M2GSNet: Multi-Modal Multi-Task Graph Spatiotemporal Network for Ultra-Short-Term Wind Farm Cluster Power Prediction

Hang Fan 1, Xuemin Zhang 1,*, Shengwei Mei 1, Kunjin Chen 1 and Xinyang Chen 2

1 State Key Laboratory of Power System and Generation Equipment, Department of Electrical Engineering, Tsinghua University, Beijing 100084, China; fanh15@mails.tsinghua.edu.cn (H.F.); meishengwei@mail.tsinghua.edu.cn (S.M.); chenkj17@mails.tsinghua.edu.cn (K.C.)
2 School of Software, Tsinghua University, Beijing 100084, China; chenxiny17@mails.tsinghua.edu.cn

* Correspondence: zhangxuemin@tsinghua.edu.cn

Received: 18 October 2020; Accepted: 6 November 2020; Published: 8 November 2020

Abstract: Ultra-short-term wind power prediction is of great importance for the integration of renewable energy. It is the foundation of probabilistic prediction and even a slight increase in the prediction accuracy can exert significant improvement for the safe and economic operation of power systems. However, due to the complex spatiotemporal relationship and the intrinsic characteristic of nonlinear, randomness and intermittence, the prediction of regional wind farm clusters and each wind farm’s power is still a challenge. In this paper, a framework based on graph neural network and numerical weather prediction (NWP) is proposed for the ultra-short-term wind power prediction. First, the adjacent matrix of wind farms, which are regarded as the vertexes of a graph, is defined based on geographical distance. Second, two graph neural networks are designed to extract the spatiotemporal feature of historical wind power and NWP information separately. Then, these features are fused based on multi-modal learning. Third, to enhance the efficiency of prediction method, a multi-task learning method is adopted to extract the common feature of the regional wind farm cluster and it can output the prediction of each wind farm at the same time. The cases of a wind farm cluster located in Northeast China verified that the accuracy of a regional wind farm cluster power prediction is improved, and the time consumption increases slowly when the number of wind farms grows. The results indicate that this method has great potential to be used in large-scale wind farm clusters.

Keywords: graph neural network; multi-modal learning; multi-task learning; NWP; wind power prediction; wind farm cluster

1. Introduction

Renewable energy, especially wind energy, has become the key to alleviating the energy problem. The installed capacity of wind power is also increasing year by year and most wind farms are integrated into grids in the form of large-scale clusters. Due to the fluctuation and intermittence of wind, wind power not only provides clean energy, but also brings severe challenges to the safe and stable operation of power systems. Accurate prediction of wind speed and wind power is a fundamental requirement and basic task to ensure the grid connection of wind power [1].

There is a tremendous amount of research about the ultra-short-term wind power prediction, which can be divided into two types of methods: Physical method and Statistic Learning method. The physical method models the wind behavior according to the equation of atmosphere movement and can simulate the nonlinear characteristic of wind process. However, the parameter of the physical
model is not easy to be obtained and it also leads to expansive computational costs. So, it is not suitable for the ultra-short-term wind power prediction directly.

The statistic learning methods can be roughly divided into four categories. The first kind of prediction method is the classical multivariate time series prediction method, which has a relatively solid statistical theoretical foundation. In essence, it expands the traditional Autoregressive Integrated Moving Average model (ARIMA) on the multivariate time series and adopts the lasso method to select the key characteristic variables [2,3]. The advantage is that the mathematical meaning is clear, and the model parameters are very convenient to adjust. Therefore, it is suitable for online application scenarios and is often used in practical fields. However, its disadvantage is that it is a linear model and does not consider the intrinsic relationships among variables. It often leads to larger error than other methods. The second kind of prediction method adopts the idea of the probabilistic graph model and models wind process as the Gaussian process [4,5]. It uses the Gaussian function as the kernel function and can model the nonlinear data. It also can provide the confidence interval of the model and output the probabilistic distribution directly. However, the Gaussian process is a non-parametric model and every inference has to use each sample for the inverse calculation. It is intractable when the data volume is large. It is more precise than the first method, but it is still hard to capture the complex spatiotemporal relationship of wind process. The third kind of prediction method is the AI method. It includes both the traditional machine learning method and modern deep learning method. The machine learning method includes Back Propagation (BP) neural network [6], Decision Tree (DT) and its advanced derivatives Xgboost [7], and Extreme Learning Machine (ELM) [8]. The deep learning method is famous for its ability to extract the abstract features and different researchers use different models to predict the related energy time series. It includes Long Short-term Memory networks (LSTM) [9], Long- and Short-term Time-series network (LSTNet) [10] and so on. The advantage is that the accuracy of the model is higher than the two methods above when the structure of the model is well designed. The disadvantage is that the training time is relatively long and Graphics Processing Unit (GPU) is needed, which is suitable for offline training models and online application. The fourth kind of prediction method is the hybrid method. It often relates to the combination of different machine learning or deep learning methods and there are many varieties [11]. Some methods also decompose time series into several more predictable components by the empirical mode decomposition or variational mode decomposition, and then establish prediction method for each decomposed subsequence to increase the accuracy [12,13]. It integrates the advantages of different methods and the structure can be adjusted according to the practical engineering scenario.

Although the methods vary in forms, the performance of prediction methods is partly decided by the data used. From the perspective of input data, in addition to using single wind farm data directly, there are also some methods that use data from multiple wind farms. The relevant research and the actual observation of the wind farm in the field show that there is obvious correlation among wind farms [14,15]. Cavalcante [16] brought out Least Absolute Shrinkage and Selection Operator-Vector Autoregression (LASSO-VAR) which can take consideration of the historical data of all the wind farms in the region. However, it is still a linear regression model. Deep learning such as classic convolutional neural network (CNN) [17] and stacked denoising auto-encoder (SDAE) [18] are introduce for the prediction of multiple wind farms. It can effectively model the time-varying and nonlinear effect among all the closely related wind farms, but it does not consider the global geographical relation of wind farms in the region when dealing with the complex spatial and temporal features.

