of art in image-based solar energy forecasting: The methodology and technology used. *Applied Sciences*, 14(13):5605, 2024.

[Verbois *et al.*, 2022] Hadrien Verbois, Yves-Marie Saint-Drenan, Alexandre Thiery, and Philippe Blanc. Statistical learning for nwp post-processing: A benchmark for solar irradiance forecasting. *Solar Energy*, 238:132–149, 2022.

[Wu *et al.*, 2021] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

[Wu *et al.*, 2022] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

[Yan *et al.*, 2021] Jichuan Yan, Lin Hu, Zhao Zhen, Fei Wang, Gang Qiu, Yu Li, Liangzhong Yao, Miadreza Shafie-khah, and João PS Catalão. Frequency-domain decomposition and deep learning based solar pv power ultra-short-term forecasting model. *IEEE Transactions on Industry Applications*, 57(4):3282–3295, 2021.

[Yoon *et al.*, 2019] Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32, 2019.

[Yu *et al.*, 2017] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

[Zhang *et al.*, 2023] Jianyi Zhang, Aashiq Muhamed, Aditya Anantharaman, Guoyin Wang, Changyou Chen, Kai Zhong, Qingjun Cui, Yi Xu, Belinda Zeng, Trishul Chilimbi, et al. Reaugkd: Retrieval-augmented knowledge distillation for pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 1128–1136, 2023.

[Zisos *et al.*, 2024] Athanasios Zisos, Dimitrios Chatzopoulos, and Andreas Efstratiadis. The concept of spatial reliability across renewable energy systems—an application to decentralized solar pv energy. *Energies*, 17(23):5900, 2024.