Modeling Dynamic Spatial Influence for Air Quality Prediction with Atmospheric Prior

Dan Lu[1], Le Wu[2] , Rui Chen[1(✉)], Qilong Han[1] , Yichen Wang[3], and Yong Ge[4]

[1] Harbin Engineering University, Harbin, Heilongjiang, China
{ludan,ruichen}@hrbeu.edu.cn
[2] Hefei University of Technology, Hefei, Anhui, China
[3] Hunan University, Changsha, Hunan, China
[4] The University of Arizona, Tucson, AZ 85721, USA
yongge@arizona.edu

Abstract. Air quality prediction is an important task benefiting both individual outdoor activities and urban emergency response. To account for complex temporal factors that influence long-term air quality, researchers have formulated this problem using an encoder-decoder framework that captures the non-linear temporal evolution. Besides, as air quality presents natural spatial correlation, researchers have proposed to learn the spatial relation with either a graph structure or an attention mechanism. As well supported by atmospheric dispersion theories, air quality correlation among different monitoring stations is dynamic and changes over time due to atmospheric dispersion, leading to the notion of dispersion-driven dynamic spatial correlation. However, most previous works treated spatial correlation as a static process, and nearly all models relied on only data-driven approaches in the modeling process. To this end, we propose to model dynamic spatial influence for air quality prediction with atmospheric prior. The key idea of our work is to build a dynamic spatial graph at each time step with physical atmospheric dispersion modeling. Then, we leverage the learned embeddings from this dynamic spatial graph in an encoder-decoder model to seamlessly fuse the dynamic spatial correlation with the temporal evolution, which is key to air quality prediction. Finally, extensive experiments on real-world benchmark data clearly show the effectiveness of the proposed model.

Keywords: Air quality prediction · Dynamic spatial correlation · Atmospheric dispersion

Supported by the National Key R&D Program of China under Grant No. 2020YFB1710200, the National Natural Science Foundation of China under Grant No. 61872105 and No. 62072136, the Fundamental Research Funds for the Central Universities under Grant No. 3072020CFT2402 and No. 3072020CFT0603, and the Opening Fund of Acoustics Science and Technology Laboratory under Grant No. SSKF2020003.

L. H. U et al. (Eds.): APWeb-WAIM 2021, LNCS 12859, pp. 384–398, 2021.
https://doi.org/10.1007/978-3-030-85899-5_28

1 Introduction

With the fast pace of urbanization and industrialization, air pollution has been an endemic threat to human health and the environment, especially in metropolitan cities. Air pollution generally refers to the release of pollutants into the air, which is detrimental to human health and the planet as a whole. To prevent human beings from long-term exposure of pollution and reduce air pollution, accurately predicting future air quality is essential. For example, policy makers can properly choose guides or policies, such as temporary traffic control or production ban for heavy-polluting factories, according to the future air quality trend in order to reduce the severity of local pollution levels.

Precisely predicting air quality, often in terms of the major pollutant PM2.5 value, is non-trivial. This is due to the fact that air quality depends on multiple complex factors, such as meteorology, road networks, and point of interests (POIs), and evolves over time. While previous works carefully designed sophisticated static features and temporal features for air quality prediction, recent studies have begun to use recurrent neural networks (RNNs) to capture the non-linear temporal evolution. In particular, researchers proposed to use an encoder-decoder framework, with the encoder fusing heterogeneous features and the decoder predicting long-term PM2.5 values [21].

Besides temporal correlation, PM2.5 values among different air quality monitoring stations naturally exhibit spatial autocorrelation, with nearby stations having similar PM2.5 values. Researchers have proposed to incorporate spatial correlation by including nearby stations' features in the input space [12,14] or by further considering the Pearson correlation of geo-context features between a target station and its neighboring stations [4,23]. Instead of having nearby stations defined by spatial distance contribute equally to a target station, attention mechanisms have been increasingly used to differentiate the weights of different monitoring stations, where the attentive weights are either static over time [5] or dynamic (i.e., having different weights at different time steps) [13]. Researchers have also proposed to leverage a graph structure to capture the higher-order spatial correlations among stations and to learn the graph structure to facilitate weather prediction [19]. These attempts have demonstrated that modeling spatial correlations among stations can boost air quality prediction performance.