In most cases, we need to predict not only the power of each wind farm, but also the regional wind power. The additive method, extrapolation method and statistical scaling up method are commonly used [19]. The superposition method is to predict the power of all wind farms in the cluster, and simply sum the results. Extrapolation is a prediction method by comparing NWP with historical meteorological data to find similar scenarios through a historical database. The statistical scale-up method is to obtain the regional wind power output by multiplying the prediction results of the reference wind farm. In addition, there are some downscale methods for the spatial–temporal correlation analysis of wind
power and wind speed [20–22]. The downscale method is about getting higher resolution wind speed or wind power from lower resolution NWP or prediction results. The downscaled NWP windspeed can provide more precise information for wind power prediction [23].

In fact, there are difficulties on two levels to build a comprehensive wind farm prediction model. The first one is how to use the complex spatial–temporal relationship effectively among the historical wind power data and NWP data of different wind farms, in order to increase the accuracy of prediction. The second is how to get the output of every single wind farm and the whole region efficiently, especially when the number of wind farms is big.

Addressing these two goals, we proposed a hybrid prediction framework based on deep learning for wind power prediction in a region, calling it the Multi-modal Multi-task Graph Spatiotemporal NETwork (M2GSNet). The main contribution is as follows:

1. We designed a spatiotemporal graph convolutional network which can extract the spatiotemporal feature of historical wind power and NWP data of wind farms in the given region. To the best of our knowledge, we are the first to employ a spectral graph neural network for the ultra-short-term wind farm cluster power prediction. Compared to the previous wind power prediction method, it can take consideration of the global geographical location and make better use of the historical wind power and NWP information of wind farms in a region. It can reduce normalized root mean square error (RMSE) in the fourth hour by 1.75%.

2. We also designed multi-task learning for the wind power prediction of all the wind farms. This can enhance the learning efficiency by combining similar learning tasks and sharing weights of some neural network layers. The power of every single wind farm and the whole region efficiently can be predicted in one model. The time consumption of 20 wind farm forecasts is only 4.1 times the time used for one wind farm. There is also great potential to expand the method to a region which contain hundreds of wind farms.

The rest of this paper is organized as follows. Section 2 analyzes the availability of NWP and formulates the problem of wind power prediction on the graph. Section 3 provides temporal and spatial dependency modeling based on graph convolution. Here, the feature of historical wind power data and NWP data of different wind farms can be extracted. Based on these features, Section 4 proposes a multi-modal multi-task graph convolutional network for wind power prediction. The experimental results are reported in Section 5. The conclusion is made in Section 6.

2. Preliminaries

2.1. Availability Analysis of NWP

The wind farm can convert the wind energy into electrical power and the energy conversion process can be depicted as follows [24]:

\[ P_E = C_p P_{\text{wind}} = C_p \times \frac{1}{2} \rho A v^3 \] (1)

where \( C_p \) represents the wind power conversion parameter and \( \rho \) is the air density. \( A \) is the area through which the wind is flowing and \( v \) is the wind speed. Equation (1) shows that the wind power is proportional to the cubic of wind speed. Therefore, when comparing the effectiveness of numerical weather prediction, the cubic of wind speed is also chosen. Generally, the height of the wind turbines is 50 m to 150 m. The height of the NWP wind speed from the meteorology department is 10 m, 30 m, 100 m and 170 m. Therefore, the wind speed at the height of 100 m is selected for comparison and analysis since this altitude is more effective for the wind power conversion. To verify the availability of the NWP, two scenarios in 2019 are selected from a wind farm (no. 1 wind farm) in our testing wind farm cluster. The time intervals of the scenarios are from 2019-03-26 08:15:00 A.M. to 2019-03-31 08:15:00 A.M. and from 2019-04-05 08:15:00 A.M. to 2019-04-10 08:15:00 A.M., separately. Both the wind
speed and the cubic of wind speed are compared in those scenarios. As the wind power and wind speed are in different ranges, both of them are normalized to make it more illustrative in Figure 1.

![Wind speed and cubic comparison](image)

**Figure 1.** Comparison between numerical weather prediction (NWP) speed and wind power. (a) Wind speed comparison in scenario 1; (b) wind speed cubic comparison in scenario 1; (c) wind speed comparison in scenario 2; (d) wind speed cubic comparison in scenario 2.

To illustrate the similarity of NWP and wind power, we use the dynamic time wrapping (DTW) distance to assess the discrepancy of them. The classical sequence similarity computation is the Euclidean distance. However, when there is a time shift in the sequence, there will be large errors between two sequences even when they are similar. DTW is a kind of commonly used method to assess the similarity of two time series and is especially suitable when the length of two time series is different or there is a time shift between them [25]. The main idea of DTW is to convert the time series similarity computation into the shortest path problem and use the dynamic programming method for the computation. The computation process of DTW distance between $L_1$ and $L_2$ is listed as Equations (2) and (3).

$$dtw(i, j) = d(i, j) + \min\left\{ \begin{array}{ll}
dtw(i - 1, j) \\
dtw(i, j - 1) \\
dtw(i - 1, j - 1)
\end{array} \right.$$

(2)

$$\begin{array}{l}
dtw(1, 1) = d(1, 1) \\
dtw(i, 1) = d(i - 1, 1), i > 1 \\
dtw(1, j) = d(1, j - 1), j > 1
\end{array}$$

(3)

where $dtw \in R^{m \times n}$ is the distance of two sequences, $d(i, j)$ is the distance of the $i$-th element in the $L_1$ and $j$-th element in $L_2$ and the distance can be measured by the absolute value. $m$ and $n$ are the length of $L_1$ and $L_2$, respectively, and $dtw(m, n)$ is the DTW distance between $L_1$ and $L_2$.

According to the definition of DTW distance, in the first scenario, the distance between wind power and 100 m wind speed in the NWP is 140.8926 and the distance between wind power and 100 m wind speed cubic in the NWP is 80.2233. In the second scenario, the distance between wind power and 100 m wind speed in the NWP is 50.8905 and the distance between wind power and 100 m wind speed cubic in the NWP is 26.6638. From the computation results, we can see that due to the NWP amplitude
and shape error, there are also some times when the discrepancy between the NWP windspeed and wind power is relatively large, such as the time interval between 2019-03-28 and 2019-03-29. The DTW distance cannot be reduced into 0. However, the cubic of wind speed can represent the tendency of the wind power more precisely than the wind speed itself. Therefore, if we use the cubic NWP windspeed as the input of the model, rather than the NWP windspeed, it is beneficial for the prediction modeling.

