Complementary Fusion of Deep Spatio-Temporal Network and Tree Model for Wind Power Forecasting (Team:HIK)

Linsen Li
Zhejiang University
Hikvision Research Institute
Hangzhou, China
lilinsen@hikvision.com

Qichen Sun
Zhejiang University
Hangzhou, China
sunqichen@zju.edu.cn

Dongdong Geng
Hikvision Research Institute
Hangzhou, China
gengdongdong@hikvision.com

Chunfei Jian
Hikvision Research Institute
Hangzhou, China
jianchunfei@hikvision.com

Dongen Wu
Zhejiang University
Hangzhou, China
wude@zju.edu.cn

Shiliang Pu
Hikvision Research Institute
Hangzhou, China
pushiliang.hri@hikvision.com

ABSTRACT

Wind power forecasting (WPF) is crucial for planning a power system with wind power integration; hence, it has received much attention recently. In WPF, spatial and temporal correlations significantly impact the forecasting performance. However, most existing WPF methods only focus on modeling spatial proximity relation while overlooking the multi-relational spatio-temporal dependence. Moreover, existing ensemble strategies used in WPF ignore the difference between sub-modules. To address these issues, this paper proposes a novel WPF method, called FDSTT. FDSTT consists of a multi-relational graph constructor module, a multi-relational graphs-based deep spatio-temporal module, a spatio-partitioned-time-phased tree module and a data-driven ensemble module. In the KDD Cup 2022 of Wind power prediction, the proposed FDSTT has win the 1st in the final Phase and 2nd in Phase 2. Furthermore, we has win 3rd place in Phase 1 with the primary version of FDSTT. Those results confirm that FDSTT can achieve consistently superior performance on different test datasets.

CCS CONCEPTS

• **Information systems** → **Data mining**.

KEYWORDS

Wind Power Forecasting; Spatio-temporal Network; Ensemble Model

ACM Reference Format:
Linsen Li, Qichen Sun, Dongdong Geng, Chunfei Jian, Dongen Wu, and Shiliang Pu. 2022. Complementary Fusion of Deep Spatio-Temporal Network and Tree Model for Wind Power Forecasting (Team:HIK). In *KDD '22: ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 14–18, 2022, Washington DC, U.S.*. ACM, New York, NY, USA, 7 pages. https://doi.org/xxx/xxxxxx.xxxxxx

KDD '22, August 14–18, 2022, Washington DC, U.S.

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/xxx/xxxxxx.xxxxxx

1 INTRODUCTION

Wind energy, an important energy alternative of fossil fuel, plays an important role in energy conservation and emission reduction. However, integrating wind power in a power system is challenging. This is because, the wind power supply is stochastic, and the stochastic character of wind power frustrates the planning and operation of the power system. To properly plan and operate the power system, estimating the wind power supply at different time scales, namely power forecasting (WPF), is important. With the WPF results, power grid dispatchers can adjust the dispatching plan in time, which improves the peak shaving capacity and reduces the tuning reserve capacity.

Recently, deep learning-based WPF methods have achieved superior performance. Existing deep learning-based WPF methods can be divided into two categories: time series forecasting-based methods and spatio-temporal model-based methods. Time series forecasting-based methods regard WPF as a time series forecasting problem, and adopt RNN-based [5, 14, 21] or transformer-based model [19] to complete forecasting. Unfortunately, this kind of method ignores the spatial information of wind farms. Correspondingly, spatio-temporal model-based methods fill this gap by means of graph neural networks, such as STGNN [20], DCRNN [15], Graph Wavenet [2]. However, those spatio-temporal model-based methods mostly focus on the short-term forecasting task while struggling under the long-term setting. Thus they do not fit our problem since our time scale in WPF is a long-term situation. Summarily, existing learning-based WPF methods have two following challenges.

Challenge I: *How to capture multi-relational spatio-temporal dependence between wind turbines?* Most existing models only utilize a single relation between turbines (e.g., Euclidean distance) to construct the adjacency graph, which ignores the complicated relation between nodes. For example, a wind turbine may be affected by both nearby and distant wind turbine statues due to geographical influences and temporal similarity influences, respectively.

Challenge II: *How to ensemble models to fit variable data distributions?* Most ensemble strategies ensemble the models statically. Nevertheless, the fixed ensemble strategy may be suitable for a special data distribution while performs poor in other data distributions. Therefore, a flexible ensemble strategy to adapt different data distribution is needed.

To tackle the above challenges, we present a novel WPF method, called FDSTT. To handle the first challenge, FDSTT embeds a multi-relation graph constructor in the deep spatio-temporal module. This constructor simultaneously constructs the spatial-aware graph and semantic-aware graph, which can capture both the local and global contextual information. To overcome the second challenge, FDSTT dynamically adjusts the model structure according to the data distributions as different models may be good at handling different data distributions, which enables our model to be relative optimal for different situations. Moreover, FDSTT integrates a tree-based model to further enhance the prediction performance.

