WINDFORMER: BI-DIRECTIONAL LONG-DISTANCE SPATIO-TEMPORAL NETWORK FOR WIND SPEED PREDICTION

Xuewei Li[1,2,3,4], *Zewen Shang*[1,2,3], *Zhiqiang Liu*[1,2,3], *Jian Yu*[1,2,3], *Wei Xiong*[5,*], *Mei Yu*[1,2,3,4]

[1] College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China.
[2] Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin, 300350, China.
[3] Tianjin Key Laboratory of Advanced Networking, Tianjin, 300350, China.
[4] Tianjin International Engineering Institute, Tianjin University, Tianjin, 300350, China.
[5] TCU School of Civil Engineering, Tianjin Chengjian University, Tianjin 300384, China

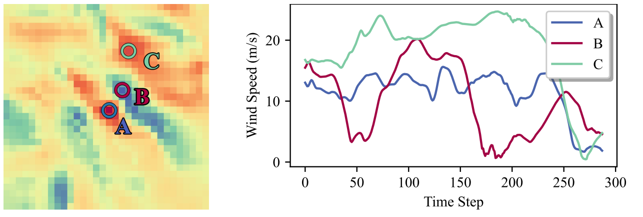
ABSTRACT

Wind speed prediction is critical to the management of wind power generation. Due to the large range of wind speed fluctuations and wake effect, there may also be strong correlations between long-distance wind turbines. This difficult-to-extract feature has become a bottleneck for improving accuracy. History and future time information includes the trend of airflow changes, whether this dynamic information can be utilized will also affect the prediction effect. In response to the above problems, this paper proposes Windformer. First, Windformer divides the wind turbine cluster into multiple non-overlapping windows and calculates correlations inside the windows, then shifts the windows partially to provide connectivity between windows, and finally fuses multi-channel features based on detailed and global information. To dynamically model the change process of wind speed, this paper extracts time series in both history and future directions simultaneously. Compared with other current-advanced methods, the Mean Square Error (MSE) of Windformer is reduced by 0.5% to 15% on two datasets from NERL.

Index Terms— Deep learning, Spatio-temporal prediction, Self-attention, Bi-ConvGRU, Feature fusion

1. INTRODUCTION

In the context of the global shift towards a low-carbon energy structure, wind energy has emerged as one of the most extensively developed and applied renewable energy sources. Wind speed prediction can improve the control foresight and safety of wind turbines, bringing significant safety value and economic benefits. Deep learning has achieved considerable results in the field of wind energy prediction, such as AGRU [1], TACBiLSTMD [2], EMD-LSTM-ARIMA [3] and other RNN methods, FC-CNN [4], U-ConvRes [5] and other CNN-like methods, as well as CNN-GRU [6], STNN [7], etc. that combine the two.



(a) MIC (b) Wind speed curve of three turbines

Fig. 1. Example of long-distance correlations. (a) shows the Maximal Information Coefficients between the wind speeds of wind turbine A located at the center and the wind speeds of each other wind turbine on a randomly selected day.

It can be observed from Figure 1 that despite wind turbine C being at a long-distance from wind turbine A, it still exhibits a notable correlation compared to the nearby wind turbine B. This suggests that the wind speeds of the turbines are not only correlated with those in their immediate vicinity but also with wind speeds at certain long-distance intervals, emphasizing the significance of these relationships.

The self-attention mechanism was first proposed in Transformer [8]. Later, Vit [9] integrated Transformer into CV tasks and achieved good results, especially in the extraction of global long-distance dependencies. Subsequently, Vit variants such as PVT [10], Conformer [11], Swin Transformer [12] appeared continuously. On this basis, this paper proposes a model suitable for dealing with wind speed problems, called Windformer. Windformer divides the wind turbines into multiple non-overlapping windows, performs correlation calculations for the wind turbines in the windows, and then shifts the windows to provide connectivity between windows. This article also designs the Channel Fusion Module to assign different weights to the data of different channels in the model to emphasize more important features. Then, Windformer uses Bi-ConvGRU to improve the ability to extract bidirectional timing information.