2.2. Power Prediction of Wind Farms Cluster

Wind power forecast is a classic time-series prediction problem. It can give out the most likely output in the next H time steps given the previous M wind power observations and NWP data of length N. It can be depicted as:

\[
P_{t+1}, \ldots, P_{t+H} = f(P_{t-M+1}, \ldots, P_t, V_{t+1}, \ldots, V_{t+N})
\]

where \(P_t \in \mathbb{R}^{n \times 1}\) is the vector of the wind power of the wind farms in a region and \(V_t \in \mathbb{R}^{N \times k}\) is the matrix of NWP. \(n\) is the number of wind farms and \(k\) is the number of NWP variables. The prediction is to find the most appropriate \(f\) by using the machine learning or deep learning method. Since the ultra-short-term wind power prediction is supposed to output the wind power in the next 4 h at the interval of 15 min, the time step \(H\) is 16 generally.

However, there exists complex spatial–temporal correlation among the wind farms which is shown in Figure 2 and it is necessary to consider the coupling relationship when designing the function. However, even though there are some methods that can take into consideration the correlation [2,17], they neglect the geographical information of different wind farms.

![Figure 2](image-url)

**Figure 2.** Spatial temporal correlation of wind farms in a region. (a) Spatial temporal correlation of wind power; (b) spatial–temporal graph modeling.

In Figure 2, we also notice that the historical wind power and NWP data of the wind farms constitute a kind of typical spatiotemporal graph data structure. The prediction function can be described as:

\[
P_{t+1}, \ldots, P_{t+H} = f(P_{t-M+1}, \ldots, P_t, V_{t+1}, \ldots, V_{t+N}, G)
\]

where \(G\) is the graph constructed by the wind farms in the cluster and we design a multi-modal multi-task spatiotemporal graph convolutional network for the approximation of the prediction function.
3. Spatial-Temporal Dependency Modeling

3.1. The Adjacent Matrix

The adjacent matrix, which reflects the geographical dispersion of the wind farms, is very
important for the spatial–temporal dependency modeling and there are many methods to construct the
adjacent matrix. As the wind farms are located in a region and the spatial dispersion is mostly based
on distance, we define the adjacent matrix based on the Gaussian kernel threshold distance function of
the wind farms.

\[
A_{i,j} = \begin{cases} 
  e^{-\frac{\text{dist}(i,j)^2}{2\text{std}^2}}, & \text{if } \text{dist}(i,j) \leq \epsilon \\
  0, & \text{otherwise} 
\end{cases}
\]  

(6)

where dist(i, j) is the geographical distance between wind farm i and wind farm j. std is the standard
deviation of the distance between n wind farms and \(\varepsilon\) is the threshold. Here, we choose the half of the
mean distance as the threshold. If the distance is smaller than the threshold, we assess that there is no
connection between the two wind farms to guarantee the sparsity of the adjacent matrix.

3.2. The Graph Convolutional Neural Network

The convolution network witnessed a great success in the image recognition for its ability to
extract the spatial feature and many research studies use it for the spatial dependency modeling of
the wind farms. However, this kind of method needs to arrange the wind farms in a specific way
and ignore the geographic location relationship among the wind farms. To deal with this case, it is
necessary to expand the spatial dependency modeling method and develop the convolutional neural
network for the graph data. There has been some research [26] about using the graph neural network
for the short-term windspeed prediction and its effectiveness has been verified. We can also use the
graph convolutional neural network for the spatial feature extraction in the ultra-short-term wind farm
cluster power prediction.

There are two methods to develop the graph neural network, namely, the spatial domain method
and spectral domain method. However, the spectral method is based on graph Fourier transformation
and has a relatively solid theoretical foundation that is more suitable for the wind power prediction.

We use the first order approximation of Chebyshev spectral filter brought out by Kipf in the graph
convolution layer [27,28] and the propagation mode between layers are as follows:

\[
X^{l+1} = \sigma \left( \hat{D}^{-\frac{1}{2}} \hat{A} \hat{D}^{-\frac{1}{2}} X^{(l)} W^{(l)} \right) \quad (7)
\]

\[
\hat{A} = A + I_n \quad (8)
\]

In this equation, \(A \in R^{n \times n}\) is the adjacency matrix formed from the location of wind farms and \(n\)
is the number of wind farms, \(\hat{D} \in R^{n \times n}\) is the degree matrix of \(\hat{A}\). \(X^{(l)} \in R^{n \times d}\) is the feature of layer
\(l\) and \(X^{(l+1)} \in R^{n \times h}\) is the updated feature of layer \(l + 1\). \(d\) and \(h\) constitute the feature dimension of
each node, which are the time series data of certain wind farms in this case. \(W \in R^{d \times h}\) is the learnable
convolutional kernel parameter. \(\sigma\) is the activation function. Through the matrix product, the feature
of each wind farm correlated with each other.

For the input layer, we use the historical wind power of length \(M\) or the NWP series of length \(N\)
as the feature of each wind farm. According to Equation (7), the first layer of the graph convolutional
network outputs a matrix with the following elements:

\[
X_p^{(l)}(i,j) = \sigma \left( \sum_{s=1}^{M} \sum_{k=1}^{n} \frac{1}{\hat{D}(i,i) \hat{D}(k,k)} \hat{A}(i,k) P_{l-1+1}(k) W(s,j) \right) \quad (9)
\]
In the equation above, $X^{(l)}_nwp(i, j)$ and $X^{(l)}_{p}(i, j)$ are the spatiotemporal feature of the historical wind power and NWP, where $i = 1, 2, \ldots, n$ and $i = 1, 2, \ldots, d$. It can be noticed that the spatial correlation is adjusted by the $\hat{A}$ and the temporal correlation is mapped by the $W$.

In this way, the temporal feature of wind farms is utilized in the network. Through this design, we can get the graph neural network that is suitable for our wind power prediction and we call it Graph Convolution Module. The structure of the graph convolutional network can be represented as Figure 3.

