Intelligent and Converged Networks ISSN 2708-6240
2022, 3(4): 364−377 DOI: 10.23919/ICN.2022.0024

Combining random forest and graph wavenet for spatial-temporal data prediction

Chong Chen*, Yanbo Xu, Jixuan Zhao, Lulu Chen, and Yaru Xue

Abstract: The prosperity of deep learning has revolutionized many machine learning tasks (such as image recognition, natural language processing, etc.). With the widespread use of autonomous sensor networks, the Internet of Things, and crowd sourcing to monitor real-world processes, the volume, diversity, and veracity of spatial-temporal data are expanding rapidly. However, traditional methods have their limitation in coping with spatial-temporal dependencies, which either incorporate too much data from weakly connected locations or ignore the relationships between those interrelated but geographically separated regions. In this paper, a novel deep learning model (termed RF-GWN) is proposed by combining Random Forest (RF) and Graph WaveNet (GWN). In RF-GWN, a new adaptive weight matrix is formulated by combining Variable Importance Measure (VIM) of RF with the long time series feature extraction ability of GWN in order to capture potential spatial dependencies and extract long-term dependencies from the input data. Furthermore, two experiments are conducted on two real-world datasets with the purpose of predicting traffic flow and groundwater level. Baseline models are implemented by Diffusion Convolutional Recurrent Neural Network (DCRNN), Spatial-Temporal GCN (ST-GCN), and GWN to verify the effectiveness of the RF-GWN. The Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) are selected as performance criteria. The results show that the proposed model can better capture the spatial-temporal relationships, the prediction performance on the METR-LA dataset is slightly improved, and the index of the prediction task on the PEMS-BAY dataset is significantly improved. These improvements are extended to the groundwater dataset, which can effectively improve the prediction accuracy. Thus, the applicability and effectiveness of the proposed model RF-GWN in both traffic flow and groundwater level prediction are demonstrated.

Key words: spatial-temporal data; random forest; graph wavenet; groundwater level prediction

1 Introduction

With the rapid development and application of technologies such as the Internet and high-precision sensors, spatial-temporal data are becoming more massive. Compared to time series data, spatial-temporal data are more complex in the (time-dependent) spatial correlations[1, 2]. Because of the complexity of spatial-temporal data and the rapid expansion of data volume, the defects of traditional data mining methods are becoming increasingly apparent[3]. Traditional methods based on statistical principles[4, 5] have difficulties in capturing spatial correlation in spatial-temporal series, such as Autoregressive Integrated Moving Average (ARIMA)[6], which only considers the time-dimensional characteristics of spatial-temporal series and generates large errors in prediction results. Traditional machine learning methods (e.g., Support Vector Machine (SVM)[7, 8] and Hidden Markov Model

• Chong Chen, Yanbo Xu, Jixuan Zhao, and Yaru Xue are with the College of Information Science and Engineering, China University of Petroleum-Beijing, Beijing 102249, China. E-mail: chenchong@cup.edu.cn; 2020215913@student.cup.edu.cn; zjxuantj@163.com; xueyaru@cup.edu.cn.
• Lulu Chen is with the Education Management Information Centre of the Ministry of Education, Beijing 100816, China. E-mail: chenlulu@moe.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2022-09-06; revised: 2022-11-23; accepted: 2022-12-02

© All articles included in the journal are copyrighted to the ITU and TUP. This work is available under the CC BY-NC-ND 3.0 IGO license: https://creativecommons.org/licenses/by-nc-nd/3.0/igo/.

(HMM)[9]) can capture nonlinear relationships in spatial-temporal data to a limited extent, but in the long term, they are difficult to predict and adapt to large-scale spatial-temporal datasets. The recent proposal of Graph Neural Network (GNN) and its growing popularity in the past few years make it possible for deep learning algorithms to be used in graph structured data, and spatial-temporal graph modeling has also received great attention.

In recent years, the most widely used model GNN is Graph Convolutional Network (GCN) and its variants[10]. For example, Graph SAmpling and aggreGatE (GraphSAGE) algorithm[11] samples the nodes, aggregates the neighbors of the nodes, and learns the nodes based on the aggregated information. Graph Attention Network (GAT)[12] introduces the attention mechanism into GCN and adds aggregation operations on neighbor nodes to achieve adaptive assignment of weights to different neighbors. There are two most common approaches to capture the spatial-temporal correlation of spatial-temporal data using GNN. One approach is to combine GCN with Recurrent Neural Network (RNN). For example, Li et al.[13] proposed a Diffusion Convolutional Recurrent Neural Network (DCRNN) model which described the diffusion process of spatial network information through diffusion graph convolution network with RNN capturing the time correlation. The other approach is to combine GCN with Convolutional Neural Network (CNN). For example, Yu et al.[14] proposed a Spatial-Temporal GCN (ST-GCN), which used a CNN-based approach that combines a GCN layer with a 1D convolutional layer. Both approaches can easily create a connection between two vertices with very little correlation. The addition of an attention mechanism has since been proposed to solve this problem[15]. However, the attention mechanism tends to ignore those vertices which have dependencies but lack edges.

The key to spatial-temporal data prediction is to accurately extract the spatial-temporal properties from the data and build a prediction model on this basis. Spatial-temporal graph modeling faces a severe challenge in extracting dynamic spatial-temporal dependencies. Random Forest (RF) is an ensemble learning method proposed by Breiman in 2001[16]. RF is composed by multiple Decision Trees (DTs), which can not only deal with classification and regression problems, but also analyze the critical measure[17]. In this paper, an RF-based Variable Importance Measure (VIM) is designed to capture the spatial correlation between vertices in the spatial-temporal graph. RF provides two kinds of importance measures: Mean Decrease Accuracy (MDA) based on Out-Of-Bag (OOB) data and Mean Decrease Impurity (MDI) based on the Gini index[18]. The MDI of a feature is calculated as the improvement of the (weighted) average of the Gini impurities generated on RF for each variable in a single decision tree[19]. The MDA index is commonly used since it has no biases and can directly assess the impact of each variable on the RF model's forecast accuracy. The ability to capture potential spatial correlations in spatial-temporal data can be achieved with the help of RF's variable importance measure.

Furthermore, current research on spatial-temporal graph modeling has shown that it is inefficient for learning temporal dependence. For example, RNN-based methods usually perform poorly in long-term prediction tasks[20]. Wu et al.[21] developed Graph WaveNet (GWN), which uses adaptive adjacency matrix to capture spatiotemporal correlation by stacking spatiotemporal layers and expands causal convolution as a time convolution layer to expand the receptive field, allowing to capture longer sequences with fewer layers. Therefore, we propose a new spatial-temporal sequence prediction model (RF-GWN), which effectively combines RF and GWN. The model can effectively extract the complex and dynamic spatial-temporal dependencies in spatial-temporal data by using the VIM of RF. In terms of temporal correlation, the long series temporal feature extraction capability of GWN is fused to stack multiple spatial-temporal layers to capture longer time series features with a shallower network, effectively alleviating overfitting. In order to verify the model, two experiments are conducted on real traffic flow datasets and spatial-temporal groundwater level datasets.

DCRNN, ST-GCN, and GWN are used as baseline models to validate the effectiveness of RF-GWN. The experiments demonstrate that RF-GWN achieves better performance compared to the baseline models.

