EasyST: Modeling Spatial-Temporal Correlations and Uncertainty for Dynamic Wind Power Forecasting via PaddlePaddle

Yiji Zhao[1,2], Haomin Wen[1,2,3], Junhong Lou[3], Jinji Fu[1,2], Jianbin Zheng[3], Youfang Lin[1,2,*]
[1]School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
[2]Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing, China. *Team advisor
[3]Artificial Intelligence Department, Cainiao Network, Hangzhou, China.
{yijizhao,wenhaomin,20271271,yflin}@bjtu.edu.cn;
{loujunhong.loujunh,jabin.zjb}@cainiao.com;

ABSTRACT

Wind Power Forecasting (WPF), which aims to predict the future wind power supply, is increasingly becoming one of the most critical issues in wind power integration and operation. Accurate wind power forecasting is key to reducing wind power fluctuations in system dispatch planning. However, accurate prediction of wind power is challenged by the following two aspects: i) complex spatial-temporal correlations and ii) huge data uncertainties. To address the above challenges, we propose H-STWPF, which can capture the spatial-temporal correlations and uncertainty for wind power forecasting. Specifically, we first develop a deterministic model (named Temporal MLP) to forecast wind power effectively and efficiently. Meanwhile, a probabilistic model (named Deep Factor) is proposed to model uncertainty factors and achieve probabilistic prediction to improve the robustness of the model. At last, to obtain more accurate predictions, a multi-stage and multi-granularity fusion mechanism is developed that can fuse models with different prediction lengths. H-STWPF is fully developed based on paddlepaddle and equipped with a unified data pipeline and training configuration to quickly implement multi-model training and integration. Experiments in the Baidu KDD Cup 2022 show the effectiveness of our model.

ACM Reference Format:
Yiji Zhao[1,2], Haomin Wen[1,2,3], Junhong Lou[3], Jinji Fu[1,2], Jianbin Zheng[3], Youfang Lin[1,2,*]. 2022. EasyST: Modeling Spatial-Temporal Correlations and Uncertainty for Dynamic Wind Power Forecasting via PaddlePaddle. In *Proceedings of Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Wind Power Forecasting (WPF), which aims to predict the future wind power supply, is increasingly becoming one of the most critical issues in wind power integration and operation [17]. Accurate wind power forecasting is the key to reducing wind power fluctuations in system dispatch planning [4, 13]. However, accurate prediction of wind power is challenged by the following two aspects:

Complex Spatial-Temporal Correlations. Wind power is affected by various kinds of spatial and temporal factors. Previous methods mainly model the WPF as a time series forecasting problem. Each turbine is considered independent of the others in those methods. However, in real scenarios, spatial correlations exist between different turbines. For example, generated wind power of a turbine usually has the same trend or shares some similarities with its spatial neighbors. How to effectively encode and model both the spatial and temporal correlations brings huge challenges to the problem.

Huge data uncertainties. Affected by stochastic factors such as the weather, wind speed and turbine internal contexts, wind power data contains huge uncertainties. A majority of previous works focus on the point prediction of wind power. Although some of them achieved promising performance, those methods cannot depict the uncertainty of the prediction. How to model the uncertainties in the data and build a probabilistic prediction is a great challenge.

To address the above challenges, we propose a Hybird Spatial-Temporal Wind Power Forecasting framework (H-STWPF). It is a hybrid deep learning model equipped with two important prediction models, i. e., a deterministic prediction model and an uncertainty prediction model. Both spatial and temporal correlations are captured in each model. In addition, in order to predict wind power more accurately within 48 hours (i.e., 288 steps), a multi-term and multi-grain fusion mechanism is designed to improve the prediction accuracy by integrating different models. In summary, the main contributions of this work are as follows:

- We propose H-STWPF, which can capture the spatial-temporal correlations and uncertainty for wind power forecasting.
- A deterministic model (named Temporal MLP) is proposed to forecast wind power effectively and efficiently. Meanwhile, a probabilistic model (named Deep Factor) is proposed to model uncertainty factors and achieve probabilistic prediction to improve the robustness of the model.
- To obtain more accurate predictions, a multi-term and multi-grain fusion mechanism is developed that can fuse models with different prediction lengths.
- H-STWPF is fully developed based on paddlepaddle and equipped with a unified data pipeline and training configuration to quickly implement multi-model training and integration. Experiments in the Baidu KDD Cup 2022 show the effectiveness of our model.