Graph convolution can extend the convolution from the traditional Euclidean distance space to the general graph data by carrying out the convolution in the spectral domain. In practice, we should decide the depth of the GCN and the adjacent matrix. In our wind power prediction problem, we use two layers of GCN for the spatial dependency modeling considering too many layers will lead to the feature embedding indistinguishably, although it can increase the size of receptive filed [27].

4. Multi-Modal Multi-Task Graph Spatiotemporal Network

4.1. Multi-Modal Learning

Historical wind power and NWP contain different types of information about the wind power to be predicted and we need to fuse the spatiotemporal feature of them in the network. Multi-modal learning is a kind of technique which can process and combine information from different sources [29]. Feature fusion aims to integrate information of different types and sources to get a consistent and common model output, which is a basic problem in the multi-modal learning field. There are three commonly used methods for the feature fusion.
4.1.1. Bilinear Fusion Method

The calculation formula is

\[ Y = X_p \ast W_m \ast X_{nwp} + b \] (11)

where \( X_p \in \mathbb{R}^{n \times d_p} \) and \( X_{nwp} \in \mathbb{R}^{n \times d_{nwp}} \) are the spatiotemporal feature of historical wind power and NWP. Tensor \( W_m \in \mathbb{R}^{d_p \times d_{nwp} \times d_\text{out}} \) is the parameter of bilinear transformation and \( b \) is the bias of bilinear fusion. Suppose \( X_p \) is a vector with dimension (128, 20) and \( X_{nwp} \) is a vector with dimension (128, 30). When the dimension of \( W_m \) is (20, 30, 40), the dimension of the fused feature is (128, 40).

4.1.2. Nonlinear Weighted Fusion

The nonlinear weighted fusion is as follows:

\[ Y = \tanh(X_p W_p + X_{nwp} U_{nwp}) \] (12)

The \( W_p \in \mathbb{R}^{d_p \times d_\text{out}} \) and \( U_{nwp} \in \mathbb{R}^{d_{nwp} \times d_\text{out}} \) are the weighted parameter of the spatiotemporal feature in the fusion.

4.1.3. Concatenate

When the correlation of features is weak, the method of direct feature concatenate can be used for feature fusion as follows.

\[ Y = [X_p, X_{nwp}] \] (13)

In the wind power prediction problem, the fused feature \( Y \in \mathbb{R}^{n \times d_\text{out}} \) is the input for the task-specific layer of each wind farm which consists of the multi-task learning.

4.2. Multi-Task Learning

Multi-task learning (MTL) has led to great success in many areas of deep learning, from natural language processing to speech recognition [30]. Traditionally, the wind power prediction model is designed for every single wind farm separately or predicts the wind power of a region directly. We then fine-tune and tweak these models to improve the performance to an acceptable level. However, when optimizing more than one loss function and the tasks are similar to each other, there is a chance that the auxiliary task will help improve the accuracy of the main task. The multi-task learning can improve the generalization ability by leveraging the domain-specific information in the training signals of related tasks. It can make better use of the entangled features when using the multi-task learning [31]. There is also research about using the multi-modal and multi-task learning for short-term wind power prediction [32]. However, multi-task learning is used for different steps rather than different wind farms. We use one of the most commonly used methods for the multi-task learning called hard parameter sharing. It shares some hidden layers and has several task-specific output layers, as can be seen in Figure 4.
In our network, we can share the parameter of the graph neural network and multi-modal learning in the previous part and use the multi-task learning for the wind power prediction of each wind farm. The task-specific layer is designed for each wind farm and the loss function of the neural network is the sum of loss in each wind farm.

\[
\hat{p}_i = Y_i W_{oi} \\
L = \sum_{i=1}^{n} ||\hat{p}_i - p_i||_1
\]

where \(\hat{p}_i \in R^{1 \times H}, i = 1, 2, \ldots, n\) is the predicted wind power of the \(i\)-th wind farm. \(Y_i \in R^{1 \times d_{out}}\) is the \(i\)-th row of multi-modal feature fusion matrix and \(W_{oi}\) is the parameter of the fully connected layers for the power prediction. The number of task-specific layers is equal to the number of wind farms. \(L\) is the loss function of the M2GSNet and it is the sum of loss function for each wind farm. We can use the absolute value of prediction power and true power to calculate. The parameters of the fully connected layers are learned in the training process. It actually plays a role in weighting the fused features. Therefore, the spatial–temporal feature is taken into consideration in the procedure above.

4.3. GCN Model for Wind Power Prediction

According to the analysis before, the NWP contains the future meteorological information and it can be found that there is a good correlation with the power. Therefore, the multi-modal learning is used to combine the temporal and spatial characteristics of the historical wind power and NWP. The multi-task learning is used for the prediction of each wind farm. The proposed multi-modal multi-task graph spatiotemporal network (M2GSNet) model for wind power prediction is as shown in Figure 5:
connected layers are learned in the training process. It actually plays a role in weighting the fused features. Therefore, the spatial–temporal feature is taken into consideration in the procedure above.

4.3. GCN Model for Wind Power Prediction

According to the analysis before, the NWP contains the future meteorological information and it can be found that there is a good correlation with the power. Therefore, the multi-modal learning is used to combine the temporal and spatial characteristics of the historical wind power and NWP. The multi-task learning is used for the prediction of each wind farm. The proposed multi-modal multi-task graph spatiotemporal network (M2GSNet) model for wind power prediction is as shown in Figure 5:

Figure 5. Multi-modal multi-task graph spatiotemporal network (M2GSNet) model for wind farm cluster power prediction.

From Figure 5, we can see that the whole structure is an encoder–decoder framework. The historical wind power and NWP data are encoded into a latent spatial and temporal feature space by using graph convolutional network and multi-modal learning. The features are decoded into the wind power of different wind farms by using multi-task learning and fully connected layers. The model consists of the following modules.

(1) GCN Module: GCN part includes two GCN modules for the historical wind power and NWP data, respectively. Each GCN module includes two layers of standard GCN and the adjacent matrix is based on distance. It is used to extract the spatial feature of the wind farm.

(2) Multi-Modal Learning Module: It is used to concatenate the spatial and temporal feature of NWP and historical wind power. It makes M2GSNet an effective hybrid prediction method by combining the advantage of model-based prediction and data-driven prediction.