In view of the importance of spatial correlation for air quality prediction, we argue that the current solutions for spatial modeling are still far from satisfactory. In fact, the well-established and widely-used atmospheric dispersion models [2,17] have pointed out that air quality correlation among different monitoring stations is inherently *dynamic* and changes over time. In particular, how air pollutants disperse from a station to another relies on not only their spatial distance and direction, but also other dynamic factors, such as wind direction and speed, leading to the notion of *dispersion-driven dynamic spatial correlation*. Atmospheric dispersion modeling provides a mathematical simulation of how air pollutants disperse in the ambient atmosphere, and is built on top of expert knowledge. For example, in the Gaussian plume model, the concentration of pollutant downwind from a source is treated as spreading outward from

the centerline of the plume following a Gaussian statistical distribution in both vertical and horizontal directions [1]. These physical models provide us a solid theoretical foundation to guide air quality prediction. By far, atmospheric dispersion models are still the dominant models used in air quality policy making. However, most related works adopted only *static* spatial correlation modeling methods. What's worse, almost all of these works relied on a purely data-driven approach, which may introduce unnecessary noise and violate well-established dispersion theories due to the black box nature of deep learning models.

In this paper, we focus on modeling the dynamic spatial influence for air quality prediction with atmospheric prior. This is particularly challenging as it is still unknown how to leverage atmospheric prior to model dynamic spatial correlation among stations and integrate these well-established theories into a data-driven air quality prediction process. To tackle these challenges, we first build a dynamic spatial graph at each time step with the simple yet effective Gaussian plume model, which can well capture the dynamic higher-order spatial correlations among monitoring stations. Then, we incorporate the embeddings learned from dynamic spatial graphs using graph convolutional networks (GCNs) into an encoder-decoder model to seamlessly fuse the dynamic spatial correlation with the temporal evolution. The key technical contribution of this paper lies in combining knowledge-driven atmospheric dispersion models with data-driven deep learning techniques for air quality prediction in an elegant way. Finally, experimental results on real-world benchmark datasets clearly demonstrate the superiority of our proposed model over the state-of-the-art methods.

2 Related Work

Air quality prediction has been a long-standing research problem with practical importance. Existing methods roughly fall into two categories: classical physical models and data-driven models. Physical models have been widely used in the early stage of air quality prediction research. They explicitly simulate the actual physical dispersion process of air pollutants and feature a rigorous mathematical foundation. Gaussian plume models [2] and Street Canyon models [17] are most widely-used physical models that estimate future pollutants' concentration by considering a few important factors, such as meteorological conditions, source term, emissions or release parameters, and terrain elevations. While these physical models work well in relatively simple conditions, they lack the capability of learning from more complex urban big data involving a large number of external factors, and fall short of expectations in practice. In this paper, we propose a novel method to integrate such atmospheric prior into a data-driven approach, resulting in better performance.

With the availability of more urban big data that can be used for air quality prediction, data-driven models have gained increasing attention. Some early studies consider Gaussian processes as a nonparametric method to predict the average pollution level [6,10]. A semi-supervised method is used to make PM2.5 inference based on an PM2.5 affinity graph structure [7]. Another semi-supervised method focuses on the spatial correlation between a target area and

its top-k nearest neighbors [4]. Multi-task learning based strategies are also used to incorporate spatio-temporal smoothness [22].