2 Methodology

2.1 Mathematical definition

In spatial-temporal graph model, the graph can be represented by $G=\{V,E,\boldsymbol{A}\}$, where V is a set of nodes, E is a set of edges, and $\boldsymbol{A}\in\mathbb{R}^{N\times N}$ is a weighted adjacency matrix representing the weighted adjacent relation of nodes. The graph G has a dynamic feature matrix $\boldsymbol{X}^t\in\mathbb{R}^{N\times M}$ at each time step t. The task is to predict the feature matrix of the graph G in the next T time steps $\boldsymbol{X}^{(t+1):(t+T)}\in\mathbb{R}^{N\times M\times T}$ given the previous S time steps $\boldsymbol{X}^{(t-S):t}\in\mathbb{R}^{N\times M\times S}$. The mapping relationship can be expressed as

$$\left[\boldsymbol{X}^{(t-S):t},G\right]\xrightarrow{f}\boldsymbol{X}^{(t+1):(t+T)} \tag{1}$$

2.2 Framework of RF-GWN

The proposed deep learning framework (RF-GWN) is displayed in Fig. 1. The framework consists of input layer, stacked spatial-temporal layers, and output layer. The proposed model can capture spatial dependencies at different temporal levels by stacking multiple spatial-temporal layers, with the bottom spatial-temporal layer receiving short-range spatial-temporal information and the top spatial-temporal layer handling long-range spatial-temporal information. The spatial-temporal layer is constructed by a graph convolution module and a Gated Temporal Convolution layer (Gated TCN). In the graph convolution module, an adaptive adjacency matrix is constructed to capture potential spatial dependencies by introducing RF. On one hand, the input is transformed by the linear layer and passed to the Gated TCN layer, and on the other hand, the dependencies between nodes are learned through RF and input to the graph convolutional neural network as an adaptive weight matrix, and different weights are assigned to different neighborhood nodes according to the importance scores as the basis for aggregating the neighborhood nodes. The output of the Gated TCN is passed to the graph convolutional neural network as the feature input to the graph convolutional neural network. The output of each layer is passed to the output layer through residual connections, and the output layer is set up with two ReLU activation layers as well as two linear layers.

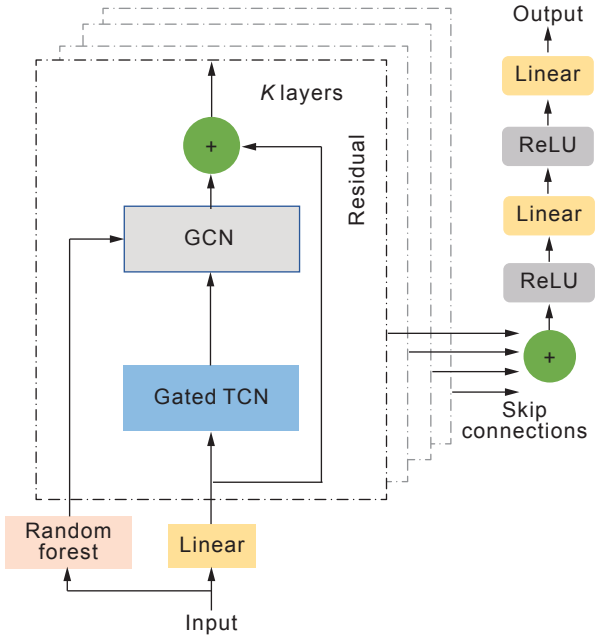

Fig. 1 Framework of RF-GWN.

2.3 Graph convolution layer combining RF

Graph convolution is used to look for hidden spatial dependencies from spatial-temporal data. To automatically infer spatial connectivity from data without any prior knowledge, we build an adaptive weight matrix base on the VIM of RF. RF is an integrated method of Classification And Regression Tree (CART) based on bootstrap samples and randomly selected features. As mentioned in introduction, the VIM of RF is calculated in two ways: MDA based on OOB data and MDI based on the Gini index[18]. We employ the nodes in the graph structure as features of the sample data in the RF model using the MDA calculation approach. Bootstrap sampling technique is carried out to extract training samples from the original data to construct the decision tree, and the rest of the data, which are the OOB data, are utilized to evaluate the accuracy of the decision tree[22, 23]. The OOB errors before and after adding random noise to a feature are calculated (Fig. 2). The average of the difference of the OOB error is the

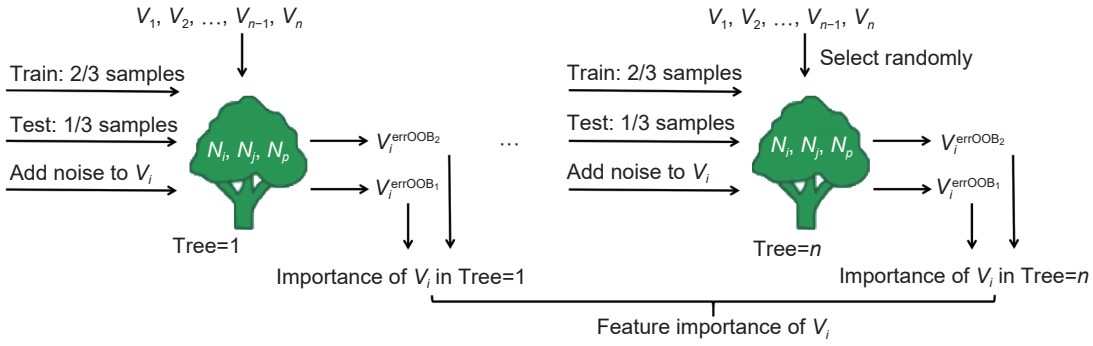

Fig. 2 Schematic of the method based on OOB sample for getting feature importance.

importance score of the feature in the whole RF. The importance score is used to measure the spatial dependency among the nodes in the graph structure. The importance scores of nodes can be described as Eq. (2).

$$\mathrm{FI}(V_i) = \frac{1}{\mathrm{nTree}} \sum_{i=1}^{\mathrm{nTree}} (V_i^{\mathrm{errOOB}_2} - V_i^{\mathrm{errOOB}_1}) \tag{2}$$

where $V_i^{\mathrm{errOOB}_1}$ is the OOB error of node V_i calculated by decision tree in RF; $V_i^{\mathrm{errOOB}_2}$ is the OOB error of node V_i with the random noise added; and nTree is the number of decision tree.