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 PROBLEM DEFINITION

The main objective of this study is to forecast wind power with respect to the dynamic features of each wind turbine, including weather conditions, internal status and other signals of wind turbines. Meanwhile, the relative positions of different wind turbines on the farm and contextual information are used to enhance the spatial correlation of each turbine. Given historical data $\mathcal{X} \in \mathbb{R}^{N \times T \times F}$ for a wind farm of length T, where N is the number of turbines and F is the feature number of each turbine, the prediction target $\mathcal{Y} \in \mathbb{R}^{N \times P \times 1}$ is the power time series generated by all turbines in the future of length P. For notational purposes, we use B to denote the batch size and C to denote the number of hidden units.

3 MODEL

3.1 Overview

Figure 1 shows the overall architecture of our method. We construct H-STWPF in the following three steps: 1) data pre-processing & feature extraction; 2) multi-model prediction; 3) multi-term & multi-grain fusion. The data pre-processing & feature extraction step first remove the noise and do the data augmentation of the original data, then extract comprehensive features based on the processed data. The multi-model prediction learns multiple deterministic and probabilistic models under different granularities. The deterministic model (named Temporal MLP) is proposed to forecast wind power effectively and efficiently. And the probabilistic model (named Deep Factor) can model uncertainty factors and achieve probabilistic prediction. At last, the multi-term & multi-grain fusion step fuses the predictions of multiple models with different characteristics for more accurate prediction.

3.2 Data Pre-processing & Feature Extraction

To avoid the influence of noise features, only *active power (i.e., Patv)* and *wind speed (i.e., wspd)* are used in our method.

3.2.1 Data Pre-processing. Normalized data helps to improve the training speed and prediction accuracy of the model, so we first preprocess the data in the following operations:

Pre-filling missing values. All missing values are initially filled by 0.

Cilpping abnormal values. In this work, abnormal values refer to values not aligned with physical rules. For example, the generated wind power is less than zero. In such cases, we clip those values into their corresponding value range.

Spatial group partitioning with turbine clustering. To capture the spatial correlation, turbines with similar power generation patterns are partitioned into the same group by clustering. Specifically, the Euclidean distance of the time series of *Patv* is calculated, and the turbines are divided into 6 clusters using the K-means algorithm, as shown in Figure 2.

Group-mean aggregation. We aggregate the data (i.e., *Patv*, *wind speed*) from multiple turbines and take their average to generate a new time series. We obtain six new mean aggregation series based on the grouping information of the turbines, called the spatial group mean series, which are fed into the feature extraction.

Global-mean aggregation. We also define a spatial global-mean series, which obtains a new series by merging the time series of all turbines. It can help our model to capture the common patterns of all turbines.

Spatial filling of missing values. Since the zero-filling approach cannot effectively exploit the spatial pattern information, we design a novel spatial filling method that uses group-mean to fill in the missing values for each turbine.

Coarse-grained aggregation of time series. In addition to the original 10min-granularity, we introduce more coarser granularities by aggregating the data in the same 20min / 30min window.

Table 1: Feature table with 10-minute granularity.

feature type	features	configure
raw values	patv,wspd	granularity = 10 min
time different features	diff	Δt = 1,2,3
time window statistics	max, min, std, var, mean, median	window = 1h, 3h, 6h, 12h, 24h, 48h, 72h

3.2.2 Feature Extraction. In this work, the extracted features can be roughly classified into three types: i) raw values ii) time difference and iii) time window statistics. All the features are listed in Table 1. We extract the features for the input three series, and get the full series features, group series features and global series features.

Raw values. The original inputs after filling missing values and clipping abnormal values. The raw values contain the basic information of the historical data.

Time difference features. It is constructed by the difference between the values of the given two time steps. Let Δt denote the duration between the two time steps. We extract multiple time difference features to fully capture the change patterns in the data.

Time window statistics. We count statistical features such as mean, variance, standard deviation, maximum value, and minimum value of a feature within a given time window.