(3) Multi-Task Learning Module: The fully connected layer is used to map the fused spatial–temporal feature of each wind farm into the wind power. Each wind farm has a specific layer. The historical wind power and NWP data are the input of the model. The output is the ultra-short-term power sequence of each wind farm. The regional wind power can be calculated by adding together the power of each wind farm.

5. Case Study

The proposed method is also tested on the measurement data of a wind farm cluster in Northeast China. The proposed model is tested on Linux server Cluster (CPU: Intel Xeon (R) CPU E5-2650 v4 @ 2.10 GHz, GPU: NVIDIA Tesla P100) and deep learning framework Pytorch (1.4.0) with GPU acceleration to speed up the training process.

5.1. Data Set and Test Description

For the test system, only the historical wind power data and the NWP data are provided. The wind speed data are not included. The historical wind power is from the field measurement and the NWP is from the meteorology station. The NWP wind speed at the height of 170, 100, 30 and 10 m are used for the analysis. According to the analysis in 2.1, the cubic NWP windspeed can reflect the tendency of
wind power more effectively. So, we use the cubic NWP windspeed rather than the NWP windspeed as the input of the model. The location of those wind farms in the cluster is as Figure 6.

![Figure 6](image)

**Figure 6.** The location of wind farms in the northeast part of China.

The red point in the figure is the wind farm and the number in the array is the wind farm number and the capacity, respectively. The whole capacity of the wind farm cluster is 2854.31 MW. The wind farms with strong correlation are linked together according to the adjacent matrix computed by Equation (6). Data from 2019 are used for the model training and testing. The training set include 8000 samples (from 2019-01-01 08:15:00 A.M. to 2019-03-25 04:00:00 P.M.) and the testing set include 2000 samples (from 2019-03-25 04:15:00 P.M. to 2019-04-15 12:00:00 P.M.). The training samples are randomly scrambled to avoid overfitting of the model. The measurement and prediction interval of the data is 15 min. However, the wind power and cubic wind speed have different units, the normalization is used as follows.

\[
x = \frac{x_m - x_{\min}}{x_{\max} - x_{\min}}
\]

The root mean square error (RMSE) and mean absolute error (MAE) are selected as the evaluation metric to assess the performance of the model on the testing set:

\[
RMSE = \sqrt{\frac{1}{K} \sum_{i=1}^{K} (x_{ti} - \hat{x}_{ti})^2}
\]

\[
MAE = \frac{1}{K} \sum_{i=1}^{K} |x_{ti} - \hat{x}_{ti}|
\]
where \( x_{ti} \) and \( \hat{x}_{ti} \) are the normalized true value and normalized predicted value in prediction scenario \( i \) at prediction time step \( t \). \( K \) is the number in the test set. To represent the prediction error of scenarios, we design another index:

\[
\text{Ind}_i = \frac{1}{16} \sum_{t=1}^{16} |x_{ti} - \hat{x}_{ti}|
\]  

(19)

In each scenario, there are 16 time steps and the maximum value of the time step is 16. The index is used to assess the similarity of a given scenario and it reflects the average deviation of true value and predicted value in each time step.

The sensitivity of some hyper-parameters is taken into account, such as learning rate and the hidden state number, which are very important for the training process [33]. However, it is impossible to do the grid search on the whole parameter space. So, the hyper-parameter is determined according to the grid search combined by human experience. The learning rate is chosen from the set \((0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4)\). The hidden layer numbers of the graph convolutional part for historical wind power and wind speed are both chosen from the set \((10, 20, 30, 40, 50, 60, 70, 80, 90, 100)\). The input lengths of the historical wind power and cubic NWP windspeed are chosen from set \((20, 30, 40, 50, 60)\). We decided the optimal value according to the prediction error in the fourth hour. The parameter combination with the lowest prediction error is the optimal value. After adjusting the structure and parameters of the model, the parameters of the final model are as follows. In the M2GSNet model, input characteristic matrix \( P_{t-M+1:t} \) is the power measurement information of each node on the graph. It includes 40 time steps \((M = 40)\), which require 10 h of power data of 20 wind farms to predict the power of the next 16 time steps \((4 \text{ h})\). Input characteristic matrix \( P_{t+1:t+N} \) selects the NWP data of each node. It includes 20 time steps \((N = 20)\), which require 5 h wind speed of 20 wind farms in the future to predict the output. NWP variables include windspeed from four different altitudes which are 10, 30, 100 and 170 m. Data of 20 wind farms in Jilin Province are used for training and prediction. So, \( X_p^{(0)} \) is a \((20 \times 40)\) matrix and \( X_{nwp}^{(0)} \) is a \((20 \times 80)\) matrix. The hidden state for the wind power GCN module is 60 and for the NWP GCN module is 40. The adjacent matrix in the graph convolution network is calculated by the distance between wind farms. The dimension of variable is labeled in Figure 5. The prediction error is calculated according to the RMSE after normalization and the specific calculation method can refer to the previous description.

The iteration epoch of model training is 200 and the training batch size is 256. The optimizer is Adadelta [34] and the learning rate is 0.1. The five-fold cross-validation is used for verification.
5.2. Baseline Model

M2GSNet is our proposed model and it has three features. First, it utilizes the feature of cubic NWP wind speed by using the multi-modal learning. Second, it adopts the spatiotemporal model for the geographical information extraction. Third, it uses the multi-task learning to predict the power of each wind farm. To illustrate the accuracy improvement by each feature, we design the baseline model and other GCN model for ablation study.

(1) MLP [6]: This is the multilayer perception model for regression and the hidden state number is 800. The historical wind power of each wind farm, including 40 time steps as the input and output, is the wind power of the same wind farm including 16 time steps. The sum of the wind power for each wind farm is the regional wind power.

(2) LSTM [9]: This includes two LSTM layers and the dropout rate is 0.25. The historical wind power of each wind farm, including 40 time steps as the input and output, is the wind power of the same wind farm including 16 time steps.

(3) ELM [35]: This uses the classic ELM model parameter. The historical wind power of each wind farm, including 40 time steps as the input and output, is the wind power of the same wind farm including 16 time steps. The sum of the wind power for each wind farm is the regional wind power.