More recent research addresses the air quality prediction problem by deep learning techniques. Modeling the temporal and/or spatial correlation is key to air quality prediction because air pollutant dispersion is inherently a spatio-temporal process. A simple idea is to directly aggregate the air quality readings, spatial features (e.g., POIs, road networks) and meteorological data from neighboring stations to improve accuracy [23]. More advanced spatial partition and aggregation methods are also introduced to better model spatial correlation [20,24]. Attention mechanisms are another popular way of capturing spatial correlations. Cheng *et al.* [5] introduce an attention mechanism to learn the contributions of different monitoring stations to a target station's PM2.5 value. Liang *et al.* [13] further learn different attentive weights for different stations at different time steps while considering the geospatial similarities between stations. In a slightly different application, Wilson *et al.* [19] propose to capture the higher-order spatial correlations of monitoring stations by graph convolution operations. In contrast, our paper considers a novel type of dispersion-driven dynamic spatial correlation that betters prediction accuracy.

As to temporal correlation, RNNs have been a widely-used choice. For example, Li *et al.* [12] employ a stacked long short-term memory (LSTM) network to extract features from historical air quality data and other auxiliary data. To support long-term air quality prediction, encoder-decoder networks are used to model the non-linear temporal evolution [14,21].

There are also some very recent studies [8,16] that address the air quality prediction problem by considering social media information (e.g., tweets) as an auxiliary data source. Our contributions are orthogonal to them and can be used to further improve their performance.

3 Problem Formulation

Similar to previous studies [5,21], we consider the problem of air quality prediction based on multi-source heterogeneous data. We brief the data sources below.

Air Quality Data. It contains hourly readings of multiple pollutants (e.g., PM2.5, PM10, O_3, NO_2, CO, SO_2, etc.) from each air quality monitoring station $s_i \in \mathcal{S}$, where $\mathcal{S}$ is the entire set of stations under consideration. We denote all stations' air quality data by $\mathcal{M}$.

Weather Data. The weather data of a station s_i at time t is denoted by $\mathbf{w}_i^t$. It contains multiple weather attributes, such as temperature, humidity, wind speed and wind direction. We consider both historical weather data of all stations, denoted by $\mathcal{W}$, and forecast weather data, denoted by $\overline{\mathcal{W}}$.

Geospatial Topology Data. We consider the geospatial topology of all stations, which is denoted by $\mathcal{T}$. It contains the latitude and longitude of each station, and thus allows to calculate the distance and direction (i.e., bearing)

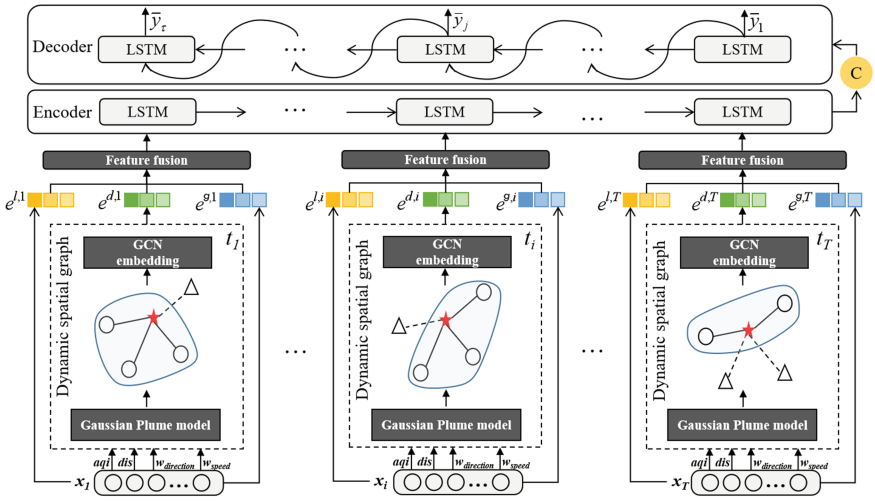


Fig. 1. Architecture overview of the proposed model

between two stations. The geospatial topology data itself is static, but we combine it with the above data sources to compute the stations' dynamic spatial correlation at each time step.

Geo-Context Data. The geo-context data $\mathbf{c}_i \in \mathcal{C}$ of station s_i includes information about road networks and point of interests (POIs) extracted from s_i's affecting area (i.e., the area surrounding s_i). Note that this type of data does not change over time.