The proposed FDSTT has win the 1st place in the Phase 3 of the Baidu KDD CUP 2022 of WPF, which solves a real-world wind power forecasting problem: estimating the wind power supply of a wind farm in 2 days without wind speed forecasting data. In this competition, a unique spatial dynamic wind power forecasting (SD-WPF) dataset is provided, which includes the spatial distribution of wind turbines, as well as the dynamic context factors like temperature, weather, and turbine internal status[23]. When evaluated on this dataset, our ensemble model FDSTT consistently outperforms other competing models online.

Overall, our contributions are summarized as follows:

- We propose an ensemble framework for long-term wind power forecasting, including a Deep Multi-relational Spatio-Temporal (DMST) module to effectively capture spatio-temporal correlations, a Spatio-partitioned Time-phased Tree (ST-Tree) module to improve prediction robustness, and a ensemble module to integrate the prediction signals between modules.
- We present a multi-relational graph constructor to capture the multi-relational dependencies among wind turbines. Based on this, the downstream models are able to leverage both the local and global contextual information.
- We design a data-driven ensemble strategy, which can dynamically adjust the ensemble architecture to fit the input distribution.
- In the KDD Cup 2022 of WPF, the proposed FDSTT has win the 1st place in the final Phase. Furthermore, FDSTT and its primary version have won 2nd and 3rd place in Phase 2 and Phase 1, respectively. Note that, FDSTT is the only solution that remains in the top three during all phases.

2 PROBLEM FORMULATION

In this section, we formally define Turbine Time Series and Spatial Correlation Graph in Definitions 1 and 2, respectively. Then, we provide a formal statement of the Spatial Dynamic Wind Power Forecasting problem.

Definition 1. **(Turbine Time Series)** *Suppose there are N wind turbines, each of which generates time series $x_i \in \mathbb{R}^{T\times D}$, where i is the turbine id, T is the number of timestamps, and D denotes the number of features. $\boldsymbol{X} = \{x_1, x_2, \ldots, x_N\} \in \mathbb{R}^{N\times T\times D}$ represents multivariate time series generated by all turbines. Additionally, we denote $x_i^t \in \mathbb{R}^D$ the status of turbine i at timestamp t.*

The wind power generated by each turbine depends not only itself status but also adjacent turbine's. Hence, we use a graph to capture the spatial correlation, as defined below.

Definition 2. **(Spatial Correlation Graph)** *A graph $\mathcal{G} = (V, E)$ is used to capture the spatial correlation, where $\mathcal{V} = \{v_i\}_{i=1}^N$ ($|\mathcal{V}| = N$) is a set of vertices (i.e., wind turbines) and E is the edge set. An edge $e_{i,j}$ is associated with a spatial distance from v_i to v_j.*

Note that, the graph can also be denoted as an adjacent matrix $A \in \mathbb{R}^{N\times N}$ to capture the spatial correlation.

Problem Statement. The wind power forecasting problem is to predict the future wind power generation. Let $\boldsymbol{X}^t \in \mathbb{R}^{N\times D}$ be the history turbine statues observed at timestamp t. Following previous multivariate time series forecasting prediction solutions, we formulate the problem as learning a function $\mathcal{F}_\theta$ to forecast the next τ steps data based on the past T steps historical data:

$$\left\{\hat{Y}^{t+1}, \hat{Y}^{t+2}, \ldots, \hat{Y}^{t+\tau}\right\} = \mathcal{F}_\theta\left(\boldsymbol{X}^t, \boldsymbol{X}^{t-1}, \ldots, \boldsymbol{X}^{t-T+1}; \mathcal{G}\right) \quad (1)$$

Here, $\hat{Y}^t \in \mathbb{R}^{N\times 1}$ denotes the predicted wind power for all turbines at timestamp t, and $\mathcal{G} = (V, E)$ represents spatial adjacent matrix.

3 THE PROPOSED APPROACH

3.1 Overview

The overall framework of FDSTT is illustrated in Figure 1. It consists of four modules: (1) Multi-relational Graph Constructor, (2) DMST module, (3) ST-Tree module, (4) Ensemble module. Specifically, (1) the Multi-relational Graph Constructor generates multi-relational graphs for spatio-temporal information aggregation; (2) the DMST module aggregates the spatial information and captures temporal patterns in an Encoder-Decoder form; (3) the ST-Tree module extracts partition features, and makes predictions for each time segment to improve the accuracy and robustness; (4) the Ensemble module dynamically fuses the output of each module in a data-driven ensemble strategy. The details of these modules are introduced in the following subsections.

3.2 Multi-relational Graph Constructor

As discussed in Challenge I, the spatio-temporal dependence of wind power is highly correlated with the complicated relations between turbines. In view of this, we present the multi-relational graph constructor to extract multi-view graph information among turbines for subsequent spatio-temporal deep learning.