*Corresponding author.

This work is supported by National Natural Science Foundation of China (Grant No. 61976155).

In summary, our main contributions are listed as follows.

(1) This paper proposes a Spatial Feature Extraction Module based on Sub-cluster Corr Blocks, which divides the wind turbine cluster into multiple sub-clusters and enables the dynamic extraction of spatial correlations over long-distance. This paper also introduces the Global-detail Fusion Module, which refines features by incorporating detailed and global information.

(2) A Temporal Feature Extraction Module based on forward and reverse bidirectional timing paths is proposed, which captures the temporal correlation of time steps based on past and future dynamic information.

(3) Windformer is experimentally validated on two large-scale wind turbine clusters, each containing over eight hundred wind turbines. The results show that Windformer performs better than other current-advanced methods in different tasks and reduces the MSE by 0.5% to 10%. Our code will be released at
https://github.com/szwszwszw123/Windformer

2. PROPOSED METHOD

2.1. Problem Definition

Windformer utilizes historical feature data from wind turbine cluster to predict future wind speeds. Following the concept introduced by Scene [4], this paper embeds meteorological data from multiple wind turbines at a certain moment into a two-dimensional grid, referred to as a Scene denoted as $S^t \in R^{F\times H\times W}$. Here, t represents the time step, F signifies the number of input features, and H, W represent the height and width of the grid, respectively. The model's input consists of multiple time steps, namely $S = \{S^1, S^2, ..., S^T\} \in R^{T\times F\times H\times W}$, where T denotes the number of time steps. The output yields predicted results $\hat{Y} \in R^L$ for each wind turbine, where L represents the number of wind turbines. The complete prediction process can be defined as follows:

$$\hat{Y} = Windformer(S) \tag{1}$$

The model architecture of Windformer is illustrated in Figure 2(a).

2.2. Time Feature Extraction Module

In wind speed prediction, the features from each time step are indispensable and must be considered collectively. In Recurrent Neural Network, each time step can only access historical features and not future ones, limiting the model's expressive capability. Therefore, Windformer employs the Bi-ConvGRU to handle input scenes, maximizing the utilization of bidirectional information. This method preserves spatial features while extracting temporal features.

GRU is a variant of LSTM. ConvGRU[13] modifies the fully connected layers in GRU to convolutional layers. To effectively utilize bidirectional information, this paper integrates the forward and backward ConvGRUs, forming a Bi-ConvGRU. This integration concatenates the hidden states of both directions at each time step and outputs them together, as shown in the equation below:

$$\begin{aligned} \overrightarrow{H_t} &= \overrightarrow{GRU}(X_t, H_{t-1}), t \in [1, T], \\ \overleftarrow{H_t} &= \overleftarrow{GRU}(X_t, H_{t+1}), t \in [1, T], \\ H_t &= [\overrightarrow{H_t}, \overleftarrow{H_t}] \end{aligned} \tag{2}$$

where $\overrightarrow{H_t}$,$\overleftarrow{H_t}$ represent the hidden states from history and future, respectively. Compared with ConvGRU, Bi-ConvGRU can extract richer time series features from history and future time steps, and the extracted features are convenient for subsequent Spatial Feature Extraction Modules to process.

Following the Bi-ConvGRU, Windformer employs the Turbine Embed Module to independently extract features for each wind turbine. This involves utilizing a linear layer to transform the features of a wind turbine at different time steps into a vector. Subsequently, this vector is input into the Spatial Feature Extraction Module.

2.3. Spatial Feature Extraction Module

2.3.1. *Shift Window Block*

Drawing inspiration from Liu et al. [12], this paper adopts the concept of shifted windows to extract spatial features. The Spatial Feature Extraction Module comprises three stages, each composed of two consecutive Shift Window Blocks, a Turbine Merge Module, and a Channel Fusion Module.

Self-attention has less inductive bias and can handle wind turbines at different distances equally, thus being able to handle both short-distance and long-distance correlations. Based on this background, this paper uses Window Multi-Head self-attention (W-MSA) and Shift Window Multi-Head self-attention (SW-MSA) [12] in Shift Window Block to capture spatial correlation. W-MSA divides the wind turbine into multiple non-overlapping windows and performs self-attention calculations inside the windows.