Now we are ready to present the problem definition.

Problem Definition. Consider a target station $s_i \in \mathcal{S}$, a historical time window T, and a forecast time window γ. Given all stations' air quality data $\mathcal{M} = \{\mathcal{M}^t\}_{t=1}^{T}$, historical weather data $\mathcal{W} = \{\mathcal{W}^t\}_{t=1}^{T}$, forecast weather data $\overline{\mathcal{W}} = \{\overline{\mathcal{W}}^t\}_{t=T+1}^{T+\gamma}$, geospatial topology data $\mathcal{T}$, and geo-context data $\mathcal{C}$, the goal is to predict the PM2.5 values of station s_i in the next γ hours, denoted by $\hat{\mathbf{y}} = (\hat{y}^{T+1}, \hat{y}^{T+2}, \cdots, \hat{y}^{T+\gamma})$. That is, we aim to learn a prediction function f such that

$$\hat{\mathbf{y}} = f(\mathcal{M}, \mathcal{W}, \overline{\mathcal{W}}, \mathcal{T}, \mathcal{C}, \Theta), \tag{1}$$

where Θ denotes the set of parameters of f to learn.

4 Proposed Method

In this section, we elaborate our proposed method that makes use of atmospheric dispersion theories to model the dynamic spatial correlations among monitoring stations in order to improve air quality prediction. The overall architecture of the proposed method is illustrated in Fig. 1.

4.1 Feature Representation

To predict the PM2.5 values of station s_i in the next γ hours, we construct three types of features as explained below.

Local Features. This set of features includes station s_i's air quality data, historical and forecast weather data, and geo-context data. Air quality data and weather data are observed or forecasted each hour, and naturally form time series. Thus, they lay the foundation for temporal correlation modeling.

Global Features. This set of features includes the local features of all nearby stations, defined by the Euclidean distance. While the air quality data and weather data of the neighboring stations change over time, global features fail to capture the dynamic spatial influence of neighboring stations on the target station s_i due to air dispersion, which is critical to achieve better prediction performance.

Dynamic Spatial Features. This is a set of novel features driven by atmospheric dispersion. Guided by atmospheric dispersion models, dynamic spatial features explicitly measure the spatial influence of neighboring stations by considering multiple external factors at each time step. We detail how to generate dynamic spatial features in the next section.

4.2 Dynamic Spatial Graph Construction

At time step t, we represent the dispersion-driven dynamic spatial influence of all other stations on a target station s_i by a weighted directed graph $\mathcal{G}_i^t = (\mathcal{S}, \mathcal{E}_i^t)$, where an edge $e_{ij} \in \mathcal{E}_i^t$ gives the dynamic spatial influence of station s_j on the target station s_i as per atmospheric dispersion modeling. Note that normally the dynamic spatial influence of station s_i on station s_j is different from that of s_j on s_i as shown in Fig. 2. In the following, we omit the superscript t as

(a) time t (b) time $t + j$

Fig. 2. Illustration of the dynamic spatial influence. s_1, s_2 and s_3 are the associated stations of the star as the northwest wind in time t, and changed to s_1 and s_5 in time $t + j$ as the northeast wind

all discussions are for time t. Previous studies normally consider the geographic proximity between stations as the key factor to model their spatial correlations. For example, the top-k nearest neighbors' features are used to predict a target station's PM2.5 value. Wilson *et al.* [19] propose to use a graph structure to explicitly model the spatial correlations. All elements in the adjacency matrix of the spatial graph are model parameters that need to be learned from training data. However, the adjacency matrix is assumed to be *static*, that is, it is fixed at different time steps. This assumption directly violates the well-established dispersion models. In addition, considering all elements in the adjacency matrix as trainable parameters substantially increases the model complexity. Therefore, Wilson *et al.* [19] further assume that the adjacency matrix is either sparse or low rank to mitigate the number of parameters. However, this assumption is not backed up by any theoretical ground. To this end, we propose a novel domain knowledge driven method that not only allows to dynamically learn a different adjacency matrix at each time step, but also fully uses atmospheric dispersion theories to mitigate model complexity.