3.2.1 Spatial-aware Graph.

The spatial correlation between wind turbines will effectively improve the prediction efficiency. Thus, we first build a spatial-aware graph G_d to capture the explicit neighboring relations in a local view. To obtain the G_d, we calculate the Euclidean distance between two nodes to get the spatial distance matrix, then take the $top - K$ nearest nodes as the neighbors of node i, denoted as $N(i)$. Hence, G_d is formulated below:

$$A(i,j) = \begin{cases} 1, & j \in N(i) \\ 0, & j \notin N(i) \end{cases} \quad (2)$$

3.2.2 Semantic-aware Graph.

Although spatial neighbors can be used to capture the local scope effects, they are limited since the environments can also vary over short distances. As an example, one turbine on the back slope and the other one on the windward slope, in this case their acceptable wind direction is opposite but the

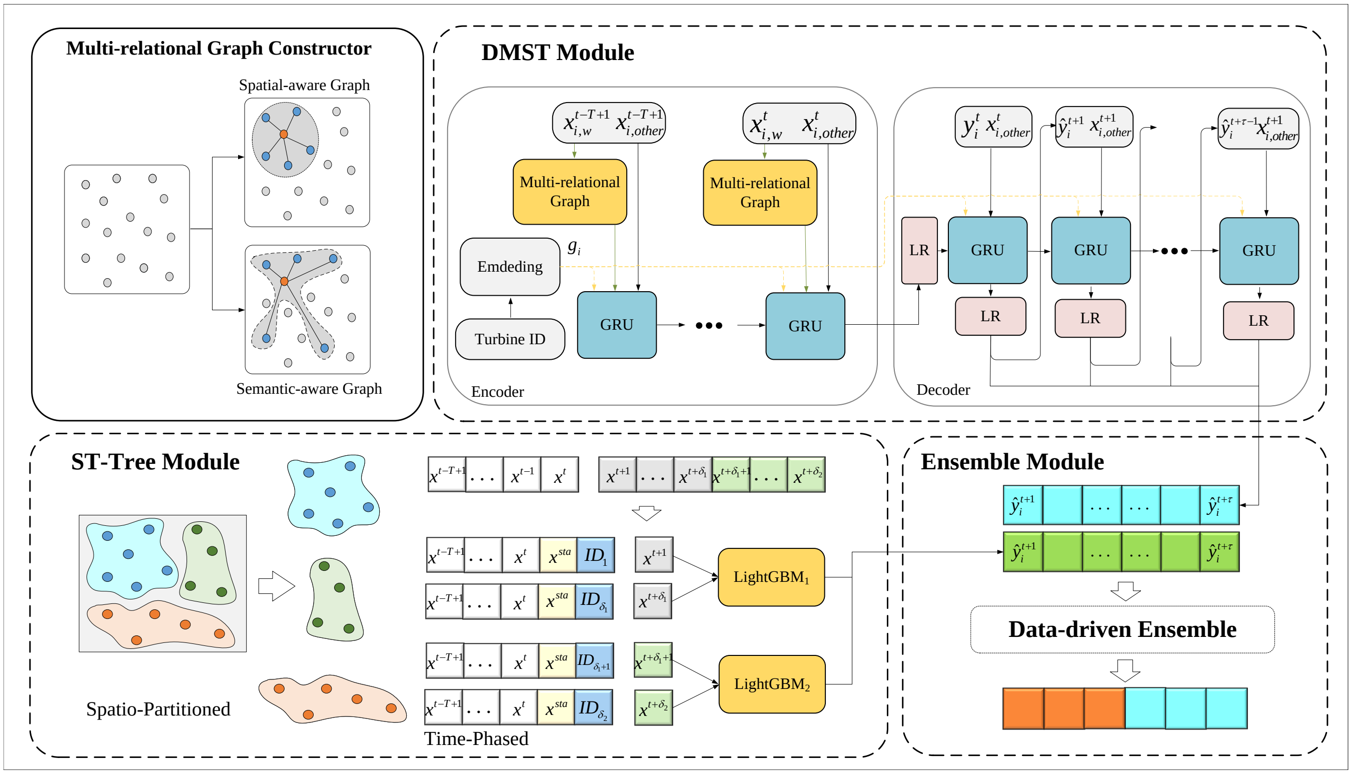

Figure 1: The overall architecture of FDSTT. Multi-relational Graph Constructor uses input turbine series to generate the graphs. The encoder in DMST utilizes the generated graphs, turbine id embedding and original turbine series to aggregate the spatio-temporal information, then predicts the wind power in decoder autoregressively. ST-Tree module predicts short-term future power for each partition. Finally, Ensemble module dynamically fuses the output of DMST and ST-Tree as the prediction.