Spatial position embedding. A trainable embedding is assigned to each turbine as follows:

$$\mathbf{SE} = \left[z_1^s, z_2^s, \ldots, z_N^s \right] \in \mathbb{R}^{N \times d}. \tag{1}$$

Temporal position embedding. We use *time-of-day* and *hour-of-day* to depict the position of each time step t and construct two trainable embeddings for them. For each time step, the two embeddings are combined by summing. All temporal position embedding vectors are denoted as:

$$\mathbf{TE} = \left[z_1^t, z_2^t, \ldots, z_{T_d}^t \right] \in \mathbb{R}^{T_d \times d}. \tag{2}$$

3.2.3 Feature normalization. We use Z-Score normalization to normalize the features in each dimension.

3.3 Spatial and Temporal Modeling

3.3.1 Spatial Correlation Modeling. There are spatial correlations between turbines, e.g., the power time series of turbines in geographical proximity can be similar. We constructed group-mean series and global-mean series to model local spatial correlation and global spatial correlation, respectively. In order that each turbine can effectively exploit spatial correlations, local and global series features are fed into the neural network along with the original series features. The input is denoted as $\mathbf{X}_i \in \mathbb{R}^{T \times F}$.

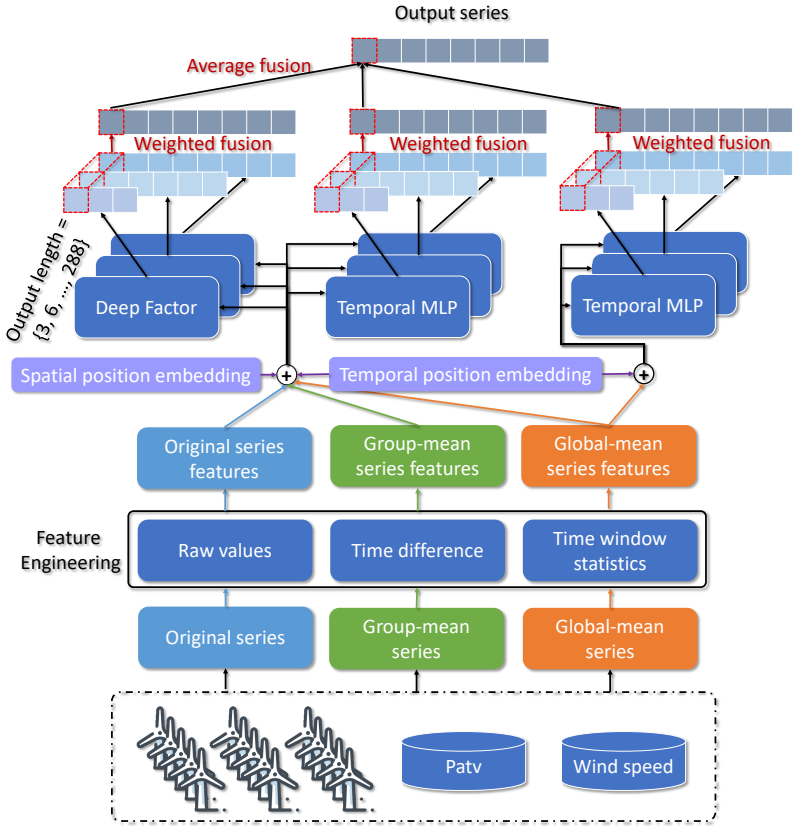

Figure 1: Overview of the Solution, which includes the following three steps: 1) data pre-processing & feature extraction; 2) multi-model prediction; 3) multi-term & multi-grain fusion. The data pre-processing & feature extraction step first remove the noise and do the data augmentation of the original data, then extract comprehensive features based on the processed data. The multi-model prediction contains the deterministic model (Temporal MLP) and the probabilistic model (Deep Factor) under different granularity to forecast wind power effectively and efficiently. The multi-term & multi-grain fusion step fuses the predictions of multiple models with different characteristics for more accurate predictions.

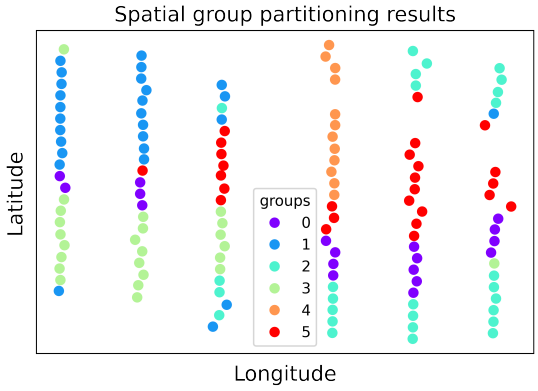

Figure 2: Spatial group partition.