For some graph structures that lack prior knowledge, the spatial connectivity between nodes is difficult to infer. To solve this problem, we propose an adaptive adjacency matrix $\boldsymbol{A}_{\mathrm{opt}}$. The model sets a threshold ψ based on the importance scores of the nodes calculated from the VIM of RF. Nodes greater than or equal to ψ are used as the actual predicted neighbor nodes to the target node. The specific value of ψ needs to be determined according to the actual graph structures. The values of ψ will be discussed in the experimental section. The relationship between the adaptive adjacency matrix $\boldsymbol{A}_{\mathrm{opt}}$ and ψ is shown in Eq. (3).

$$\boldsymbol{A}_{\mathrm{opt}} = \begin{cases} \mathrm{FI}(V_i, J_j), & \mathrm{FI}(V_i, J_j) \geqslant \psi; \\ \boldsymbol{0}, & \mathrm{FI}(V_i, J_j) < \psi \end{cases} \quad (j = 1, 2, ..., N; j \neq i) \tag{3}$$

where $\mathrm{FI}(V_i, J_j)$ represents the importance score of the target node V_i relative to neighbor node V_j. The adjacency matrix $\boldsymbol{A}_{\mathrm{opt}}$ only describes the connectivity of the nodes, which can not adequately express the degree of the spatial dependencies between them. To solve this problem, we use the importance score as an adaptive weight matrix $\boldsymbol{R} \in \mathbb{R}^{N \times N}$. The weight matrix requires no prior knowledge and will be utilized as a model parameter throughout the model's training. We propose the graph convolution layer combining RF as

$$\boldsymbol{Z} = f(\boldsymbol{X}, \boldsymbol{A}_{\mathrm{opt}}) = \mathrm{Sigmoid}((\boldsymbol{A}_{\mathrm{opt}} \otimes \boldsymbol{R})\boldsymbol{X}\boldsymbol{W}) \tag{4}$$

where $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ denotes the input signals, $\boldsymbol{Z} \in \mathbb{R}^{N \times M}$ denotes the outputs, $\boldsymbol{W} \in \mathbb{R}^{D \times M}$ denotes the model parameter matrix, and $\otimes$ is the element-wise product.

By incorporating RF, the model is forced to only focus on the spatial dependencies between nodes with neighboring relationships, and the accuracy of prediction is improved. The redundant parameters in the model are reduced and the computational efficiency is improved. Figure 3 shows a schematic diagram of the graph convolution layer with RF.

2.4 Gated TCN

We employ dilated causal convolution as the temporal convolution layer to capture the temporal relationships. Dilated causal convolution can be sampled at intervals in the input of convolution process, and the sampling rate can be controlled artificially. As shown in Fig. 4, it can extend the convolution kernel's receptive field and lower network depth to some amount. Also, dilated casual convolution networks can efficiently handle long-range sequences without recursion manner, reducing the occurrence of gradient disappearance and explosion[24].

Mathematically, given a 1-dimensional input sequence $\boldsymbol{x}$ and a filter f, the dilation causal convolution operation of $\boldsymbol{x}$ with f at step t can be expressed as Eq. (5).

$$\boldsymbol{x} * f(t) = \sum_{s=0}^{K-1} f(s)\boldsymbol{x}(t - d \times n) \tag{5}$$

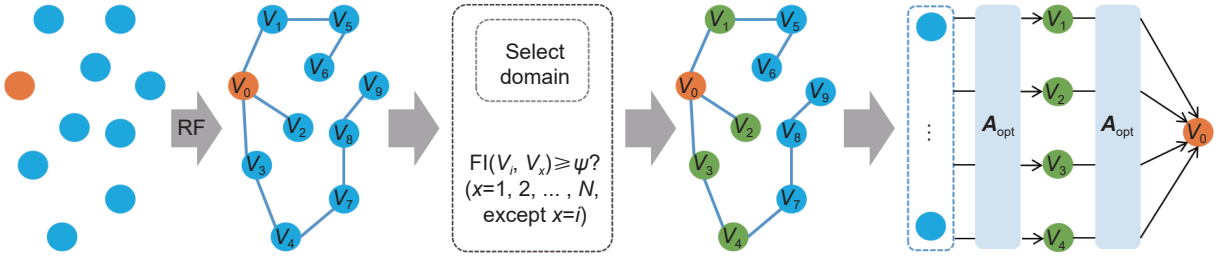

Fig. 3 Graph convolution layer with RF.

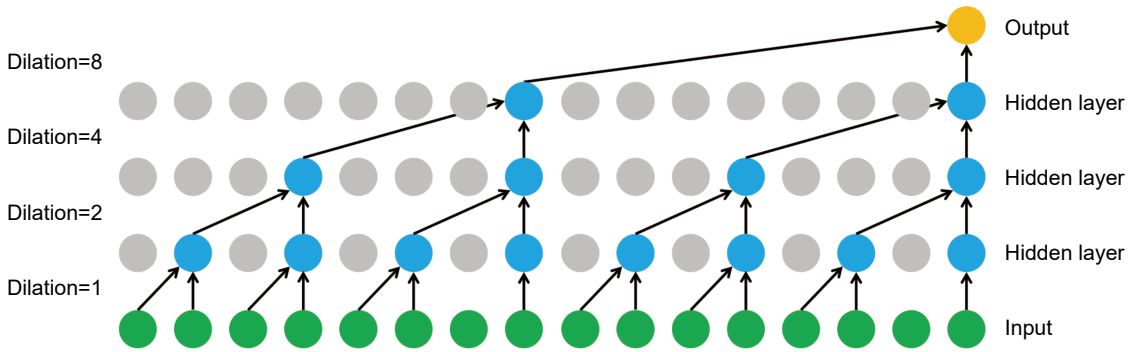

Fig. 4 Dilated casual convolution.

where d represents the dilation factor. By stacking dilated causal convolution layers with dilation factors in an increasing order, the receptive field of TCN also increases[25].

The temporal convolution module selects the gating mechanism to control the flow of information in each layer of the temporal convolutional network[24]. As shown in Fig. 5, Gated TCN is composed of two TCN modules. According to Ref. [26], a TCN module is a dilated causal convolution and a hyperbolic tangent activation function. Another TCN module is a dilated causal convolution and a Sigmoid activation function to control the ratio of the delivered information.

The formulation of Gated TCN is presented as follows:

$$h = g(\Theta_1 * \boldsymbol{X} + b) \otimes \sigma(\Theta_2 * \boldsymbol{X} + c) \tag{6}$$

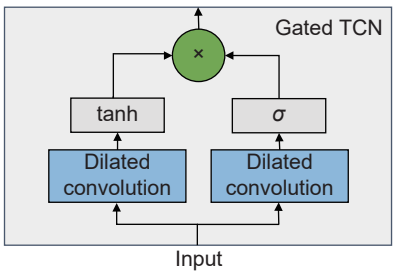

Fig. 5 Gated TCN module.

where $\boldsymbol{X} \in \mathbb{R}^{N \times M \times S}$ is the input, Θ_1, Θ_2, b, and c are the model parameters of the extended causal convolution, $g(\cdot)$ is the output activation tanh function, and $\sigma(\cdot)$ is the Sigmoid function that determines the information ratio to the next layer.

3 Experiment

3.1 Experimental setups

In order to verify the proposed model RF-GWN, experiments are conducted on two real traffic datasets (METR-LA and PEMS-BAY) and a spatial-temporal groundwater dataset. METR-LA and PEMS-BAY (in Section 3.4) are typical spatial-temporal datasets, which are often used to test the feature extraction ability of algorithms for spatial-temporal data. The groundwater dataset (in Section 3.5) is a small dataset with hundreds of records. To the best of our knowledge, existing researches for this kind of problems only considered the time-relations between data. However, it is obvious that spatial relations exist between different geographical locations where observation boreholes are deployed. Therefore, the two kinds of datasets are used to verify the effectiveness and applicability of RF-GWN.

The data are normalized using *Z*-score as

$$Z = \frac{X - \overline{X}}{K} \tag{7}$$

where X is the sample to be processed, $\overline{X}$ represents the average value of the sample, and K is the standard deviation of the sample.

The hardware and software environments for developing the proposed model and conducting the experiments are shown in Table 1.