relative distance between them is small. To this end, we further construct a semantic-aware graph G_s, to capture the similar temporal pattern between turbines in a global view. In this term, we propose the differential similarity rather than directly calculations of the similarity over origin sequences. The differential similarity can be explicitly captured via the similarities between sequence variation patterns, while the global semantic neighbors can be integrated to break through the local shackles. Let $x_{i,w} \in \mathbb{R}^{T\times 1}$ represent the wind speed sequence of turbine i, we can calculate the differential similarity as follows:

$$Sim(i,j) = \sum_{t=1}^{T} \left((x_{i,w}^{t} - x_{i,w}^{t-1}) \cdot (x_{j,w}^{t} - x_{j,w}^{t-1}) \right) \tag{3}$$

After computing the similarity between nodes, we obtain the $top-K$ most similar nodes as semantic neighbors for each node, and define the G_s the same as G_d does.

3.3 Deep Multi-relational Spatio-Temporal Network

3.3.1 Spatial Features. The spatial-aware graph and the semantic-aware graph reflect the internode correlations from distinct perspectives. To enhance the performance of WPF, we combine spatial-aware graph and the semantic-aware graph when capturing spatial features. For each wind turbine, we aggregate the wind speed information from 1-hop neighbors in both G_s and G_d, which helps target turbine to extract a more informative representation. The aggregation process can be formulated as:

$$\hat{x}_{i,w}^{t}(d) = \mathrm{CONCAT}(x_{i,w}^{t}, \mathrm{AGGREGATE}(x_{u,w}^{t}, u \in \mathcal{N}_d(i))) \tag{4}$$

$$\hat{x}_{i,w}^{t}(s) = \mathrm{CONCAT}(x_{i,w}^{t}, \mathrm{AGGREGATE}(x_{u,w}^{t}, u \in \mathcal{N}_s(i))) \tag{5}$$

$$\hat{x}_{i,w}^{t} = \mathrm{CONCAT}(\hat{x}_{i,w}^{t}(d), \hat{x}_{i,w}^{t}(s)) \tag{6}$$

where $x_{i,w}^{t}$ is the wind speed of wind turbine i at timestamp t. $\mathcal{N}_d(i)$ is the set of neighbor nodes of node i in graph G_d. $\mathcal{N}_s(i)$ is the set of neighbor nodes of node i in graph G_s. $\hat{x}_{i,w}^{t}$ is the aggregated spatial features of wind turbine i at timestamp t.

3.3.2 Turbine Embedding. If the prediction model is built for each wind turbine, it will incur problems such as too many models and poor generalization of models. Therefore, it is more appropriate to build deep learning model based on the wind farm rather than a single wind turbine. To make power prediction for a specific wind turbine, and allow the model to share parameters between wind turbines, we need to provide the model which wind turbine the current data comes from through the model input. We represent each turbine with a learnable latent vector by adding an embedding layer to the model. The embedding layer can reduce the dimension of the wind turbine one-hot sparse vector representation into a dense vector representation.

$$g_i = e_i E, E \in \mathbb{R}^{N\times d_E} \tag{7}$$

where g_i is the latent vector of the ith wind turbine, E is learnable embedding matrix, e_i is one-hot verctor of the ith wind turbine, d_E is the latent vector dimension, N is the number of wind turbines in the wind farm. After model training, the more similar the latent vectors, the more similar the power patterns of the wind turbine.

3.3.3 Temporal Features. The sequence-to-sequence (Seq2Seq) model based on recurrent neural network (RNN) is widely used in multi-horizon sequence forecasting scenarios. Seq2Seq is mainly composed of RNN-based encoder and decoder, and RNN is commonly used with its variants: long short-term memory network (LSTM) or gated recurrent unit (GRU). GRU uses update gate to adjust the balance between forget gate and input gate in LSTM, avoiding the redundancy of forget gate and input gate. Thus, it is simpler in network structure than LSTM. In our model, GRU is used to build Seq2Seq model.

In terms of model encoder input, we assume that the wind turbine features corresponding to timestamp $\tilde{t} \leq t$ are available at timestamp t. We use the T-length time window $[t-T+1, \cdots, t-1, t]$ to capture the temporal features. For wind turbine i at timestamp t, we first extract the spatial feature of the wind speed $x_{i,w}^t$ to obtain $\hat{x}_{i,w}^t$, and then concat it with the wind turbine embedding g_i and other wind turbine feature $x_{i,other}^t$ (e.g., temperature, weather, turbine internal states, etc.). The function is as follows:

$$x_i^t = \mathrm{CONCAT}(g_i, \hat{x}_{i,w}^t, x_{i,other}^t) \tag{8}$$

$$h_i^t = \mathrm{GRU}(x_i^t, h_i^{t-1}) \tag{9}$$

where h_i^t is hidden feature at timestamp t for wind turbine i.