3.3.2 Temporal Correlation Modeling. Since the state of the future time step depends on the historical time step, we construct two network components to capture both short-term and long-term time series correlations.

Gated Recurrent Unit (GRU). GRU [2] has been shown to be excellent at capturing short-term temporal dependencies. Therefore, GRU is used as one of the main components in our model.

Temporal Linear Layer (TLinear). It directly flattens and blends information from multiple time steps to capture long-term temporal dependencies. It is efficient and cost-effective compared to the currently popular self-attention structures [14], which helps us ensemble more models and benefit from them. Given the hidden layer representation $\mathcal{H} \in \mathbb{R}^{B\times N\times T\times C}$ of all turbines, the time-linear layer is formulated as:

$$\tilde{\mathcal{H}} = \text{TLinear}_{TC\to C}(\mathcal{H}) \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{TC\times C}$ and $\boldsymbol{b} \in \mathbb{R}^{C}$ are trainable parameters.

3.3.3 Input with Spatial and Temporal Position Embedding. Since MLP and GRU do not have the ability to learn spatial and temporal inductive biases, inspired by Transformer [14], the learnable spatial and temporal position embeddings are added to the input $\mathcal{X}$ =

$(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N) \in \mathbb{R}^{B\times N\times T\times F}$ as follows:

$$\mathcal{H}^0 = \text{Linear}_{F\to C}(\mathcal{X}) \oplus \text{Linear}_{d\to C}(\mathbf{SE}) \oplus \text{Linear}_{d\to C}(\mathbf{TE}), \quad (4)$$

where $\oplus$ denotes the broadcast addition operation.

3.4 Deterministic Forecasting Model

We design a deterministic prediction model, called **Temporal MLP**, which directly regresses the value of the prediction target. We design the model directly using MLP instead of self-attention because the proposed TLinear layer is able to capture long-term temporal dependencies.

As shown in Figure 1, we first construct a multi-layer MLP network to enhance the representation of the model. Each network layer consists of a linear layer and a GLU [3] activation unit as follows:

$$\mathcal{H}^l = \text{GLU}(\text{Linear}_{C\to 2C}(\mathcal{H}^{l-1})). \quad (5)$$

Then, the TLinear layer is used to extract temporal features based on the features extracted from each network layer, and the obtained features are then fed into a linear layer and GLU activation units to boost the nonlinearity:

$$\begin{aligned}\tilde{\mathcal{H}}^l &= \text{TLinear}_{TC\to C}(\mathcal{H}^l) \in \mathbb{R}^{B\times N\times C},\\ \tilde{\mathcal{H}}^l &= \text{GLU}(\text{Linear}_{C\to 2C}(\tilde{\mathcal{H}}^l)) \in \mathbb{R}^{B\times N\times C}.\end{aligned} \quad (6)$$

Finally, all temporal features are collected and fed into an MLP consisting of two layers to directly predict multiple future time steps as follows:

$$\begin{aligned}\mathcal{O} &= \text{GLU}(\text{Linear}_{(L+1)C\to C'}(\tilde{\mathcal{H}}^0 \parallel \cdots \parallel \tilde{\mathcal{H}}^L)) \in \mathbb{R}^{B\times N\times C'},\\ \hat{\mathcal{Y}} &= \mathit{Reshape}(\text{Linear}_{C'\to P}(\mathcal{O})) \in \mathbb{R}^{B\times N\times P\times 1}.\end{aligned} \quad (7)$$

Based on the above network, we obtain the final prediction result $\hat{\mathcal{Y}}$. In our experiments, we set C=64, L=2, and use a *Dropout* operation with a missing rate of 0.3 for $\mathcal{O}$ during the training process.

In general, using evaluation metrics as loss functions in the training process can lead to more accurate prediction results. Since RMSE may fail to compute the gradient, we construct a pseudo-score loss consisting of MSE and MAE as follows:

$$\mathcal{L}_{\text{score}} = \frac{1}{2SP}\left(\sum_{s=1}^{S}\sum_{i=1}^{N}\sum_{t=1}^{P} m_{s,i,t}\cdot\left[\left(y_{s,i,t}-\hat{y}_{s,i,t}\right)^2 + \left|y_{s,i,t}-\hat{y}_{s,i,t}\right|\right]\right) \quad (8)$$

where S is the number of samples. y and $\hat{y}$ are the actual and predicted wind power, respectively. Note that y and $\hat{y}$ are scaled by dividing by 1000 before calculating the loss in order to be consistent with the online evaluation standard. In addition, $m \in \{0, 1\}$ denotes the mask, and if the predicted target is abnormal, $m = 0$.