(4) LSTNet [10]: This uses the standard LSTNet structure and parameter and it can take consideration of the spatiotemporal relationship of wind farms. However, it cannot make use of the geographical information of the wind farms when extracting the spatial feature of the wind farms. The input only includes the historical wind power of each wind farm. The output is the wind power of each wind farm including 16 time steps.

(5) LSTNet_NWP: This uses the same structure and parameter with LSTNet. However, it also uses the cubic NWP wind speed as input and it concatenates the spatial–temporal feature of historical wind power and NWP. The output is the wind power of each wind farm including 16 time steps.

We also compare different M2GSNet models for the ablation study and the characteristic of each model is as shown in Table 1.

| Model Type | Method      | Multi-Modal | Multi-Task | Adjacent Matrix | Prediction Granularity       |
|------------|-------------|-------------|------------|-----------------|------------------------------|
| M2GSNet    | Cubic wind speed | Yes | Distance based | Power of each wind farm |
| w/o CW     | Raw wind speed   | Yes | Distance based | Power of each wind farm |
| w/o AD     | Cubic wind speed | Yes | Correlation based | Power of each wind farm |
| w/o W      | No wind speed    | Yes | Distance based | Power of each wind farm |
| w/o MT1    | Cubic wind speed | No | Distance based | Regional power |
| w/o MT2    | Cubic wind speed | No | Distance based | Power of each wind farm |

Where M2GSNet means the model that uses the information of the cubic NWP wind speed, the w/o CW means the GCN model that only uses the raw data of NWP but without using the cubic NWP wind speed. The w/o AD means the GCN model that uses the information of cubic NWP wind speed but uses the wind speed time series correlation to define the graph. The w/o W means the GCN model that does not use the NWP. The w/o MT1 means the GCN model that does not use the multi-task learning and predicts the wind power of the region directly. The w/o MT2 means the GCN model that does not use the multi-task learning and predict the wind power of each wind farm separately. We use same model structure but train the model individually. The training hyper-parameter is the same as the description above and only the model structure is different.
5.3. The Main Prediction Results

5.3.1. The Prediction Results for Regional Wind Power

The prediction results of several structures of the M2GSNet are listed in Table 2. From Table 2, it is obvious that M2GSNet is the model that performs best. Besides, methods which take consideration of the NWP are better than those that do not include NWP data. From the results, the prediction error of the M2GSNet method is smaller than LSTM by over 2 percent in the fourth hour. This means it can reduce more than 50 MW prediction error for the whole cluster, which is vital progress for the operation center of the power grid.

| Method       | 1 h  | 2 h  | 3 h  | 4 h  |
|--------------|------|------|------|------|
|              | RMSE | MAE  | RMSE | MAE  |
| MLP          | 4.66 | 3.29 | 8.03 | 5.92 |
| LSTM         | 5.02 | 3.61 | 7.93 | 5.94 |
| ELM          | 4.52 | 3.29 | 7.74 | 5.79 |
| LSTNet       | 5.35 | 3.76 | 8.22 | 6.02 |
| LSTNet_NWP   | 4.85 | 3.51 | 7.92 | 5.95 |
| M2GSNet      | 4.40 | 3.22 | 7.11 | 5.47 |
| w/o CW       | 4.69 | 3.52 | 7.43 | 5.80 |
| w/o AD       | 4.84 | 3.66 | 7.68 | 6.05 |
| w/o W        | 4.77 | 3.60 | 7.59 | 5.94 |
| w/o MT1      | 4.73 | 3.49 | 7.46 | 5.78 |
| w/o MT2      | 5.01 | 3.72 | 7.42 | 5.72 |

LSTNet is a kind of deep learning method that takes consideration of the spatiotemporal relationship of wind farms in the cluster. It is an improved version of spatiotemporal prediction model [17]. From Table 2, we can see that LSTNet is indeed better than the MLP, LSTM and ELM which do not consider the spatiotemporal relationship. However, the M2GSNet method is better than the LSTNet due to its ability to extract the geographical location information feature.

Even for the M2GSNet, when using multi-task learning to predict the wind power of each wind farm, the results are better than predicting the regional wind power directly (w/o MT1) or predicting the wind power of each and summing them together (w/o MT2) which proves the effectiveness of multi-task learning. Besides, the result of using the cubic NWP windspeed in the multi-modal learning is better than the result of using NWP windspeed directly.

5.3.2. The Prediction Results of Each Wind Farm

The M2GSNet is not only convenient for predicting the regional wind power, but it also can output the detail power of each wind farm by one training session. To verify the effectiveness of M2GSNet on the single wind farm power prediction, the RMSEs of every wind farm on the 16 time steps are calculated. The RMSE data of 20 wind farms in 1 h are used for the boxplot analysis. We displayed the results of MLP, LSTM, ELM, LSTNet and M2GSNet in Figure 7.
Figure 7. Boxplot of prediction results for wind farms.

From Figure 7, we can notice that the prediction errors of single wind farms are much higher than some other wind farms. However, due to the “smooth effect” of the wind farm cluster, the prediction error of the cluster is much smaller than the individual wind farms. This result is very meaningful for the power grid dispatching center. In addition, according to the mean value, max value and minimum value, the performance of the M2GSNet is much better than the other methods, especially in the 3rd hour and 4th hour in the statistical sense. However, due to the NWP feature fusion, the prediction error of M2GSNet in the 1st hour is a little higher than the other methods that do not consider the NWP. It also enlightens us to design a mechanism to dynamically select the models. For example, for the ultra-short-term prediction within 1 h, we can choose a model with lower RMSE.

5.3.3. Ablation Study

(1) The Comparison of Different Concatenate Method

The feature fusion of historical wind power and NWP is very important for the wind power prediction and there are three commonly used feature fusion methods. The prediction results of the three methods are listed in Figure 8.

From Figure 8, it is obvious that the feature fusion method of bilinear and concatenate is better than the nonlinear Tanh method. The prediction error of the concatenate method is slightly lower than the bilinear method, especially in the interval of 0.5 h–3.5 h. Considering that the bilinear method is more complex and has lower training efficiency, the concatenate method is chosen as the feature fusion method in our network.
5.3.3. Ablation Study

(1) The Comparison of Different Concatenate Method

The feature fusion of historical wind power and NWP is very important for the wind power prediction and there are three commonly used feature fusion methods. The prediction results of the three methods are listed in Figure 8.