3.2 Baseline

To evaluate the ability of RF-GWN in capturing the spatial-temporal correlations, three traditional baseline models are compared, including GWN[21], DCRNN[13] and ST-GCN[14], where GWN combines dilated causal convolution with adaptive adjacency matrix to extract spatial dependence, DCRNN combines GCN with RNN in an encoder-decoder manner, and ST-GCN combines graph convolution with 1D convolution.

3.3 Performance criteria

Three indexes are implemented in this work to assess the performance of the proposed model, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The indexes are defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{8}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{9}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{10}$$

where n is the number of training or test samples; y_i and $\hat{y}_i$ are the actual value and predicted value, respectively. The average amount of the error between the model's predicted value and actual value is measured by the RMSE. The average of the absolute errors between the predicted value and actual value is calculated using MAE. MAPE is used to calculate the percentage error between the expected and actual value. For the three indexes, the lower the RMSE, MAE, and MAPE, the better the performance of the model.

Table 1 Environments of the experiment.

Item	Configuration
System	Ubuntu 18.04.4 LTS
Software	Python 3.7.6
	PyCharm community edition 2 019.2.5
	PyTorch 1.1.0
Hardware	Intel(R) Xeon(R) Silver 4110 CPU @ 2.10 GHz
	16 GB RAM
	NVIDIA GEFORCE RTX 2080 Ti

3.4 Experiments on traffic dataset

3.4.1 Dataset

METR-LA records four-month traffic speed statistics of 207 sensors on Los Angeles county highways from March 1st, 2012 to June 30th, 2012. PEMS-BAY contains six-month traffic speed information of 325 sensors in the Bay area from January 1st, 2017 to May 31st, 2017. In this paper, the same data preprocessing is adopted for all the methods: the data are processed into 5-min time interval and normalized by *Z*-score. At the same time, the datasets are divided into training, verification, and test datasets in chronological order with the percentage of 70%, 10%, and 20%, respectively. The summary of the datasets is shown in Table 2.

3.4.2 Parameter

The ideal threshold ψ setting in traffic flow prediction is the first topic we cover in this section. Figure 6 compares the performance of 15-min advance predictions for the METR-LA and PEMS-BAY datasets when different thresholds are set for RF-GWN, where the thresholds increase from 0 to 0.15 with an interval of 0.01. It can be seen that the lowest RMSE and MAE are obtained when the threshold value is 0.08 and 0.11 in METR-LA and PEMS-BAY datasets. The learning rate is set to 0.001 and the batch size is set to 64 for both datasets. Weight decay is set to 0.0001 in order to prevent overfitting. Dropout with p=0.3 is applied to the outputs of the graph convolution

Table 2 Summary of METR-LA and PEMS-BAY.

Dataset	Number of nodes	Number of time steps
METR-LA	207	34 272
PEMS-BAY	325	52 116

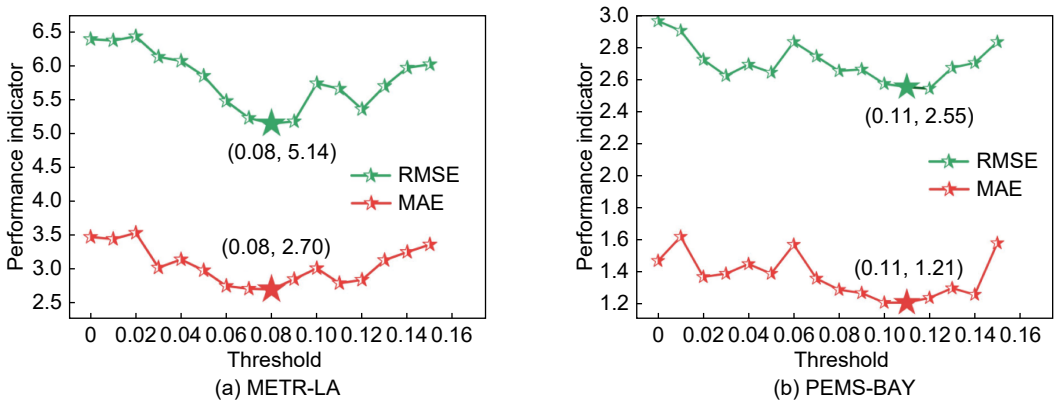

Fig. 6 Evolution of RMSE and MAE.

layer. We optimize model parameters of RF-GWN by minimizing the Mean Absolute Error (MAE) loss with stochastic gradient descent. In addition, the hyperparameters of the RF for conducting experiments are shown in Table 3.

3.4.3 Results and analysis

(1) Effects of adaptive weight matrix

On the METR-LA dataset, the accuracy of the adaptive adjacency matrix produced via RF-based model learning is verified. Figure 7a shows the case marked on the map, and Fig. 7b is a partial heat map of the adaptive adjacency matrix in this paper, where the horizontal coordinates only show the first 50 nodes. According to Fig. 7b, different columns in the adaptive adjacency matrix are more different, and some columns have more high-value points than others. For example, columns 14, 34, and 47 in Fig. 7b have fewer high-value points compared to columns 9, 41, and 43, which indicates that nodes 9, 41, and 43 have an impact on more nodes in the graph. The reliability of the results can be confirmed by the actual geographic traffic distribution map in Fig. 7a. Nodes 9, 41, and 43 are located near the intersection of several major roads, while nodes 14, 34, and 47 are located on a roadway farther from the intersection, and it is clear that traffic speeds at the intersection node have an impact on traffic speeds at more nodes. This demonstrates that the model's spatial dependencies can be captured more precisely by the adaptive adjacency matrix learned

Table 3 RF hyperparameters.

Parameter	Value	Parameter	Value
Bootstrap	True	Criterion	MSE
Max_features	Auto	Min_samples_split	2
OOB_score	False	n_estimators	10
n_jobs	1	Min_samples_leaf	1

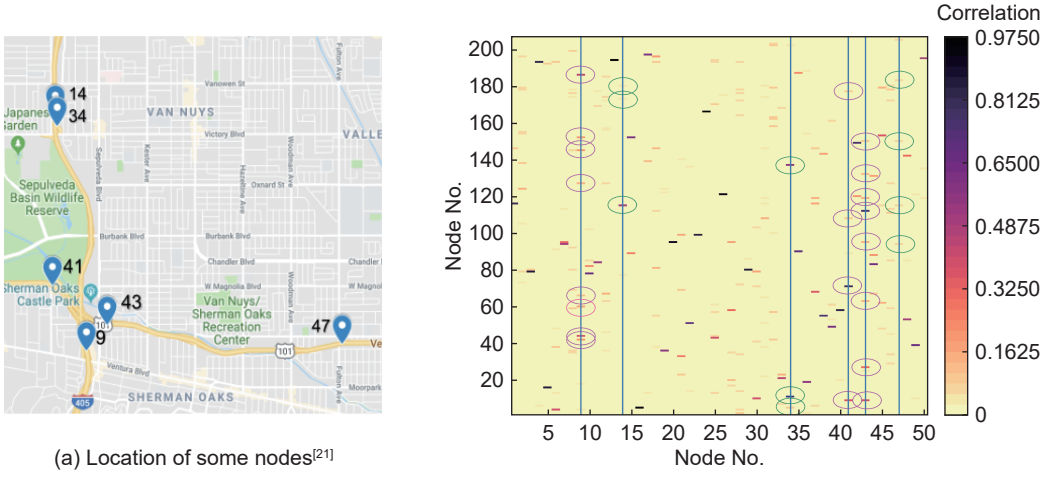

Fig. 7 Part of self-adaptive adjacency matrix.

based on RF. It also shows that RF-GWN can still function well even in the presence of a flawed graph structure.