3.5 Probabilistic Forecasting Model

Affected by stochastic factors such as the weather, wind speed and turbine internal contexts, wind power data contains huge uncertainties. To describe the prediction uncertainty, we introduce a probabilistic prediction model **Deep Factor** [15]. Unlike the Temporal MLP output point prediction results, Deep Factor describes each prediction target as a probability $p(y|x)$. Given a forecast length P, our goal is to calculate the joint predictive distribution of future observations, $p(\{y_{i,T+1:T+P}\}_{i=1}^N|\{\boldsymbol{x}_{i,1:T}\}_{i=1}^N)$.

The key assumption of Deep Factor is that each time series is governed by a fixed global (non-random) and a random component. In particular, we assume the following generative process:

$$\text{fixed effect}: \quad \boldsymbol{\sigma}_t = \text{Linear}_{C\to 1}(\text{GRU}(\mathbf{H}_{t-1})) \in \mathbb{R}^{BN}, \quad (9)$$

$$\text{random effect}: \quad \boldsymbol{\mu}_t = \text{Linear}_{C\to 1}(\text{GRU}(\mathbf{H}_{t-1})) \in \mathbb{R}^{BN}, \quad (10)$$

$$\text{emission}: \quad \boldsymbol{y}_t \sim p(\boldsymbol{y}_t|\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t),$$

The observation model p can be any parametric distribution, such as Gaussian, Poisson, or Negative Binomial. In this work, we use the Gaussian distribution. Then, we use the negative log-likelihood loss to train the model as follows:

$$\begin{aligned}\mathcal{L}_{\text{NLL}} &= -\sum_{s=1}^{S}\sum_{i=1}^{N}\sum_{j=1}^{P} m_{s,i,t}\cdot\log p(y_{s,i,t}|\mu_{s,i,t},\sigma_{s,i,t})\\ &= \sum_{s=1}^{S}\sum_{i=1}^{N}\sum_{j=1}^{P} m_{i,j}^{s}\cdot\left(\frac{\log\sigma_{s,i,t}^2}{2}+\frac{\left(y_{s,i,t}-\mu_{s,i,t}\right)^2}{2\sigma_{s,i,t}^2}+C\right),\end{aligned} \quad (11)$$

where C is a constant which can be omitted during training. $m_{s,i,t}$ is the mask, refer to Eq. (8) for details. In the inference stage, we use a fixed effect $\mu_{s,i,t}$ as the final prediction.

3.6 Hybrid Model

Owing to the huge data uncertainties, the characteristics of wind power fluctuations are so complex that a single model can hardly capture them comprehensively. Therefore, we train several above-proposed models and develop a hybrid model with a multi-term & multi-grain fusion mechanism to fully take advantage of different models. As shown in Table 2, the candidate fusion models can be broadly classified by the following two dimensions:

Table 2: Model Table.

model	term	pred length	granularity
Temporal MLP (global-mean)	short	3,6,9,12,24,34	10min
	middle	72	20min
	long	144,288	10min
	long	144,288	30min
Temporal MLP (all-turbine)	short	24,34	10min
Deep Factor	short	3,6,9,12,24,34	10min
	middle	108	30min
	long	144,288	10min

Multi-term. Firstly, the models can be divided into short-term, middle-term, and long-term models according to their prediction length. Different models are designed to learn the patterns in their corresponding terms.

Multi-grain. Secondly, the models can be divided into 10min-grain (the original grain), 20min-grain, and 30min-grain according to the input and output grain of the model. Coarser granularity enables the model to process more long-term data with the same input length, thereby increasing the receptive field of the model. In addition, the granularity should not be too large, because much original information will be lost under large granularity, making it difficult

for the model to learn effective trend information. Therefore, we use 20min and 30min besides the original granularity (i.e., 10min).

Moreover, it is worth mentioning that Temporal MLP has two versions: global-mean and all-turbine. Temporal MLP (global-mean) only predict the average value of all turbines, then broadcasts the average value to all turbines as their predictions. On the contrary, Temporal MLP (all-turbine) predicts the values of all the turbines at once.