From Figure 8, it is obvious that the feature fusion method of bilinear and concatenate is better than the nonlinear Tanh method. The prediction error of the concatenate method is slightly lower than the bilinear method, especially in the interval of 0.5 h – 3.5 h. Considering that the bilinear method is more complex and has lower training efficiency, the concatenate method is chosen as the feature fusion method in our network.

(2) The Comparison of Training Time Consumption under Different Wind Farm Numbers

The training time of the M2GSNet is crucial because it determines whether it can be utilized in the large-scale renewable energy cluster which includes hundreds, even thousands, of small wind farms. Therefore, we compare the training time of M2GSNet under different wind farm numbers. The results are in Figure 9.

According to the results, the training time for one wind farm is 66 min. So, if each wind farm uses one specific model, the training time is more than 1200 min in this case. However, when multi-task learning is used, the training time reduces to 271 min. Thus, it can be seen that by using multi-task learning, it saves a lot of training time and resources. Notably, when the wind farm number increases, the increase rate of training time is actually decreased. Therefore, when more wind farms are considered, the advantages of multi-task learning will be more remarkable.
5.3.4. The Remarkable Error Analysis in Test Set

The prediction error of the M2GSNet is analyzed as Figure 10. In this figure, the M2GSNet is compared with the MLP and LSTM since they are the most commonly used machine learning and deep learning methods. The prediction results are visualized.

![Figure 10. The prediction error analysis.](image)

In the left array of the figure, it is the 4th-hour prediction results in the test set by three methods. The predicted values are compared with the true value. In the right array, the location of scenarios with larger prediction errors are visualized. The prediction error analysis is also very important because it can tell us which kind of scenario is difficult to be predicted. Then we can design methods to deal with it in the future. For each time step, we used different colors and different sizes to represent the prediction errors. The darker the color and the smaller the size, the smaller the prediction error. Since the of most of scenarios are smaller than 0.15, we classify the prediction errors into four categories. If the prediction error according to in Equation (19) is smaller than 0.05 p.u, it is the first category. This type includes the scenarios that are predicted rather accurately. If the the prediction error is between 0.05 and 0.1, it is the second category, and the color of this category in the figure is at 5. If the prediction error is between 0.1 and 0.15, it is the third category, and the color of those scenarios is at 10. If the prediction error is higher than 0.15, it is the fourth category and the color of them is at 20. So, different colors and sizes can reflect the prediction results of the same scenario. We counted the ratio of the different categories in Table 3.

| Method | Sample Type Divided by $\text{Ind}_i$ |
|--------|-------------------------------------|
|        | [0,0.05] | (0.05,0.10] | (0.10,0.15] | >0.15 |
| M2SGNet | 52.1 | 34.4 | 11.85 | 1.65 |
| LSTM | 50 | 33 | 11.55 | 5.45 |
| MLP | 53.05 | 31.15 | 10.55 | 5.25 |
From Figure 10 and Table 3, it is obvious that the M2GSNet method has better performance since it has less points belonging to the high prediction error category. However, it also can be found from Figure 10 that most light color and large size points are located in the turning point of the wind fluctuation process which means it is hard to predict and often leads to higher prediction errors.

6. Conclusions

In this paper, we bring out a spatiotemporal deep learning network for the ultra-short-term wind power prediction. Through the case study, we can draw the following conclusions:

1. Adding a numerical weather forecast by virtue of multi-modal learning, especially the third power of wind speed as auxiliary information, can improve the accuracy of forecasts.

2. The spatiotemporal graph neural network can extract the spatial–temporal feature of the wind farms effectively and is helpful in improving the accuracy of predictions compared to the other methods.

3. By using the multi-task learning method, prediction accuracy can be improved, and the training time can also be reduced compared to additive methods.

In the follow-up study, we can consider designing a comprehensive method which can classify the wind process in advance and define the dynamic graph according to the spatial–temporal relationship among wind farms to further increase the accuracy.

Author Contributions: Conceptualization, H.F. and X.Z.; methodology, H.F. and X.Z.; software, H.F., K.C. and X.C.; data curation, H.F.; writing—original draft preparation, H.F., K.C. and X.C.; writing—review and editing, X.Z. and S.M.; visualization, H.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by National Key R&D Program of China (technology and application of wind power/photovoltaic power prediction for promoting renewable energy consumption, 2018YFB0904200) and complementary S&T Program (ultra-short-term prediction of wind speed considering spatiotemporal correlation model of wind process, DUKZZZ-YBHT-2019-JSC0405-0089) of Inner Mongolia Power (Group) Co, Ltd.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xue, Y.S.; Yu, C.; Zhao, J.H.; Li, K.; Liu, X.Q.; Wu, Q.W.; Yang, G.Y. A review on short-term and ultra-short-term wind power prediction. *Autom. Electr. Power Syst.* **2015**, *36*, 141–151.
2. Dowell, J.; Pinson, P. Very-short-term probabilistic wind power forecasts by sparse vector autoregression. *IEEE Trans. Smart Grid* **2015**, *7*, 763–770. [CrossRef]
3. Messner, J.W.; Pinson, P. Online adaptive lasso estimation in vector autoregressive models for high dimensional wind power forecasting. *Int. J. Forecast.* **2019**, *35*, 1485–1498. [CrossRef]
4. Wytock, M.; Zico Kolter, Z. Large-scale probabilistic forecasting in energy systems using sparse gaussian conditional random fields. In Proceedings of the 52nd IEEE Conference on Decision and Control, Florence, Italy, 10–13 December 2013.
5. Wytock, M.; Zico Kolter, Z. Sparse gaussian conditional random fields: Algorithms, theory, and application to energy forecasting. In Proceedings of the International conference on machine learning, Atlanta, GA, USA, 17–19 June 2013.
6. Li, J.X.; Mao, J.D. Ultra-short-term wind power prediction using bp neural network. In Proceedings of the 2014 9th IEEE Conference on Industrial Electronics and Applications, Hangzhou, China, 9–11 June 2014.
7. Demolli, H.; Dokuz, A.S.; Ecemis, A.; Gokcek, M. Wind power forecasting based on daily wind speed data using machine learning algorithms. *Energy Convers. Manag.* **2019**, *198*, 1–12. [CrossRef]
8. Tan, L.; Han, J.; Zhang, H.T. Ultra-short-term wind power prediction by salp swarm algorithm-based optimizing extreme learning machine. *IEEE Access* **2020**, *8*, 44470–44484. [CrossRef]
9. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
10. Lai, G.K.; Chang, W.C.; Yang, Y.M.; Liu, H.X. Modeling long-and short-term temporal patterns with deep neural networks. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018.
11. Ju, Y.; Sun, G.Y.; Chen, Q.H.; Zhang, M.; Zhu, H.X.; Rehman, M.U. A model combining convolutional neural network and lightgbm algorithm for ultra-short-term wind power forecasting. *IEEE Access* 2019, 7, 28309–28318. [CrossRef]