In terms of model decoder input, we use the h_t^i of the encoder as the initial hidden state of the GRU in the decoder, and then concat predicted power $\hat{y}_i^{t+1}$ at the previous timestamp t with the wind turbine embedding and other known future feature (e.g., time and wind turbine position) as model input at timestamp $t+1$. The function is below:

$$x_i^{t+1} = \mathrm{CONCAT}(g_i, \hat{y}_i^{t+1}, x_{i,other}^{t+1}) \tag{10}$$

$$h_i^{t+1} = \mathrm{GRU}(x_i^{t+1}, h_i^t) \tag{11}$$

$$\hat{y}_i^{i+2} = \mathrm{LR}(h_i^{t+1}) \tag{12}$$

where $\hat{y}_i^{i+2}$ is the predicted power value at timestamp $t+2$. The decoder autoregressively predicts the wind turbine power for τ time steps.

3.4 Spatio-Partitioned Time-Phased Tree Model

3.4.1 Tree Model. Tree models are widely used in various scenarios because of good robustness. Gradient Boosting Decision Tree (GBDT) is an iterative-based decision tree algorithm. The basic idea is to set the base learner as a decision tree, use the negative gradient information of the loss function to iteratively generate the base learner f_k, and accumulate the trained base learners to form the final model. Given a data set $D = \{x_i, y_i\}$, GBDT learns K tree via fitting the pseudo-residuals of all the previous trees. The loss function in the training process is as:

$$L(\Phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \tag{13}$$

where y_i is the ground truth of the ith sample, $\hat{y}_i$ is the predicted value of the ith sample, f_k is the ith decision tree, and l is the loss function (commonly used losses for regression problems are *MSE* and *MAE*). Ω is the regular term, and the regular term penalizes the complexity of each decision tree.

There are already second-order variant models based on GBDT such as XGBoost [4], LightGBM [12], and CatBoost [17]. Due to LightGBM's fast training speed and good ability to handle noisy data, we use LightGBM to build a spatio-partitioned time-phased tree model (ST-Tree).

3.4.2 Spatio-Partitioned. In the spatial dimension, considering the wake effect of the wind farm and the location distribution of wind turbines, wind turbines under similar wind conditions are likely to have similar states. Thus, we cluster the wind turbines in the wind farm based on the wind speed similarity, and build a separate tree model for each cluster partition, which not only improves the accuracy but also increases the robustness of the model. Here, we use pearson correlation coefficient to calculate the similarity between turbines, formulated as:

$$\mathrm{PearsonSim}(i,j) = \frac{\sum_{t=1}^{T}\left(x_{i,w}^t - \bar{x}_{i,w}\right)\left(x_{j,w}^t - \bar{x}_{j,w}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{i,w}^t - \bar{x}_{i,w}\right)^2\left(x_{j,w}^t - \bar{x}_{j,w}\right)^2}} \tag{14}$$

where $x_{i,w}^t$ represents the wind speed feature of turbine i at timestamp t, and $\bar{x}_{i,w}$ denotes the average wind speed of turbine i. We utilize K-means to spatially cluster the wind farm based on the wind speed similarity for each wind turbine pair.

3.4.3 Time-Phased. In the temporal dimension, considering that this forecasting task belongs to a long-term series forecasting problem, the tree model only supports single-output forecasting, which is different from the Seq2Seq model. Building a tree model for each timestamp would lead to too many models and the risk of model overfitting. Therefore, tree models are constructed by segmenting time steps, as shown in Figure 1. We build a tree model at predicted timestamp $t+1$ to $t+\delta$, and add a column of unique ID feature to the sample features in order to distinguish different timestamps. Further, the wind turbine features of the previous day or week before the predicted timestamps can also be input as model features.

The advantage of the time-phased tree model is that it may effectively reduce the number of models, and can also enhance the robustness of the model according to different time-phased modes.

3.5 Ensemble Model

In the regression task, the model ensemble methods mainly include the averaging method and the learning method. The average method is to directly perform a simple average or weighted average for the predicted values of each model. The learning method is to use the meta learner to learn the combination strategy according to the data instead of artificially specifying, mainly including Blending and Stacking. Since the wind turbine data distribution is unstable, the learning method could cause the model overfitting, and the simple weighted average fusion strategy cannot be applied to the variable wind turbine data distribution. Hence, we propose a data-driven ensemble strategy such as Algorithm 1, based on the following characteristics of long-term WPF:

Algorithm 1: Data-driven Ensemble Strategy

Input: $X_p \in \mathbb{R}^{N\times T\times 1}$ is the power value of model input, $\hat{Y}_1 \in \mathbb{R}^{N\times \tau\times 1}$ is the predicted value of DMST, $\hat{Y}_2 \in \mathbb{R}^{N\times \tau\times 1}$ is the predicted value of tree model, ϕ is the baseline value, Δ_{low} and Δ_{up} are the judgment threshold of the current power level, β is timestamp.
Output: $\hat{Y}$ is the predicted value

```
Initialize s = Mean(X_p[:, −5 :, :])
Initialize Ŷ = 0
if s < Δ_low or s > Δ_up then
    Ŷ[:, : β, :] = Ŷ_2[:, : β, :]
else
    Ŷ[:, : β, :] = 0.5 × Ŷ_2[:, : β, :] + 0.5 × Ŷ_1[:, : β, :]
Ŷ[:, β :, :] = Ŷ_1[:, β :, :] + φ
```