Utilizing only W-MSA would lead to a lack of continuity in information between different windows. Each wind turbine could only observe other wind turbines within its own window, making it difficult to extract wind speed information from long-distance. SW-MSA addresses this issue effectively by introducing a shift in the windows, establishing connectivity between adjacent windows, as depicted in Figure 2 (b).

To further enhance the model's ability to represent wind speed features and prevent gradient explosion, the Shift Window Block also incorporates a Multi-Layer Perceptron followed by LayerNorm. Subsequently, two residual connections are introduced, as depicted in Figure 2 (b).

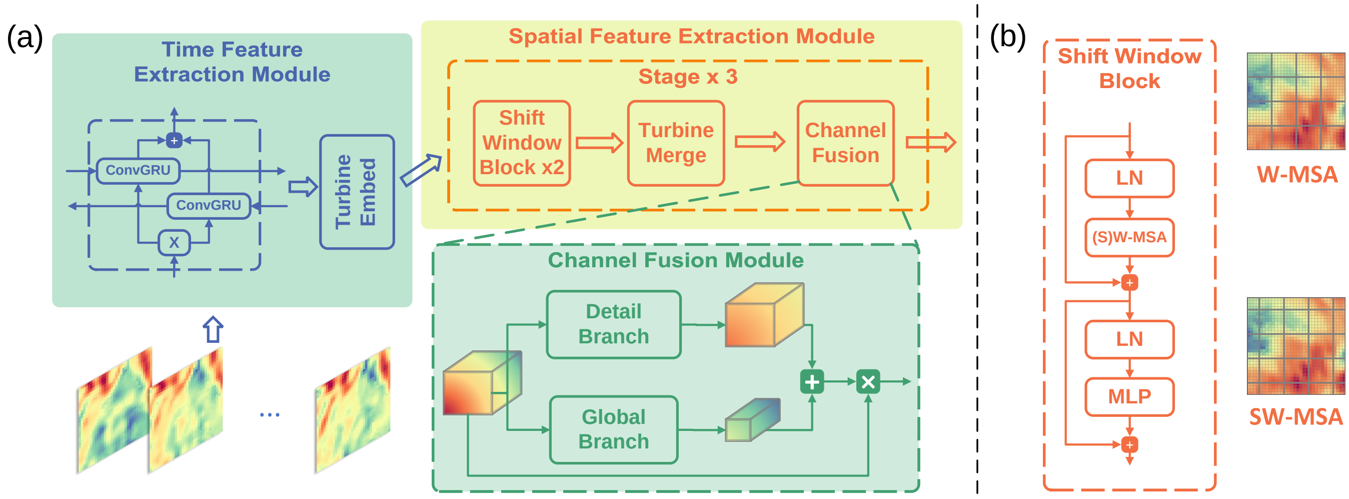


Fig. 2. (a):Architecture of Windformer. (b): Internal Architecture of Shift Window Block. A stage contains two consecutive Shift Window Blocks, with the first using W-MSA, and the second using SW-MSA to provide connectivity for the window.

2.3.2. Turbine Merge Module

After passing through two Shift Window Blocks, the generated features are input into the Turbine Merge Module. The Turbine Merge Module aggregates 2×2 adjacent data points into one, allowing the model to extract features of different granularities.

2.3.3. Channel Fusion Module

The Shift Window Block primarily focuses on modeling internal spatial relationships within the scene, but its modeling of multi-channel information across scenes is insufficient. As the stages deepen, the amount of information contained in an individual data point increases, leading to a growing number of channels. Merely linearly combining these channel features will cause the loss of feature information, so more detailed fusion modeling of different regions of different channels is required.