The first step is to select an appropriate atmospheric dispersion model that can be seamlessly integrated into a data-driven approach. Eulerian and Lagrangian models are used to predict air pollution in urban areas, which assume pollutants to be evenly distributed within the boundary [3]. Computational fluid dynamic (CFD) models are used to better understand fluid dispersion, but can also be used in urban air quality prediction [9]. The Gaussian plume model is one of the most widely-used models to assess the impacts of emission sources on local and urban air quality [2]. The dispersion of pollutants can be described in both horizontal and vertical directions by a Gaussian distribution, which well suits our setting. As such, we choose the Gaussian plume model as the domain model for modeling dispersion-driven dynamic spatial correlation. The spatial dynamics of pollutant dispersion in a Gaussian model can be described by the following equation [15]:

$$c(r,s) = \frac{Q}{2\pi\sigma_y\sigma_z\bar{u}} \exp\left(-\frac{1}{2}\left(\frac{Y}{\sigma_y}\right)^2\right) \exp\left(-\frac{1}{2}\left(\frac{h_e - z_r}{\sigma_z}\right)^2\right), \tag{2}$$

where $c(r,s)$ is the concentration at point $r = (x_r, y_r, z_r)$ due to the emissions at point $s = (x_s, y_s, z_s)$, Q is the emission rate, Y is the crosswind distance between r and s, σ_y and σ_z are the Gaussian plume dispersion parameters, which are a function of the downwind distance X, $\bar{u}$ is the average horizontal wind speed, and h_e is the effective emission height (i.e., $h_e = z_s + \Delta h$, and Δh is the emission plume rise, which is a function of emission parameters and meteorological conditions).

Based on the available data (see the experiment section for more details), we adapt Eq. (2) as follows. First, since emission sources are unavailable in the dataset, we consider other stations as the second-hand pollutant sources for the target station [20]. Second, since all stations in the data are *point sources* (i.e., without elevation information), we ignore all items related to height in Eq. (2). Third, we propose to use a data-driven method to learn a function $\phi(X)$

to determine σ_y. Then the dispersion-driven spatial influence of station s_j on target station s_i at time t can be formulated as:

$$c(s_i, s_j) = \frac{Q_j}{2\pi\phi(X)\bar{u}_j} \exp\left(-\frac{1}{2}\left(\frac{Y}{\phi(X)}\right)^2\right), \tag{3}$$

where Q_j is the air quality of station s_j , $\bar{u}_j$ is the horizontal wind velocity at s_j, and X and Y are the downwind and crosswind distances between s_i and s_j, respectively. Here we model $\phi(\cdot)$ as a linear function, that is, $\phi(X) = \gamma X$, where γ is a learnable scalar. We can also model $\phi(\cdot)$ as a more complicated function that can be learned by a multi-layer perceptron (MLP). But our experiments indicate that a linear formulation already strikes a reasonable trade-off between performance and model complexity.

We further normalize the influence of station s_j on target station s_i among all other stations:

$$a_{ij} = \frac{c(s_i, s_j)}{\sum_{s_k \in (\mathcal{S} - s_i)} c(s_i, s_k)}. \tag{4}$$

a_{ij} is the weight of the edge e_{ij}. All a_{ij} values form the adjacency matrix $\mathbf{A}_i$ of the dynamic spatial graph $\mathcal{G}_i$ for target station s_i.

It can be seen that with the help of atmospheric dispersion theories, we successfully reduce the number of learnable parameters of a dynamic spatial graph from $O(|\mathcal{S}|^2)$, where $|\mathcal{S}|$ is the number of stations, to $O(1)$. Note that the learnable parameter γ is shared among all time steps.