- Different models have different predictive ability for different time periods. The ability of the ST-Tree model for short-term forecasting is better than that of the DMST, and the ability of the DMST for medium-term and long-term prediction is better than that of the ST-Tree model.
- When the current wind power is too large or too small, the short-term forecasting curve of the DMST is steep, otherwise the effect of the DMST in the short-term prediction is also available.
- DMST learns accurately wind power trends in medium-term and long-term prediction, while the baseline prediction is less accurate. When the current wind power is large, the predictions of DMST are conservative.

4 EXPERIMENTS

4.1 Experimental Settings

4.1.1 Datasets. In the Baidu KDD CUP 2022, organizers provide a unique spatial dynamic wind power forecasting (SDWPF) dataset. The SDWPF dataset is collected from the Supervisory Control and Data Acquisition (SCADA) system of a wind farm. The SCADA data are sampled every 10 minutes from each wind turbine in the wind farm which consists of 134 wind turbines. The SDWPF dataset contains 4,727,520 data records for 245 days. The dataset includes critical external features and essential internal features which can indicate the operating status of each wind turbine, and the relative positional distribution of all wind turbines in the wind farm is also provided.

4.1.2 Parameter Settings. In the multi-relational graph constructor module, K is set to be 5 in both spatial-aware graphs and semantic-aware graphs. In the DMST module, the model is trained by means of the Adam optimizer [13]. The learning rate is set to be 0.001. The encoder GRU hidden size is 64, and the decoder GRU hidden size is 32. The wind turbine embedding size is 5. The historical time steps length T is 432, and predicted time steps length τ is 288. In the encoder, $x^t_{i,other}$ contains ten features: Etmp, Itmp, Wdir, Ndir, Pab1, Prtv, Patv (the meaning of the columns is found in [23]), Hour and turbine position. In the decoder, $x^t_{i,other}$ includes two features: hour and turbine position. In the ST-Tree module, the number of spatial partitions is 3, the number of temporal phase is 1, and its length δ_1 is 30. The tree model is built based on LightGBM, and the learning rate is 0.1. Besides, the depth of the tree is 15, and the number of nodes tree is 20. When training the model, we adopt wind speed and power at timestamps from $t-30+1$ to t, while we obtain Etmp, Itmp, Wdir, Ndir, Pab, Prtv, TurbID and hour at timestamp t. Also, we adopt statistical features (max, min, mean and median) over last 14 days and the predicted timestamp ID. In the ensemble module, Δ_{low} is set to 350 and Δ_{up} is set to 700 in the data driven ensemble strategy. Besides, ϕ is 15, and β is 30.

Table 1: Online scores with different models. The footnotes (e.g., 1st) in the table denote the online rank of the corresponding method in Baidu KDDCup 2022.

Method	Phase 1	Phase 2	Phase 3
AutoFormer	45.5570	–	–
SCINet	46.4679	–	–
AGCRN	41.3100	–	–
GRU (Baseline)	42.3019	46.9968	–
ST-Tree	**40.7903** (3rd)	45.1745	–
GWNET	–	48.8300	–
DCRNN	–	47.3043	–
ASTGCN	–	48.0889	–
DST	–	44.4205	–
DMST	–	44.2845	–
FDSTT (w/o avg)	–	44.0942	45.0405
FDSTT (w/o ϕ)	–	44.0732	45.0169
FDSTT	–	**44.0536** (2nd)	**44.9171** (1st)

4.2 WPF Performance

In this subsection, we report the WPF performance of FDSTT on both the official KDD CUP 2022 test dataset and the self-constructed test dataset. The results are presented in Sections 4.2.1 and 4.2.2, respectively.

4.2.1 Performance on Official KDD CUP 2022 Test Dataset. To verify the effectiveness of our proposed FDSTT method, we compare it with time series forecasting and spatio-temporal forecasting methods, including GRU, AutoFormer, SCINet, GWNET, AGCRN, DCRNN, ASTGCN and DST. Table 1 depicts the results obtained on the official KDD CUP 2022 test dataset for the three phases. From this table, we can observe that our proposed FDSTT model outperforms the classic models during all the three phases, which verifies the superiority of our methods. Moreover, FDSTT has win the 1st in the final Phase and the 2nd in Phase 2 of the KDD CUP 2022. Besides, the primary version of FDSTT has win the 3rd place in Phase 1. These results demonstrate that FDSTT can achieve consistently superior performance over different test datasets.