To address the aforementioned concern, Windformer employs the Channel Fusion Module at the end of each stage to further refine channel features and enhance their expressive capability. Drawing inspiration from the method by Dai et al. [14], this paper incorporates detailed context into the SENet [15] to combine detailed and global wind speed information. The Channel Fusion Module comprises two branches: Global and Detail. In Global Branch, the feature map is firstly pooled globally, and then the global wind speed information is obtained through two linear layers. In Detail Branch, the feature map is convolved twice, and finally, the global and detailed wind speed information is added. Specifically, the detailed information $D(X)$ and the global information $G(X)$ can be computed as follows:

$$\begin{aligned} X' &= Pool(X), \\ D(X) &= BN(Conv(ReLU(BN(Conv(X))))), \\ G(X) &= BN(Linear(ReLU(BN(Linear(X'))))) \end{aligned} \tag{3}$$

where BN refers to the BatchNorm layer. After computing $G(X)$ and $D(X)$, the enhanced features can be expressed as follows:

$$X' = X \otimes M(X) = X \otimes \sigma(L(X) + G(X)) \tag{4}$$

where $\otimes$ denotes element-wise multiplication and σ represents the Sigmoid function.

3. EXPERIMENT AND ANALYSIS

3.1. Problem Definition

3.1.1. Datasets and Model

This paper utilizes two dense wind turbine clusters from NREL [1] as our datasets. The two clusters include 850 and 878 wind turbines respectively, with a grid resolution of 2km × 2km. This paper set prediction tasks for time intervals of 30 minutes, 60 minutes, and 90 minutes to assess the model's generalization capability. The input features include five meteorological features: wind speed, wind direction, air pressure, temperature, and air density.

During model training, this paper set the batch size to 256 and utilizes the AdamW optimizer with a learning rate of 4×10^{-3} and a weight decay of 10^{-4}.

Table 1. Quantitative Evaluation Results on Dataset1. To adapt to the single-step prediction task and Scene size, this paper makes partial adjustments to other deep learning models, removing Reverse scheduled sampling in PredRNN and setting the Inception kernel size in SimVP to 3 and 5.

Model	MSE			MAE		
	30 min	60 min	90 min	30 min	60 min	90 min
XGBR	2.597	4.977	6.945	0.999	1.521	1.880
KNN [16]	2.881	5.638	7.960	1.069	1.645	2.039
SVR [17]	3.114	5.669	7.688	1.029	1.566	1.926
RegressionTree [18]	2.989	5.440	7.400	1.118	1.619	1.961
STNN [7] (2020)	1.287	3.249	4.966	0.702	1.212	1.566
MAU [19] (2021)	1.365	3.131	4.711	0.738	1.712	1.536
PredRNN v2 [20] (2022)	1.282	3.150	4.703	0.700	1.200	1.545
SimVP [21] (2022)	1.233	3.005	4.586	0.698	1.168	1.505
Windformer	**1.197**	**2.914**	**4.536**	**0.681**	**1.149**	**1.491**

Table 2. Quantitative Evaluation Results on Dataset2

Model	MSE			MAE		
	30 min	60 min	90 min	30 min	60 min	90 min
XGBR	0.906	1.604	2.251	0.584	0.845	1.050
KNN [16]	0.988	1.781	2.532	0.619	0.904	1.130
SVR [17]	0.984	1.753	2.448	0.59	0.858	1.066
RegressionTree [18]	0.975	1.710	2.387	0.624	0.886	1.093
STNN [7](2020)	0.472	0.962	1.550	0.416	0.633	0.838
MAU [19] (2021)	0.467	0.917	1.389	0.415	0.608	0.799
PredRNN v2 [20](2022)	0.484	0.957	1.517	0.423	0.637	0.853
SimVP [21](2022)	0.435	0.892	1.324	0.396	0.605	0.764
Windformer	**0.430**	**0.883**	**1.318**	**0.396**	**0.592**	**0.761**

3.1.2. Experimental Results and Comparison

Table 1 & 2 provides a quantitative comparison of errors for various methods. Taking the 60-minute prediction task as an example, for two datasets, Windformer achieves MSE values of 2.914 and 0.883, respectively. Compared to machine learning methods, the MSE reduction for the two datasets ranges from 41.449% to 48.596%, and 44.932% to 50.411%, respectively. Moreover, as the prediction horizon extends, the reduction in errors becomes more pronounced, emphasizing the significance of extracting spatial correlations.