Fusion. After getting the models that capture different patterns, a key problem is how to aggregate them to form the final prediction. An intuitive method is to average the predictions of different methods as the final result. However, models with different prediction lengths can perform differently. It is not wise to assign the same weights to them. For example, suppose there is a model with a prediction length of 12, and another model with a prediction length of 288. Intuitively, the 12-model should take more weight than the 288-model in the prediction range 0-12. Based on the above analysis, in this work, We assign different weights to models with different prediction lengths.

Table 3: Example of Fusion Weight.

	1-12	13-24	25-288
$\mathcal{M}_{12}$	50%	0	0
$\mathcal{M}_{24}$	40%	60%	0
$\mathcal{M}_{288}$	10%	40%	100%

Here we use an example to describe the idea of fusion. Suppose that there are three models with prediction length 12, 24, 288 (denoted by $\mathcal{M}_{12}, \mathcal{M}_{24}, \mathcal{M}_{288}$, respectively). Then an example weight matrix will be constructed as shown in Table 3. In the prediction length range 0-12, $\mathcal{M}_{12}, \mathcal{M}_{24}, \mathcal{M}_{288}$ contributes 50%, 40%, 10%, respectively. Therefore, the first time step in the prediction $y_{t=1}$ is calculated by Equation 12. While in the prediction length range 12-24, $\mathcal{M}_{24}, \mathcal{M}_{288}$ contributes 60%, 40%, respectively. So $y_{t=13}$ is calculated by Equation 13. Note that $\mathcal{M}_{12}$ is assigned weight 0 in the prediction length 12-24, because $\mathcal{M}_{12}$ has no prediction in the length range 12-24. The weight matrix in our models is illustrated by a heat map shown in Figure 3.

$$y_{(t=1)} = 0.5 * \mathcal{M}_{12}(t=1) + 0.4 * \mathcal{M}_{24}(t=1) + 0.1 * \mathcal{M}_{288}(t=1). \quad (12)$$

$$y_{(t=13)} = 0.6 * \mathcal{M}_{24}(t=13) + 0.4 * \mathcal{M}_{288}(t=13). \quad (13)$$

At last, because the wind power will be largely affected by seasons, for example, the power generation in summer will be higher than that in winter on average. However, the season is unknown in the online inference process. Therefore, we design a factor to artificially scale the predicted results, to make the predictions closer to the distribution produced by the season of the online test data.

4 EXPERIMENT

We extensively evaluate the proposed model in the Baidu KDD Cup 2022.

Dataset. The dataset we use is collected by Longyuan Power Group[1], which contains the wind power data in 245 days of 134 turbines

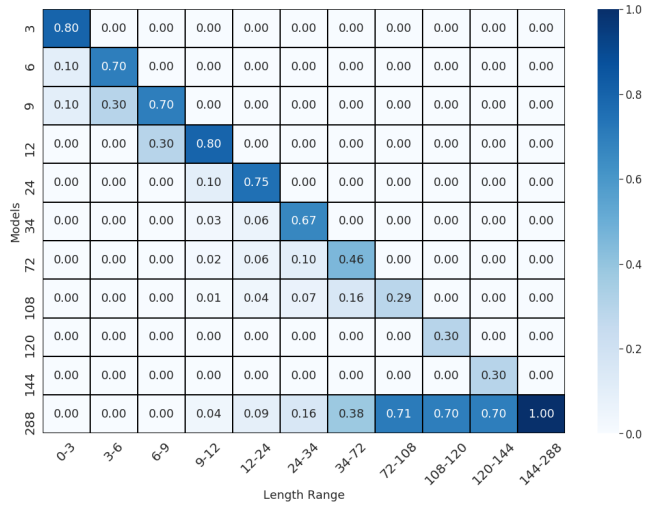

Figure 3: Fusion Weight in Our Solution.

from the same wind farm. And each record is sampled every 10 minutes. In addition to the generated wind power, the data also provides other information such as the spatial distribution of wind turbines, as well as the dynamic context factors like temporal, weather, and turbine internal status.

Experimental Setting. In the experiment, we split the given dataset into training, validation in chronological order by 215 days and 30 days, and test the model by the online test data. Our model is trained by the Adam [8] optimizer with an initial learning rate of 2e-4. The batch size is set to 32. And the training process is early stopped within 10 epochs. The model is implemented by PaddlePaddle [9] and trained on NVIDIA V100 GPUs provided by AI Studio[2].