4.3 Dynamic Spatial Graph Embedding

After constructing the dynamic spatial graph $\mathcal{G}_i$ for target station s_i at time t, we need a way to convert $\mathcal{G}_i$ into a low dimensional space so that its spatial information can be effectively fused with the temporal correlation in an encoder-decoder network. We omit the subscript i when it is clear from the context. We consider graph convolutional networks (GCNs) [11] for this purpose due to its flexibility and good performance. For a K-layer graph convolutional network, the output of the l-th layer can be represented as $\mathbf{H}^l \in \mathbb{R}^{|\mathcal{S}| \times d^{(l)}}$. Each row in $\mathbf{H}^l$ represents the embedding of a station whose dimension is $d^{(l)}$. The embedding of a station after the $(l+1)$-th layer will be computed as the aggregation of its connected stations' embeddings from the l-th layer. This operation performed in a GCN layer can be formulated as:

$$\mathbf{H}^{(l+1)} = \sigma(\mathbf{A}\mathbf{H}^{(l)}\mathbf{W}^{(l)}), \tag{5}$$

where $\sigma(\cdot)$ is a non-linear activation function, $\mathbf{A}$ is the adjacency matrix of the dynamic spatial graph $\mathcal{G}$, and $\mathbf{W}^{(l)} \in \mathbb{R}^{d^{(l)} \times d^{(l+1)}}$ is a layer-specific trainable transformation matrix for the l-th layer. The target station s_i's embedding in $\mathbf{H}^K$ is used as part of the input to encoder-decoder network.

4.4 Encoder-Decoder Based Spatio-Temporal Fusion

To support long-term air quality prediction, we use an encoder-decoder LSTM model [18] to fuse spatial and temporal information and infer future PM2.5 values. Let $\mathbf{e}_i^{l,t}$ and $\mathbf{e}_i^{g,t}$ denote station s_i's local feature embeddings and global feature embeddings at time t, where $\mathbf{e}_i^{g,t} = \sum_{s_j \in (\mathcal{S}-s_i)} \mathbf{e}_j^{l,t}$. Let $\mathbf{e}_i^{d,t}$ denote the embedding learned from the GCN over the dynamic spatial graph $\mathcal{G}_i^t$. We concatenate $\mathbf{e}_i^{l,t}$, $\mathbf{e}_i^{g,t}$ and $\mathbf{e}_i^{d,t}$, and feed it into the LSTM cell for time t in the encoder part. For ease of presentation, we drop all the subscripts. Then the hidden state h^t can be learned by

$$h^t = \text{LSTM}(\mathbf{e}^{l,t} \parallel \mathbf{e}^{g,t} \parallel \mathbf{e}^{d,t}, h^{t-1}), \tag{6}$$

where $\parallel$ means the concatenation operation, and h^{t-1} is the hidden state at time $t-1$. The resultant hidden state h^t is regarded as the latent representation of the air quality status of s_i at time t.

The last hidden state h^T produced from the encoder part encapsulates the information of all historical data and serves as the initial hidden state of the decoder. The input to an LSTM cell for time t in the decoder consists of the forecast weather data at station s_i, denoted by $\bar{\mathbf{w}}^t$, and the predicted PM2.5 value of station s_i at time $t-1$, denoted by $\hat{y}^{t-1}$. Similarly, we concatenate $\bar{\mathbf{w}}^t$ and $\hat{y}^{t-1}$, and calculate the hidden state h^t at time t as

$$h^t = \text{LSTM}(\bar{\mathbf{w}}^t \parallel \hat{y}^{t-1}, h^{t-1}). \tag{7}$$

4.5 Model Learning

Since we are tasked with a regression problem, we employ the mean squared error (MSE) as the objective function, which measures the average of squared distances between predicted PM2.5 values and the actual ones. We apply L_2 regularization to mitigate overfitting. Formally, the objective function $\mathcal{L}$ we optimize is:

$$\mathcal{L} = \frac{1}{M} \sum_{i=1}^{M} (y_i - f(\mathbf{x}_i, \Theta))^2 + \lambda \left\| \Theta \right\|_2, \tag{8}$$

where M is the number of training instances, $\mathbf{x}_i$ is a training instance, Θ is the set of trainable parameters in our proposed model, and λ is the regularization parameter. Early stopping is also used to reduce overfitting.