12. Naik, J.; Dash, S.; Dash, P.K.; Bisoi, R. Short term wind power forecasting using hybrid variational mode decomposition and multi-kernel regularized pseudo inverse neural network. *Renew. Energy* 2018, 118, 180–212. [CrossRef]

13. Liu, H.; Mi, X.W.; Li, Y.F. An experimental investigation of three new hybrid wind speed forecasting models using multi-decomposing strategy and elm algorithm. *Renew. Energy* 2018, 118, 180–212. [CrossRef]

14. Xue, Y.S.; Chen, N.; Wang, S.M.; Wen, F.S.; Lin, Z.Z.; Wang, Z. Review on wind speed prediction based on spatial corelation. *Autom. Electr. Power Syst.* 2017, 41, 161–169.

15. Ye, L.; Zhao, Y.N. A review on wind power prediction based on spatial corelation approach. *Autom. Electr. Power Syst.* 2014, 38, 126–135.

16. Cavalcante, L.; Bessa, R.J.; Reis, M.; Browell, J. Lasso vector autoregression structures for very short-term wind power forecasting. *Wind Energy* 2017, 20, 657–675. [CrossRef]

17. Zhu, Q.M.; Chen, J.F.; Shi, D.Y.; Zhu, L.; Bai, X.; Duan, X.Z.; Liu, Y.L. Learning temporal and spatial correlations jointly: A unified framework for wind speed prediction. *IEEE Trans. Sustain. Energy* 2019, 11, 509–523. [CrossRef]

18. Yan, J.; Zhang, H.; Liu, Y.Q.; Han, S.; Li, L.; Lu, Z.Z. Forecasting the high penetration of wind power on multiple scales using multi-to-multi mapping. *IEEE Trans. Power Syst.* 2018, 33, 3276–3284. [CrossRef]

19. Peng, X.S.; Xiong, L.; Wen, J.Y.; Cheng, S.J.; Deng, D.Y.; Feng, S.L.; Wang, B. A summary of the state of the art for short-term and ultra-short-term wind power prediction of regions. *Proc. CSEE* 2016, 36, 6315–6326.

20. Liu, Y.C.; Chen, D.; Li, S.W.; Chan, P.W. Discerning the spatial variations in oﬀshore wind resources along the coast of China via dynamic downscaling. *Energy* 2018, 160, 582–596. [CrossRef]

21. González-Aparicio, I.; Monforti, F.; Volker, P.; Zucker, A.; Careri, F.; Huld, T.; Badger, J. Simulating European wind power generation applying statistical downscaling to reanalysis data. *Appl. Energy* 2017, 199, 155–168. [CrossRef]

22. Gaitan, C.F.; Cannon, A.J. Validation of historical and future statistically downscaled pseudo-observed surface wind speeds in terms of annual climate indices and daily variability. *Renew. Energy* 2013, 51, 489–496. [CrossRef]

23. Li, L.; Liu, Y.Q.; Yang, Y.P.; Han, S. Short-term Wind Speed Forecasting Based on CFD Pre-calculated Flow Fields. *Proc. CSEE* 2013, 33, 27–32.

24. Manyonge, A.W.; Ochieng, R.M.; Onyango, F.N. Mathematical modelling of wind turbine in a wind energy conversion system: Power coefficient analysis. *Appl. Math. Sci.* 2012, 6, 4527–4536.

25. Qin, L.; Xiong, Y.D.; Liu, K.P. Weather division-based wind power forecasting model with feature selection. *IET Renew. Power Gener.* 2019, 13, 3050–3060. [CrossRef]

26. Khodayar, M.; Wang, J.H. Spatio-temporal graph deep neural network for short-term wind speed forecasting. *IEEE Trans. Sustain. Energy* 2018, 10, 670–681. [CrossRef]

27. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* 2016, arXiv:1609.02907.

28. Li, Y.G.; Yu, R.; Shahabi, C.; Liu, Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv* 2017, arXiv:1707.01926.

29. Zhang, C.; Yang, Z.; He, X.; Deng, L. Multimodal Intelligence: Representation Learning, Information Fusion, and Applications. *IEEE J. Sel. Top. Signal Process.* 2020, 14, 478–493. [CrossRef]

30. Caruana, R. Multitask Learning. *Machine Learning.* 1997, 28, 41–75. [CrossRef]

31. Bengio, Y.S.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 2013, 35, 1798–1828. [CrossRef]

32. Chen, J.F.; Zhu, Q.M.; Li, H.Y.; Zhu, L.; Shi, D.Y.; Li, Y.H.; Duan, X.Z.; Liu, Y.L. Learning heterogeneous features jointly: A deep end-to-end framework for multi-step short-term wind power prediction. *IEEE Trans. Sustain. Energy* 2019, 11, 1761–1772. [CrossRef]

33. Bengio, Y. *Practical Recommendations for Gradient-Based Training of Deep Architectures*. Neural Networks: Tricks of the Trade; Springer: Berlin/Heidelberg, Germany, 2012; pp. 437–478.
34. Zeiler, M.D. Adadelta: An adaptive learning rate method. *arXiv* 2012, arXiv:1212.5701.
35. Zhao, Y.N.; Ye, L.; Li, Z.; Song, X.R.; Lang, Y.S.; Su, J. A novel bidirectional mechanism based on time series model for wind power forecasting. *Appl. Energy* 2016, 177, 793–803. [CrossRef]

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).