(2) Accuracy

The predicted values (30-min-ahead) versus actual values of the RF-GWN model on the two traffic datasets are plotted in Figs. 8 and 9 . As can be observed, the predicted value of RF-GWN always falls in the middle of the actual values, which can accurately predict the missing part of the actual values and achieve stable prediction. In particular, RF-GWN can effectively extract local features and accurately predict the traffic speed of the entire traffic network.

Tables 4 and 5 show the performance comparison of the model prediction in this paper with the other three models in advance of 15 min, 30 min, and 60 min on the two datasets. As can be observed from Tables 4 and 5, the RMSE, MAE, and MAPE indexes of the RF-GWN model are significantly better than the convolution-based method ST-GCN and also better than the recursive-based method DCRNN. Although the prediction performance of the RF-GWN model on the METR-LA dataset in the 15-min horizons is lower than that of GWN, the difference between the two is almost negligible. This may be due to the small sample size of the METR-LA dataset. In particular, on the PEMS-BAY dataset, compared to the best model GWN in the benchmark model, a large improvement is achieved in the 15-min and 30-min horizons, and more pronounced in the 60-min horizons, even surpassing GWN's improvement over its predecessor. In addition, the number of parameters in RF-GWN (2.48×10^5) only accounts for around half of ST-GCN. Compared to ST-GCN which processes 228 nodes, the proposed

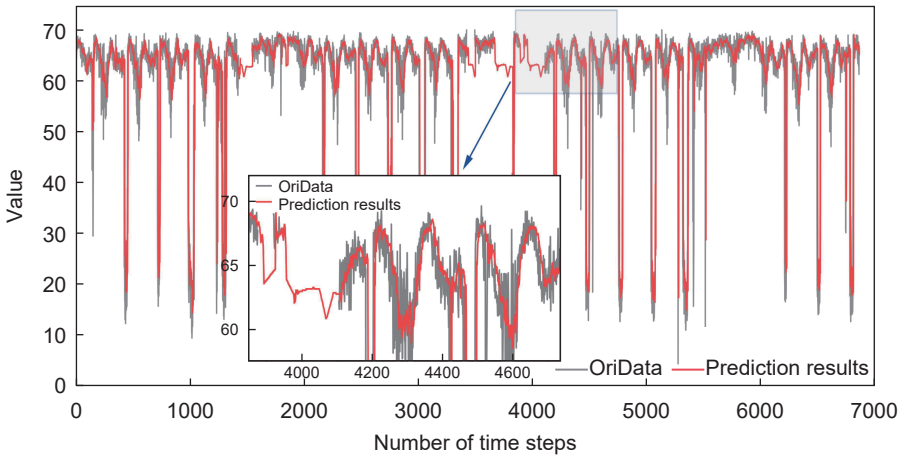

Fig. 8 RF-GWN prediction curves (METR-LA).

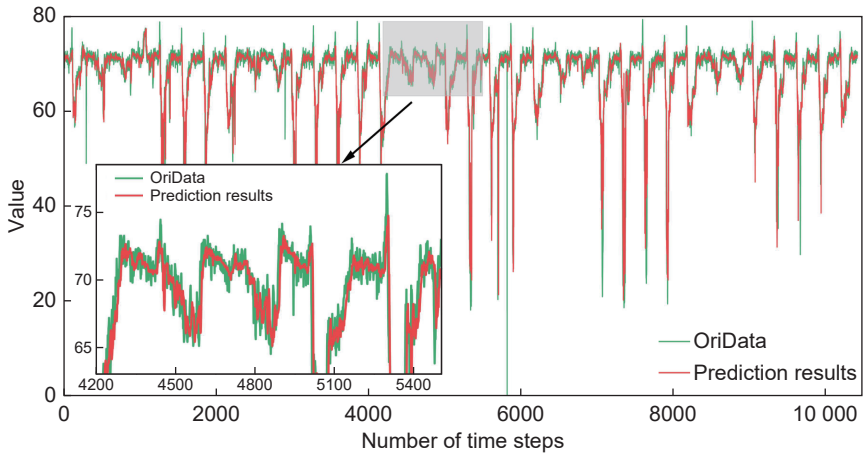

Fig. 9 RF-GWN prediction curves (PEMS-BAY).

Table 4 Comparison of performance criteria for different models on METR-LA dataset.

Model	15 min			30 min			60 min		
	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
DCRNN	2.77	5.38	7.30	3.15	6.45	8.80	3.60	7.60	10.50
ST-GCN	2.88	5.74	7.62	3.47	7.24	9.57	4.59	9.40	12.70
GWN	**2.69**	5.15	**6.90**	3.07	6.22	8.37	3.53	7.37	**10.01**
RF-GWN	2.70	**5.14**	6.94	**3.06**	**5.99**	**8.30**	**3.51**	**7.19**	10.09

Table 5 Comparison of performance criteria for different models on PEMS-BAY dataset.

Model	15 min			30 min			60 min		
	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
DCRNN	1.38	2.95	2.90	1.74	3.97	3.90	2.07	4.74	4.90
ST-GCN	1.36	2.96	2.90	1.81	4.27	4.17	2.49	5.69	5.79
GWN	1.30	2.74	2.73	1.63	3.70	3.67	1.95	4.52	4.63
RF-GWN	**1.21**	**2.55**	**2.58**	**1.44**	**3.13**	**3.13**	**1.80**	**4.08**	**4.23**

RF-GWN processes 325 nodes in PEMS-BAY dataset. This indicates that our model significantly reduces redundant parameters. The experimental results of RF-GWN confirm that by introducing RF to improve the model's ability to capture spatial-temporal dependencies, our method can make more accurate predictions than GWN, and the ability is improved at each time stage relative to the baseline model, especially the superiority in long-term prediction.

3.5 Experiments on groundwater dataset

3.5.1 Dataset

The dataset is provided by the " China Western Environment and Ecology Science Data Center" (http://www.ncdc.ac.cn/). The study area was selected from the middle reaches of the Heihe River basin (38°38′N−39°53′N, 98°53′E−100°44′E), which covers an area of roughly 9016 km^2 (refer to Fig. 2 in a previous research carried out by Chen et al.[3] in Heihe River Basin). This area was chosen because of the abundance of groundwater observation data and the large area and time span contained in these data, making it an ideal spatial-temporal dataset. The annual precipitation in the area is low and concentrated, and it belongs to the area with scarce groundwater resources. The changes of groundwater level within one year are affected by precipitation, evapotranspiration, surface water, and agricultural irrigation. The annual low point of groundwater level is from March to May, and the peak is from June to September. The observation dataset records historical groundwater level data from 42 water level observation boreholes in the area. The dataset was collected from January 1986 to December 2008, recording 276 sets on a monthly basis.

Since some stations had missing values in the groundwater level data (Fig. 10), the Lagrangian linear interpolation method is used to repair and extend the original data. The processed data are expanded to 2750 sets. In addition, we normalized the expanded groundwater level data using a Z-score normalization process and returned to the actual values to assess the prediction accuracy. 80% of the dataset is selected as the training set, 10% as the validation set, and the remaining 10% as the test set.