Metrics. Following the setting of the KDD Cup, we evaluate the prediction results for each wind turbine and then sum the prediction scores as the final score of the model. The evaluation score $s^i_{t_0}$ for wind turbine i at the time step t_0 is defined as:

$$s^i_{t_0} = \frac{1}{2}\left(\sqrt{\frac{\sum_{j=1}^{288}\left(y^i_{t_0+j} - \overline{y}^i_{t_0+j}\right)^2}{288}} + \frac{\sum_{j=1}^{288}\left|y^i_{t_0+j} - \overline{y}^i_{t_0+j}\right|}{288}\right), \quad (14)$$

where y and $\overline{y}$ is the actual and predicted wind power, respectively. The overall score of a model S_{t_0} at time t_0 is the sum of the prediction score on all wind turbines.

Results. We report the evolution process of our model in the second phase, which includes the following version:

- v1: Only includes the Temporal MLP (global-mean) model.
- v2: add the multi-term prediction based on v1.
- v3: add the multi-grain prediction based on v2.
- v4: predict the value for each turbine (i.e., Temporal MLP (all-turbine)), not the average value of all turbines (i.e., Temporal MLP (global-mean)), compared with v3.
- v5: Add the probabilistic forecasting model Deep Factor based on v4.

[1] https://aistudio.baidu.com/aistudio/competition/detail/152/0/datasets

[2] https://aistudio.baidu.com/aistudio/index

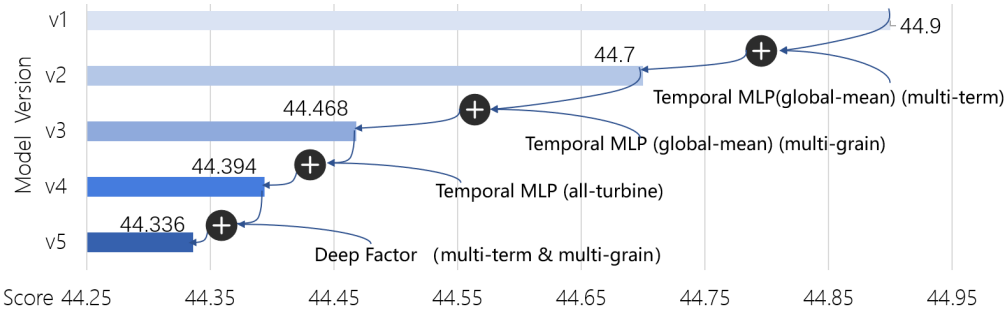

Figure 4: Model performance in the 2 phase.

The performance of different versions is reported in Figure 4. v2 largely outperforms v1, indicating that adding the multi-term prediction can greatly boost the model's prediction ability, especially the performance in short-term prediction. Compared v3 and v2, we can see that performance of v3 is significantly better than v2, which proves the effectiveness of the multi-grain prediction. Adding the multi-grain prediction can benefit the middle-term prediction.

v4 directly predicts the wind power of each turbine rather than the global average. In this way, more turbine-specific patterns can be captured by the model, thus improving the model's performance. At last, by adding the probabilistic forecasting model, the performance improves sightly, proving the effectiveness of modeling the uncertainties of wind power. And in the 3rd phase, the final score of our model is 45.173.

5 RELATED WORK

One of the common approaches for wind power forecasting is the physical model, such as Numerical Weather Prediction (NWP) [5], which predicts wind speed based on the atmosphere around and inside the wind farm, then forecasts the power by the relationship between wind speed and the output. Most studies consider wind power a time series and forecast with traditional statistical models, machine learning models, and deep learning models. Traditional statistical models include autoregressive moving average (ARMA) [11], autoregressive integrated moving average (ARIMA) [19], and fractional-ARIMA (f-ARIMA) [7]. ML contains least squares support vector machine (LSSVM) [16] and Kalman filter[1], etc. With the development of deep learning, more and more models use it to forecast wind power, such as long short-term memory (LSTM) [18, 20], gated recurrent unit(GRU) [10], etc. They are good at dealing with complex nonlinear problems and have better predictive power than traditional statistical models. Some researchers combine different models to describe different aspects of WP fluctuation, including stacking-based model [6] and weight-based model [12].