4.2.2 Performance on Self-constructed Test Dataset. We have constructed an offline test dataset to complete fast evaluation. This offline test dataset contains randomly sampled data from the last 30 days in the training set. Specifically, we randomly set 50 time points, and then, we extract the data of past 14 days and future 2 days for each time point to construct the offline test dataset.

Using this self-constructed test dataset, we compare FDSTT with the baselines, including GRU, AutoFormer, SCINet, GWNET, AGCRN, DCRNN, ASTGCN, and DST. The comparison results are reported

Table 2: The performance of different methods.

Method	MAE	RMSE	Score
GRU (baseline)	37.0174	47.0442	42.0308
AutoFormer	42.8972	54.3160	48.6066
SCINet	40.2794	47.8892	44.0843
GWNET	43.1812	55.1324	49.1568
AGCRN	40.9344	52.4761	46.7052
DCRNN	39.6786	48.4222	44.0504
ASTGCN	37.9236	46.2108	42.0672
DST	36.8172	46.1150	41.4661
FDSTT (ours)	**36.3872**	**45.5246**	**40.9559**

Table 3: Ablation study of different modules of FDSTT .

Method	MAE	RMSE	Score
DMST	36.8119	46.0136	41.4128
w/o mg	36.8172	46.1150	41.4661
ST-Tree	37.0405	46.3101	41.6753
w/o sp	37.3759	46.3857	41.8808
Ensemble	36.3872	45.5246	40.9559
w/o ϕ	36.4568	45.6260	41.0414
w/o avg	36.3410	45.8520	41.0965

in Table 2. Obviously, our method achieves the best performance. The baselines have inferior performance because:

- The time series forecasting models (namely, GRU, AutoFormer, SCINet) ignore the spatial features of wind turbines.
- The CNN-based spatio-temporal methods (viz..., GWNET, ASTGCN), are not good at long-term time series forecasting when the periodicity of the data is weak.
- The RNN-based spatio-temporal models (i.e., DCRNN, AGCRN) are not suitable for spatial aggregation of all features of wind turbines, and wind speed should be used instead.
- The baselines ignore the unstable data distribution, in which case it is necessary to control the complexity of the model and propose an effective ensemble strategy.

4.3 Analysis

To evaluate the effect of the key components for our proposed model, we conduct an ablation study on these components based on our self-constructed test dataset. The ablation settings are as follows:

- **w/o mg**: DMST whose graph is generated based on only spatial distance and without considering semantic neighbors.
- **w/o sp**: ST-Tree model whose spatial dimension is not partitioned by cluster.
- **w/o ϕ**: Model ensemble strategy without adding baseline value ϕ according to the current data distribution.
- **w/o avg**: Model ensemble strategy without averaging model predictions.

The results are shows in Table 3. From this table, it is observed that all the components are indispensable for the superior performance of our proposed method.

5 RELATED WORK

Wind power forecasting has been extensively studied over the past decades. Related studies can be divided in to five categories: physical modeling methods, statistical model methods, machine learning methods, deep learning methods, and the combination methods.

Physical modeling methods use physical models to predict wind power by simulating wind, pressure, air density [8, 9], which have achieved proper performance in the mid-long term WPF. The relationship between wind power and meteorological factors is $P = C_p A \rho v^3/2$, where P is the output power of wind turbines, v is the wind speed A is rotor swept area, ρ is the air density, C_p is the power factor of wind turbine. Wind speed is a key factor as it varies substantially. Administrators of a wind farm will change pitch angle of blade to protect wind turbine under certain circumstances.

Several researchers regard WPF as a time-series problem, and use statistical models to predict wind power [7, 8, 10, 18]. Machine learning methods are also widely used such as XGBoost, SVR, RF or combinations of these methods [3, 6]. Deep learning methods also become popular in recent years [1, 2, 5, 11, 14, 15, 20, 21], e.g., GRU, Graph Wavenet, AGCRN, DCRNN, STGNN, and DST. Besides these single models or methods, researchers also combine different models together to achieve proper prediction performance [16, 22].

6 CONCLUSION

In this paper, we propose a novel ensemble model FDSTT for long-term WPF, which consists of a multi-relational graph constructor module, a deep multi-relational graphs-based spatial-temporal module, a spatio-partitioned-time-phased tree module, and a data-driven ensemble module. In the multi-relational graph constructor module, we construct multi-relational graphs based on distance neighbors and semantic neighbors. In the deep multi-relational spatio-temporal network (DMST), we use wind speed data for spatial feature extraction via a multi-relational graph, and adopt a GRU-based Seq2Seq model to capture the temporal features. In the spatio-partitioned time-phased tree model (ST-Tree), the wind farm is spatially clustered based on the correlation of the wind speed data, and the model is built by time phase instead of each timestamp. It not only effectively reduces the number of models, but also enhances the model robustness. Based on this model ensemble strategy, we propose a data driven ensemble strategy for unstable data distribution according to the characteristics of different models in WPF. Finally, our proposed model FDSTT won the 1st in the final Phase of the KDD Cup 2022 competition and the 2nd in the Phase 2 of the KDD Cup 2022 competition.