3.5.2 Parameter

One of the key parameters of the RF-GWN model is the threshold value ψ. Different ψ could have a great impact on the prediction effect of the model. In this section, we have carried out extensive studies to determine the threshold's ideal value. The remaining hyperparameters of the model for predicting groundwater level are comparable to those for predicting traffic flow.

The node importance scores calculated by the VIM of RF provide ideas for the value of the threshold ψ. Considering the calculation volume and prediction accuracy, the top 7 observation boreholes with the highest importance score for the target observation borehole are selected, and the sum of the importance scores is taken as the optimal threshold value for the

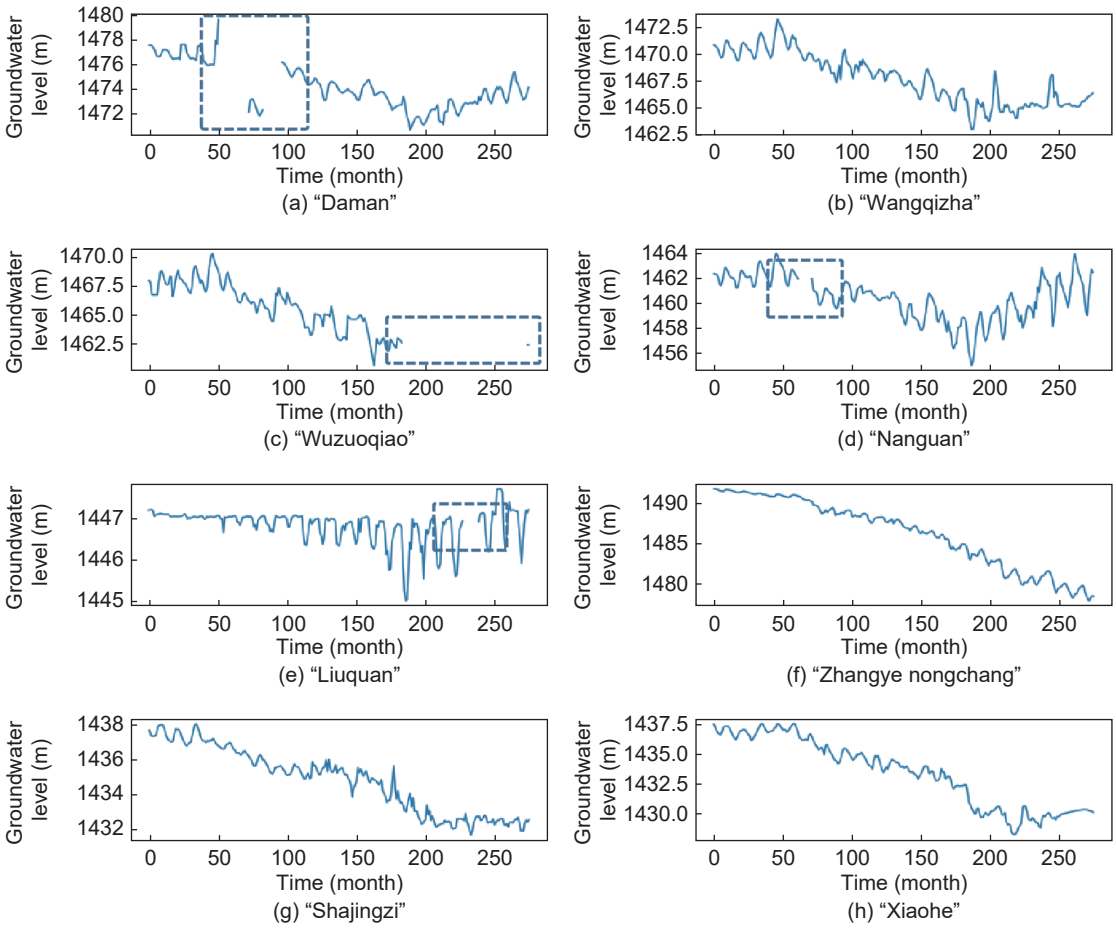

Fig. 10 Observed groundwater level in different boreholes.

target observation borehole. In our experiments, the optimal ψ of the 42 water level observation boreholes are obtained separately as shown in Fig. 11, and it can be seen that the mean value of the 42 thresholds is 0.099 965, the median value is 0.099 85, and the mode value is 0.1.

We tested the effects of ψ=0.099 965, ψ=0.099 85, and ψ=0.1 on the model performance, respectively. The performance of RF-GWN on the groundwater dataset for different ψ is compared in Table 6. The results indicate that the prediction results are similar when the median or mode is selected as the threshold value, and both are better than the mean value, so the parameter threshold ψ is set to 0.1 in this experiment.

3.5.3 Results and analysis

(1) Effects of adaptive weight matrix

Next, the spatial dependence capture capability of our model is evaluated. Taking the observation boreholes "Wangqizha" and "22" as the target observation boreholes, Fig. 12 shows the importance score of each observation borehole to the target

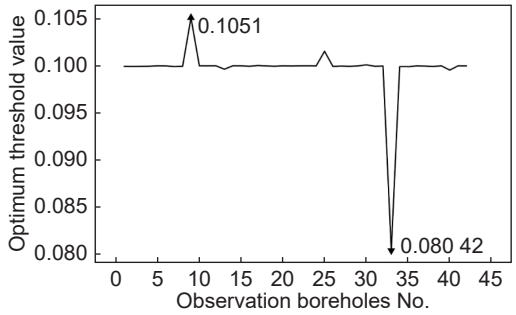

Fig. 11 Optimum threshold value of each observation boreholes.

Table 6 Comparison of performance criteria for different thresholds on groundwater dataset.

Threshold	RMSE	MAE	MAPE (%)
Average value	0.089 86	0.078 32	0.005 719
Median value	0.088 92	0.075 16	0.005 360
Mode value	0.088 81	0.075 04	0.005 349