Recall that $\Theta = \{\Theta_1, \Theta_2\}$ consists of two subsets of parameters, where $\Theta_1 = \{\gamma, \{\mathbf{W}^{(l)}\}_{l=1}^K\}$ includes the parameters to learn the embeddings from a dynamic spatial graph, and Θ_2 includes the parameters of forget gates, input gates, and output gates in the encoder-decoder LSTM network.

5 Experiments

In this section, we conduct a comprehensive experimental study to demonstrate that our proposed method outperforms the state-of-the-art competitors. In addition, we provide a case study to intuitively show the benefits of dispersion-driven dynamic spatial modeling.

5.1 Datasets

We utilize the following real datasets in the experiments, which are commonly used in extensive literature.

Air Quality Data. We collect air quality data, including AQI, PM2.5, PM10, O_3, NO_2, CO, SO_2, from all 35 ground-based air quality monitoring stations in Beijing.[1] Since PM2.5 is the major pollutant widely used by government agencies for public communication, we predict the PM2.5 values in the experiments. We use linear interpolation to fill in missing values that occur within 3 h. Continuous missing data spanning over 3 h are discarded [21].

Meteorological Data. Following the previous study [21], we consider grid-based weather data obtained from the Global Data Assimilation System (GDAS).[2] The spatial resolution of the grid data is 0.25°. We extract the region with latitudes between 39.5° and 40.75° and longitudes between 115.75° and 117.25°, which covers all the monitoring stations in Beijing. We select five weather attributes: temperature, humidity, wind speed, and wind directions (including wind-u and wind-v in GDAS). As suggested in [21], we conduct a temporal linear interpolation to convert the 3-hourly raw data to hourly data.

POIs. POI types and density in a region directly affect its air quality. Similar to [23], we consider 12 types of POIs from Amap of Beijing,[3] and compute the number of POIs in each category within the affecting region of a station as a feature.

Road Networks. We download the road network data of Beijing from OpenStreetMap (OSM).[4] There are five types of roads, namely primary road, secondary road, tertiary road, residential road and footway road. Similarly, we calculate the number of each type of roads as a feature.

In addition, similar to the previous study [21], we extract 3 time features, including hour of day, day of week, and month, from the timestamp of each data point.

5.2 Experimental Settings

We process air quality data and meteorological data from January 1st, 2016 to January 31st, 2018, together with POI and road network data. The portions of training, validation, and test data are split by the ratio 8:1:1. In particular, training data and test data are split in temporal order in order to avoid data leakage. The historical time window T is set to 48, and we aim to predict the PM2.5 values in the next 24 h. We use 64 hidden units (i.e., the dimension of a hidden state) in an LSTM cell for feature representation, and optimize the objective function using the Adam

[1] http://beijingair.sinaapp.com.
[2] https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-data-assimilation-system-gdas.
[3] https://lbs.amap.com/api/webservice/download.
[4] https://www.openstreetmap.org/.

Table 1. Performance comparisons of different models

	1–6 h		7–12 h		13–18 h		19–24 h	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Naive approach	14.87	26.33	26.00	43.16	32.21	50.70	35.45	54.79
LSTM	14.17	20.91	25.88	33.83	32.67	40.23	37.03	44.08
Seq2seq	14.13	21.39	23.99	32.59	30.14	38.55	33.61	41.89
DeepAir [20]	19.18	25.15	23.13	29.64	25.20	31.88	28.43	35.37
GeoMAN [13]	14.03	19.10	19.42	25.06	22.95	29.31	24.23	32.14
WGC-LSTM [19]	12.78	18.24	18.05	23.59	18.92	28.00	25.42	29.74
MGED-Net [21]	13.44	17.35	18.05	22.83	20.95	26.01	21.91	26.88
Our method	**10.82**	**15.71**	**16.54**	**21.10**	**17.52**	**24.54**	**19.13**	**24.91**