ACKNOWLEDGMENTS

The authors would like to acknowledge Yunjun Gao, Yuren Mao, Lu Chen, Ziquan Fang, Pengfei Wang, and Yuntao Du from Zhejiang University as well as Feng Wan and Weihao Jiang from Hikvision Research Institute, for their wonderful collaboration and patient support. Also, this work was supported in part by the National Key Research and Development Program of China under Grant No. 2021YFC3300303, and the NSFC under Grants No. (62025206, 61972338, and 62102351). Shiliang Pu is the advisor of the work.

REFERENCES

[1] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. 2020. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in neural information processing systems* 33 (2020), 17804–17815.

[2] Xu Bingbing, Shen Huawei, Cao Qi, Qiu Yunqi, and Cheng Xueqi. 2019. Graph wavelet neural network. In *The International Conference on Learning Representations.*

[3] Aditya Chaudhary, Akash Sharma, Ayush Kumar, Karan Dikshit, and Neeraj Kumar. 2020. Short term wind power forecasting using machine learning techniques. *Journal of Statistics and Management Systems* 23, 1 (2020), 145–156.

[4] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, et al. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2* 1, 4 (2015), 1–4.

[5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encode-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* 1724–1734.

[6] Halil Demolli, Ahmet Sakir Dokuz, Alper Ecemis, and Murat Gokcek. 2019. Wind power forecasting based on daily wind speed data using machine learning algorithms. *Energy Conversion and Management* 198 (2019), 111823.

[7] Mario J Duran, Daniel Cros, and Jesus Riquelme. 2007. Short-term wind power forecast based on ARX models. *Journal of Energy Engineering* 133, 3 (2007), 172–180.

[8] M De Felice, A. Alessandri, and P. M. Ruti. 2013. Electricity demand forecasting over Italy: Potential benefits using numerical weather prediction models. *Electric Power Systems Research* 104, Complete (2013), 71–79.

[9] Ulrich Focken, Matthias Lange, and Hans-Peter Waldl. 2001. Previento-a wind power prediction system with an innovative upscaling algorithm. In *Proceedings of the European wind energy conference, Copenhagen, Denmark*, Vol. 276.

[10] Bri-Mathias Hodge, Austin Zeiler, Duncan Brooks, Gary Blau, Joseph Pekny, and Gintaras Reklatis. 2011. Improved wind power forecasting with ARIMA models. In *Computer Aided Chemical Engineering*. Vol. 29. Elsevier, 1789–1793.

[11] Tianyu Hu, Wenchuan Wu, Qinglai Guo, Hongbin Sun, Libao Shi, and Xinwei Shen. 2019. Very short-term spatial and temporal wind power forecasting: A deep learning approach. *CSEE Journal of Power and Energy Systems* 6, 2 (2019), 434–443.

[12] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).

[13] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference for Learning Representations.*

[14] Jiangyuan Li and Mohammadreza Armandpour. 2022. Deep Spatio-Temporal Wind Power Forecasting. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 4138–4142.

[15] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2018. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *The International Conference on Learning Representations.*

[16] Xiaosheng Peng, Hongyu Wang, Jianxun Lang, Wenze Li, Qiyou Xu, Zuowei Zhang, Tao Cai, Shanxu Duan, Fangjie Liu, and Chaoshun Li. 2021. EALSTM-QR: Interval wind-power prediction model based on numerical weather prediction and deep learning. *Energy* 220 (2021), 119692.

[17] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems* 31 (2018).

[18] Shuxin Tian, Yang Fu, Ping Ling, Shurong Wei, Shu Liu, and Kunpeng Li. 2018. Wind power forecasting based on arima-lgarch model. In *2018 International Conference on Power System Technology (POWERCON).* IEEE, 1285–1289.

[19] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems* 34 (2021), 22419–22430.

[20] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence.*

[21] Ceyhun Yildiz, Hakan Acikgoz, Deniz Korkmaz, and Umit Budak. 2021. An improved residual-based convolutional neural network for very short-term wind power forecasting. *Energy Conversion and Management* 228 (2021), 113731.

[22] Yu Zhang, Yanting Li, and Guangyao Zhang. 2020. Short-term wind power forecasting approach based on Seq2Seq model using NWP data. *Energy* 213 (2020), 118371.

[23] Jingbo Zhou, Xinjiang Lu, Yixiong Xiao, Jiantao Su, Junfu Lyu, Yanjun Ma, and Dejing Dou. 2022. SDWPF: A Dataset for Spatial Dynamic Wind Power Forecasting Challenge at KDD Cup 2022. *Techincal Report* (2022).