When compared to other current-advanced methods, the MSE of Windformer is reduced by 0.5% to 15%. This demonstrates that the Shift Window Block efficiently extracts correlations among wind turbines at varying distances.

It can be observed from Figure 3 that the prediction curve of Windformer is highly consistent with the real result, and the stability is stronger. The prediction curve of other methods has obvious hysteresis, while the curve of Windformer is highly consistent with the label.

3.1.3. Ablation Experiment

In this study, this paper conducted ablation experiments on various modules using Dataset 1. First, this paper investigates the effectiveness of the Time Feature Extraction Module. It can be observed from Table 5 that Bi-ConvGRU achieved the best performance, and the improvement is obvious compared to ConvGRU and Bi-GRU. This illustrates the need to consider spatial features in the temporal feature extraction stage, and also illustrates the importance of bidirectional temporal feature extraction.

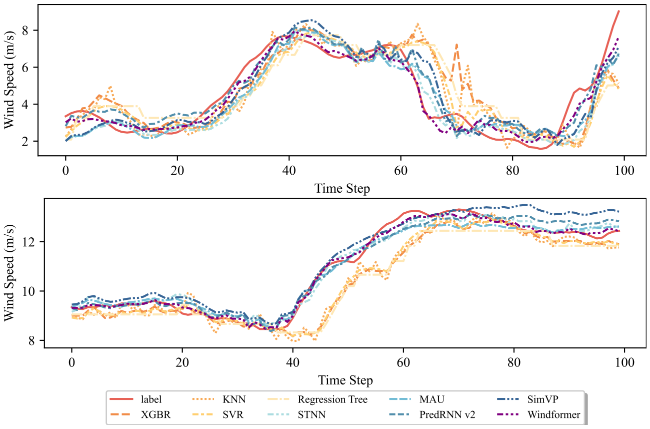


Fig. 3. Prediction curves of different methods on a randomly selected wind turbine on two datasets.

Table 3. Ablation study of Time Feature Extraction Module.

Module	MSE			MAE		
	30 min	60 min	90 min	30 min	60 min	90 min
Empty Module	1.273	3.124	4.754	0.714	1.196	1.726
Bi-ConvRNN	1.214	3.013	4.699	0.686	1.162	1.518
Bi-ConvLSTM	1.217	2.920	4.552	0.69	1.15	1.495
Bi-GRU	1.246	2.982	4.761	0.683	1.158	1.629
ConvGRU	1.230	3.019	4.566	0.692	1.173	1.515
Bi-ConvGRU	**1.197**	**2.914**	**4.536**	**0.681**	**1.149**	**1.491**

Secondly, this paper studies the effectiveness of the Spatial Feature Extraction Module. It can be observed from Table 4 that the error of the Shift Window Block is significantly lower than that of the CNN Module and the Empty Module. Removing the shift window (using two consecutive W-MSA) also reduces the prediction accuracy, which is a testament to the model's ability to capture long-distance correlations.

Table 4. Ablation study of Spatial Feature Extraction Module.

Module	MSE			MAE		
	30 min	60 min	90 min	30 min	60 min	90 min
Empty Module	1.619	3.760	5.478	0.795	1.314	1.663
CNN Module	1.235	2.985	4.626	0.701	1.171	1.513
Window	1.352	3.264	4.729	0.752	1.267	1.572
Shift Window	**1.197**	**2.914**	**4.536**	**0.681**	**1.149**	**1.491**

Subsequently, this paper ablates each Branch in the Channel Fusion Module. From the experiment, it can be observed that the removal of Global Branch or Detail Branch will affect the accuracy of the prediction, which shows that both global feature and detail feature will affect the prediction results, and need to be paid attention to at the same time.

4. CONCLUSION

The model's ability to extract long-distance correlations significantly influences its prediction accuracy, while dynamically incorporating history and future information aids in

[1] https://www.nrel.gov

Table 5. Ablation study of Channel Fusion Module.