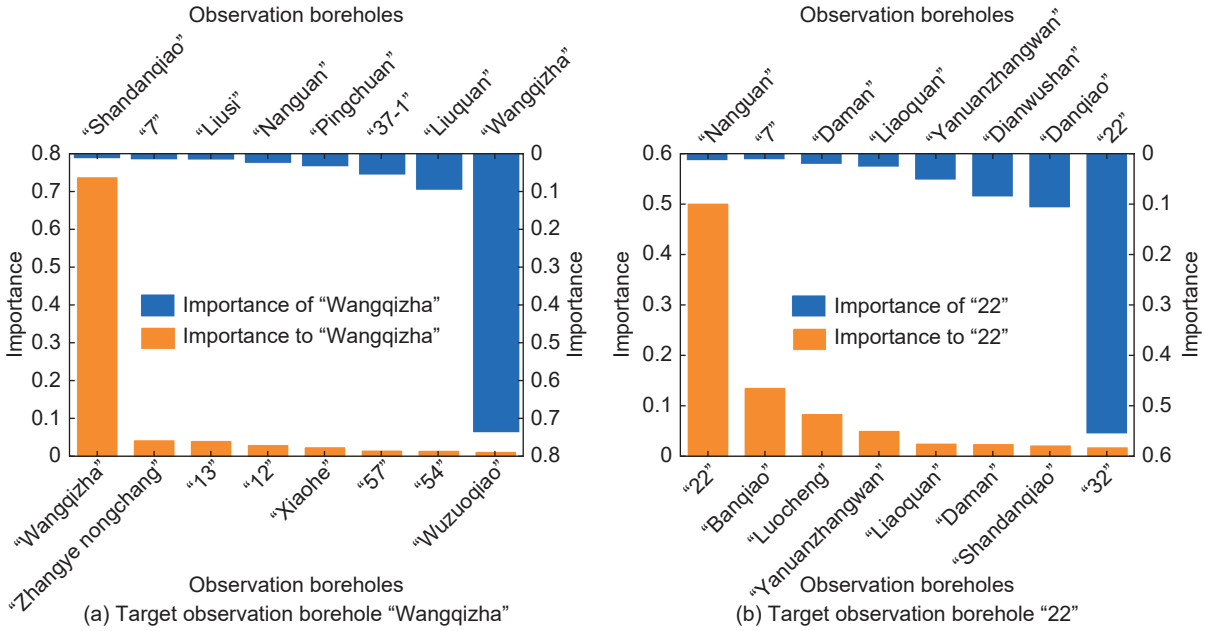

Fig. 12 Evaluation of the importance score of/to the target observation borehole.

observation borehole (yellow bar) and the importance score of the target observation borehole to the other observation borehole (blue bar) learned by our model. It can be found that the experimentally obtained spatial dependence correlates with the actual geographic distribution of the boreholes shown in the study area, which is also in accordance with the natural law, in which the flow of groundwater is inevitably influenced by topographic conditions, such as the undulations of the local terrain. It can be found from the experiment that the number of observation boreholes affecting each target observation borehole varies, but most of the observation boreholes are affected by only 6−7 nearby observation boreholes and the magnitude of the effect is also different (which provides a basis for threshold ψ setting). Our model is capable of adaptively capturing boreholes with high influence on the water level of the target borehole in terms of spatial characteristics, and assigning different weights according to the degree of influence. We believe that this is due to the advantage of introducing the spatial dependence extraction capability of RF and the automatic assignment of weights. This mechanism can significantly capture the spatial dependence between nodes in the graph structure, which is beneficial to the prediction accuracy. Compared with global prediction, the RF-GWN model can reduce the computational effort.

(2) Accuracy

The predicted values and actual values of RF-GWN and baseline models (DCRNN, ST-GCN, and GWN) are plotted on the snapshot of the test data in Fig. 13. The size of predicted time window is set to 12 month, which denotes the length of the output sequence. Taking two target observation boreholes (observation boreholes "Wangqizha" and "22") as examples, it can be seen that the two target observation boreholes have different patterns of variation, for example, the observation borehole in Fig. 13a has an obvious rising trend of groundwater level, and the observation borehole in Fig. 13b has a larger fluctuation of groundwater level. RF-GWN can extract complex groundwater level change patterns from different observation boreholes to obtain better prediction performance. The results show that the prediction curves of the RF-GWN model reflect more details of the local fluctuations and can better extract local features and adapt to the complex groundwater level change trends of different nodes. As shown in the enlarged region in Fig. 13, the prediction results are more accurate than those of GWN.

To evaluate the overall prediction performance of RF-GWN model and other baseline models on the groundwater dataset, Fig. 14 shows the average of the predicted RMSE of different models over 42

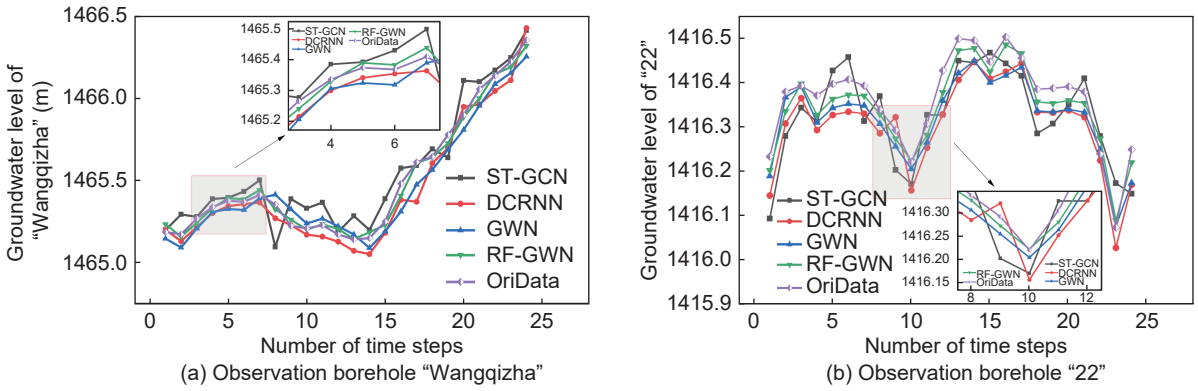

Fig. 13 Prediction results of groundwater level of target observation boreholes.

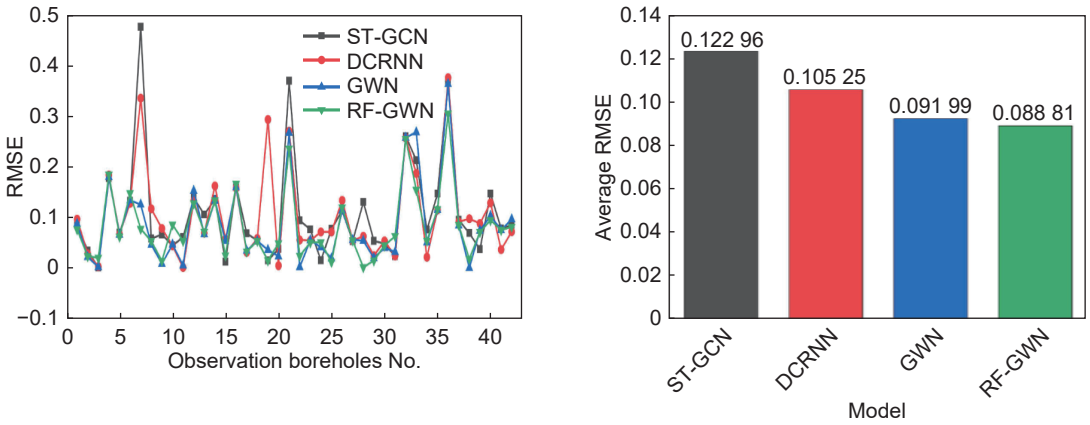

Fig. 14 RMSE of 42 observation boreholes.

observation boreholes. It can be seen that the prediction accuracy of RF-GWN is better than the classical spatial-temporal graph neural network models such as DCRNN and GWN, proving that RF-GWN can model the groundwater spatial-temporal data more effectively. In Table 7, the mean value of index of 42 observation boreholes and the average training time cost of each epoch are recorded. It can be observed that ST-GCN has the highest efficiency during the training phase. DCRNN consumes significantly more time than other methods because of the requirement of long sequence training in recurrent networks. RF-GWN and GWN consume almost the same amount of time in the training phase. In conclusion, RF-GWN has improved the prediction performance by introducing RF, however, there is still room for further improvement in computational efficiency.