6 CONCLUSION

In this paper, we propose a hybrid model named H-STWPF, which can capture the spatial-temporal correlations and uncertainty for wind power forecasting. The model is fully developed based on paddlepaddle, and experiments in the Baidu KDD Cup 2022 show the effectiveness of the proposed model.

REFERENCES

[1] Federico Cassola and Massimiliano Burlando. 2012. Wind speed and wind energy forecast through Kalman filtering of Numerical Weather Prediction model output. *Applied energy* 99 (2012), 154–166.
[2] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
[3] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *International conference on machine learning (ICML)*. PMLR, 933–941.
[4] ZhaoYang Dong, Kit Po Wong, Ke Meng, Fengji Luo, Fang Yao, and Junhua Zhao. 2010. Wind power impact on system operations and planning. In *IEEE PES general meeting*. IEEE, 1–5.
[5] Kazutoshi Higashiyama, Yu Fujimoto, and Yasuhiro Hayashi. 2018. Feature extraction of NWP data for wind power forecasting using 3D-convolutional neural networks. *Energy Procedia* 155 (2018), 350–358.
[6] Yun Ju, Guangyu Sun, Quanhe Chen, Min Zhang, Huixian Zhu, and Mujeeb Ur Rehman. 2019. A model combining convolutional neural network and LightGBM algorithm for ultra-short-term wind power forecasting. *IEEE Access* 7 (2019), 28309–28318.
[7] Rajesh G Kavasseri and Krithika Seetharaman. 2009. Day-ahead wind speed forecasting using f-ARIMA models. *Renewable Energy* 34, 5 (2009), 1388–1393.
[8] Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
[9] Yanjun Ma, Dianhai Yu, Tian Wu, and Haifeng Wang. 2019. PaddlePaddle: An open-source deep learning platform from industrial practice. *Frontiers of Data and Domputing* 1, 1 (2019), 105–115.
[10] Zhewen Niu, Zeyuan Yu, Wenhu Tang, Qinghua Wu, and Marek Reformat. 2020. Wind power forecasting using attention-based gated recurrent unit network. *Energy* 196 (2020), 117081.
[11] S Rajagopalan and Surya Santoso. 2009. Wind power forecasting and error analysis using the autoregressive moving average modeling. In *2009 IEEE Power & Energy Society General Meeting*. IEEE, 1–6.
[12] Jingjing Song, Jianzhou Wang, and Haiyan Lu. 2018. A novel combined model based on advanced optimization algorithm for short-term wind speed forecasting. *Applied Energy* 215 (2018), 643–658.
[13] Bart C Ummels, Madeleine Gibescu, Engbert Pelgrum, Wil L Kling, and Arno J Brand. 2007. Impacts of wind power on thermal generation unit commitment and dispatch. *IEEE Transactions on energy conversion* 22, 1 (2007), 44–51.
[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[15] Yuyang Wang, Alex Smola, Danielle Maddix, Jan Gasthaus, Dean Foster, and Tim Januschowski. 2019. Deep factors for forecasting. In *International conference on machine learning*. PMLR, 6607–6617.
[16] Yun Wang, Jianzhou Wang, and Xiang Wei. 2015. A hybrid wind speed forecasting model based on phase space reconstruction theory and Markov model: A case study of wind farms in northwest China. *Energy* 91 (2015), 556–572.
[17] Yun Wang, Runmin Zou, Fang Liu, Lingjun Zhang, and Qianyi Liu. 2021. A review of wind speed and wind power forecasting with deep neural networks. *Applied Energy* 304 (2021), 117766.
[18] Ruiguo Yu, Jie Gao, Mei Yu, Wenhuan Lu, Tianyi Xu, Mankun Zhao, Jie Zhang, Ruixuan Zhang, and Zhuo Zhang. 2019. LSTM-EFG for wind power forecasting based on sequential correlation features. *Future Generation Computer Systems* 93 (2019), 33–42.
[19] Kalid Yunus, Torbjörn Thiringer, and Peiyuan Chen. 2015. ARIMA-based frequency-decomposed modeling of wind speed time series. *IEEE Transactions on Power Systems* 31, 4 (2015), 2546–2556.
[20] Jinhua Zhang, Jie Yan, David Infield, Yongqian Liu, and Fue-sang Lien. 2019. Short-term forecasting and uncertainty analysis of wind turbine power based on long short-term memory network and Gaussian mixture model. *Applied Energy* 241 (2019), 229–244.