Module	MSE			MAE		
	30 min	60 min	90 min	30 min	60 min	90 min
Empty Module	1.292	3.022	4.708	0.753	1.295	1.621
Global Branch	1.275	2.993	4.612	0.691	1.171	1.513
Detail Branch	1.264	3.010	4.579	0.698	1.162	1.523
Channel Fusion Module	**1.197**	**2.914**	**4.536**	**0.681**	**1.149**	**1.491**

comprehending wind speed fluctuation trends. In this context, this paper proposes Windformer, which can effectively extract the long-distance correlation between wind turbines according to the shifted windows and dynamically combine the informative time steps. With this design, Windformer significantly reduces the prediction error and outperforms other current-advanced models.

5. REFERENCES

[1] Zhewen Niu, Zeyuan Yu, Wenhu Tang, Qinghua Wu, and Marek Reformat, "Wind power forecasting using attention-based gated recurrent unit network," *Energy*, vol. 196, pp. 117081, 2020.

[2] Zhengjing Ma and Gang Mei, "A hybrid attention-based deep learning approach for wind power prediction," *Applied Energy*, vol. 323, pp. 119608, 2022.

[3] Ming-De Liu, Lin Ding, and Yu-Long Bai, "Application of hybrid model based on empirical mode decomposition, novel recurrent neural networks and the arima to wind speed prediction," *Energy Conversion and Management*, vol. 233, pp. 113917, 2021.

[4] Ruiguo Yu, Zhiqiang Liu, Xuewei Li, Wenhuan Lu, Degang Ma, Mei Yu, Jianrong Wang, and Bin Li, "Scene learning: Deep convolutional networks for wind power prediction by embedding turbines into grid space," *Applied energy*, vol. 238, pp. 249–257, 2019.

[5] Bruno Quaresma Bastos, Fernando Luiz Cyrino Oliveira, and Ruy Luiz Milidiu, "U-convolutional model for spatio-temporal wind speed forecasting," *International Journal of Forecasting*, vol. 37, no. 2, pp. 949–970, 2021.

[6] Peng Kou, Chen Wang, Deliang Liang, Song Cheng, and Lin Gao, "Deep learning approach for wind speed forecasts at turbine locations in a wind farm," *IET renewable power generation*, vol. 14, no. 13, pp. 2416–2428, 2020.

[7] Yongqi Liu, Hui Qin, Zhendong Zhang, Shaoqian Pei, Zhiqiang Jiang, Zhongkai Feng, and Jianzhong Zhou, "Probabilistic spatiotemporal wind speed forecasting based on a variational bayesian deep learning model," *Applied Energy*, vol. 260, pp. 114259, 2020.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[10] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao, "Pvtv2: Improved baselines with pyramid vision

transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 1–10, 2022.

[11] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye, "Conformer: Local features coupling global representations for visual recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 367–376.

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.

[13] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville, "Delving deeper into convolutional networks for learning video representations," *arXiv preprint arXiv:1511.06432*, 2015.

[14] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard, "Attentional feature fusion," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3560–3569.

[15] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[16] Raik Becker and Daniela Thrän, "Completion of wind turbine data sets for wind integration studies applying random forests and k-nearest neighbors," *Applied energy*, vol. 208, pp. 252–262, 2017.

[17] Kuilin Chen and Jie Yu, "Short-term wind speed prediction using an unscented kalman filter based state-space support vector regression approach," *Applied energy*, vol. 113, pp. 690–705, 2014.

[18] Alberto Torres-Barrán, Álvaro Alonso, and José R Dorronsoro, "Regression tree ensembles for wind energy and solar radiation prediction," *Neurocomputing*, vol. 326, pp. 151–160, 2019.

[19] Zheng Chang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, Yan Ye, Xiang Xinguang, and Wen Gao, "Mau: A motion-aware unit for video prediction and beyond," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26950–26962, 2021.

[20] Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, S Yu Philip, and Mingsheng Long, "Predrnn: A recurrent neural network for spatiotemporal predictive learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2208–2225, 2022.

[21] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li, "Simvp: Simpler yet better video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3170–3180.