4 Conclusion

In this work, an improved spatial-time-series forecasting model RF-GWN is proposed, which combines RF and GWN for the first time to efficiently capture spatial-temporal dependencies. RF-GWN captures potential spatial correlations through the VIM of RF and uses the dilation causal convolution as a temporal convolution layer, enabling the temporal

Table 7 Comparison of prediction results and training time of different models.

Model	RMSE	MAE	MAPE (%)	Training time of each epoch (s)
ST-GCN	0.122 96	0.107 84	0.007 611	2.12
DCRNN	0.105 25	0.093 31	0.006 625	39.75
GWN	0.095 99	0.079 72	0.005 618	5.64
RF-GWN	**0.088 81**	**0.072 04**	**0.005 349**	**5.91**

convolution layer to capture longer sequences using fewer layers. Experiments are conducted on real datasets in two domains, traffic flow and groundwater level, and analyzed in detail from three aspects: adaptive weight matrix, prediction results, and evaluation index comparison of prediction results from multiple baseline models. The results show that the introduction of RF can make more accurate predictions, and prove the applicability and effectiveness of the model for predicting two types of spatial-temporal distribution data: traffic flow and groundwater level. However, due to the small amount of groundwater sample data used, the expanded dataset failed to reach tens of thousands, the future work of the model may include exploring applications on large datasets.

References

[1] G. Zhang, H. He, and D. Katabi, Circuit-GNN: Graph neural networks for distributed circuit design, in *Proc. 36th International Conference on Machine Learning*, Long Beach, CA, USA, 2019, pp. 7364–7373.

[2] F. R. K. Chung, Spectral graph theory, CBMS Regional Conference Series in Mathematics, https://doi.org/10.1090/cbms/092, 1997.

[3] C. Chen, W. He, H. Zhou, Y. Xue, and M. Zhu, A comparative study among machine learning and numerical models for simulating groundwater dynamics in the Heihe River Basin, northwestern China, *Scientific Reports*, vol. 10, no. 1, p. 3904, 2020.

[4] Y. Lin and Y. Yang, Stock markets forecasting based on fuzzy time series model, in *Proc. 2009 IEEE International Conference on Intelligent Computing & Intelligent Systems*, Shanghai, China, 2009, pp. 782–786.

[5] B. Gui, X. Wei, Q. Shen, J. Qi, and L. Guo, Financial time series forecasting using support vector machine, in *Proc. 2014 Tenth International Conference on Computational Intelligence and Security*, Kunming, China, 2014, pp. 39–43.

[6] S. H. Holan, Long-memory time series theory and methods, *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1715–1716, 2008.

[7] E. E. Osuna, *Support Vector Machines: Training and Application*. Cambridge, MA, USA: Massachusetts Institute of Technology, 1998.

[8] S. R. Sain, The nature of statistical learning theory, *Technometrics*, vol. 38, no. 4, p. 409, 1996.

[9] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[10] T. N. Kipf and M. Welling, Semi-supervised classification with graph convolutional networks, arXiv preprint arXiv: 1609.02907, 2016.

[11] W. L. Hamilton, R. Ying, and J. Leskovec, Inductive representation learning on large graphs, in *Proc. 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 1025–1035.

[12] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, Graph attention networks, arXiv preprint arXiv: 1710.10903, 2017.

[13] Y. Li, R. Yu, C. Shahabi, and Y. Liu, Diffusion convolutional recurrent neural network: Data-driven traffic forecasting, arXiv preprint arXiv: 1707.01926, 2018.

[14] B. Yu, H. Yin, and Z. Zhu, Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting, in *Proc. Twenty-Seventh International Joint Conference on Artificial Intelligence*, Melbourne, Australia, 2018, pp. 3634–3640.

[15] J. Zhang, X. Shi, J. Xie, H. Ma, I. King, and D. -Y. Yeung, GaAN: Gated attention networks for learning on large and spatiotemporal graphs, arXiv preprint arXiv: 1803.07294, 2018.

[16] L. Breiman, Random forests, *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[17] S. Song, R. He, Z. Shi, and W. Zhang, Variable importance measure system based on advanced random forest, *Computer Modeling in Engineering & Sciences*, vol. 128, no. 1, pp. 65–85, 2021.

[18] T. Hastie, R. Tibshirani, and J. Friedman, Random forests, in *the Elements of Statistical Learning*. New York, NY, USA: Springer, 2009, pp. 587–604.

[19] M. Loecher, From unbiased MDI feature importance to explainable AI for trees, arXiv preprint arXiv: 2003.12043, 2020.

[20] Y. Seo, M. Defferrard, P. Vandergheynst, and X. Bresson, Structured sequence modeling with graph convolutional recurrent networks, in *Proc. 25th International Conference on Neural Information Processing*, Siem Reap, Cambodia, 2018, pp. 362–373.

[21] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, Graph wavenet for deep spatial-temporal graph modeling, arXiv preprint arXiv: 1906.00121, 2019.

[22] A. Fisher, C. Rudin, and F. Dominici, All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously, *J. Mach. Learn. Res.*, vol. 20, no. 177, pp. 1–81, 2019.

[23] D. H. Wolpert and W. G. Macready, An efficient method

to estimate bagging's generalization error, *Machine Learning*, vol. 35, no. 1, pp. 41–55, 1999.

[24] Z. Wu, S. Pan, G. Long, J. Jiang, X. Chang, and C. Zhang, Connecting the dots: Multivariate time series forecasting with graph neural networks, in *Proc. 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, virtual event, CA, USA, 2020, pp. 753–763.

[25] G. Jin, C. Liu, Z. Xi, H. Sha, Y. Liu, and J. Huang, Adaptive dual-view wavenet for urban spatial–temporal event prediction, *Information Sciences*, vol. 588, pp. 315–330, 2022.

[26] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, Language modeling with gated convolutional networks, arXiv preprint arXiv: 1612.08083, 2016.

Chong Chen received the BSc, MSc, and PhD degrees from Lanzhou University in 2010, 2012, and 2017, respectively. He is currently an associate professor at the College of Information Science and Engineering, China University of Petroleum-Beijing. His research interests include numerical modeling, data assimilation, and machine learning. He is a member of China Computer Federation (CCF) and Chinese Association for Artificial Intelligence (CAAI).

Yanbo Xu is currently pursuing the master degree of engineering in China University of Petroleum-Beijing. Her major is electronic information engineering. Her research interests include machine learning and information prediction.

Jixuan Zhao is currently pursuing the master degree of engineering in the China University of Petroleum-Beijing. His major is electronic information engineering. His research interests include machine learning, graph neural networks, and information prediction.

Lulu Chen received the master degree of engineering from China University of Petroleum-Beijing in 2021. She is an engineer of the Education Management Information Centre of the Ministry of Education now. Her major is information and communication engineering. Her research interests include machine learning, graph neural networks, and information prediction.

Yaru Xue received the BS degree in information technology from Central China Normal University, Wuhan, China in 1994, the MS degree in information and communication engineering from Lanzhou University, Lanzhou, China in 2001, and the PhD degree in geophysics from China University of Petroleum-Beijing, Beijing, China in 2009. From 2009 to 2010, she was a visiting scholar in the Geophysics Department of University of Illinois at Urbana-Champaign. She is an associate professor in the College of Information Science and Engineering, China University of Petroleum-Beijing. Her research interests lie in seismic data processing, including radon transform, seismic data interpolation and denoising, and applications of machine learning in exploration geophysics.