Table 2. Performance comparison of different spatial correlation modeling methods

	1–6 h		7–12 h		13–18 h		19–24 h	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Without dynamic spatial	15.52	23.53	20.02	25.81	21.13	30.71	24.80	32.66
Fixed **A**	15.30	19.99	19.63	24.71	22.03	29.97	25.07	32.12
Shared **A** [19]	12.78	18.24	18.05	23.59	18.92	28.00	25.42	29.74
Our method	**10.82**	**15.71**	**16.54**	**21.10**	**17.52**	**24.54**	**19.13**	**24.91**

optimizer with learning rate 0.001. To address overfitting, we use L_2 regularization with the regularization coefficient of 0.0001, and employ early stopping according to the validation error. Our code is implemented in PyTorch.

5.3 Compared Methods

We compare our proposed model with a wide range of representative approaches described below.

- **Naive approach** uses the PM2.5 value of the current time step as the predicted values for all future hours.
- **LSTM** uses a typical LSTM model to predict the 24 h' PM2.5 values.
- **Seq2seq** is an encoder-decoder network with stacked LSTMs in both encoder and decoder.
- **DeepAir** [20] is a distributed fusion network, which consists of 5 subnets powered by a FusionNet structure. Then, these subnets are merged to generate prediction results according to their weights.
- **GeoMAN** [13] is based on an encoder-decoder architecture with a multi-level attention mechanism. External factors are fused with the output of the encoder as the input to the decoder.
- **WGC-LSTM** [19] is a weighted graph convolutional LSTM network, which considers the adjacency matrix of the spatial graph as model parameters. The adjacency matrix is static and shared among all time steps.

– **MGED-Net** [21] is a multi-group encoder-decoder network with multiple encoders and a single decoder. All features are divided into different groups by correlation and merged by the encoder fusion strategy.

5.4 Experimental Results

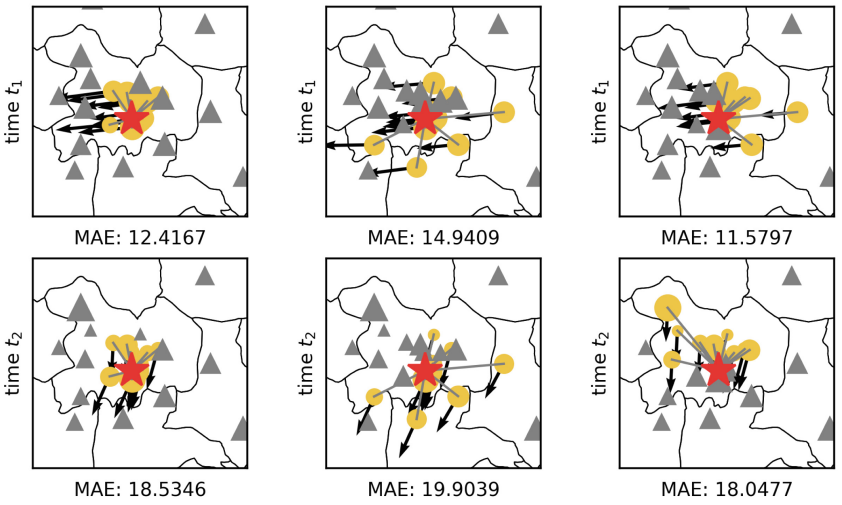
Following the previous studies [13,21], we use two widely-used evaluation metrics, *root mean squared error* (RMSE) and *mean absolute error* (MAE), to measure the performance of different prediction models. Similarly, we report the prediction results in four time intervals (1–6 h, 7–12 h, 13–18 h, and 19–24 h).

We report the main experimental results in Table 1. Among all models, our proposed model obtains the best results in all four time intervals on both metrics. Specifically, our method shows 12.6% to 19.4% improvement and 5.6% to 9.4% improvement over the state-of-the-art approach MGED-Net on MAE and RMSE, respectively. Compared to LSTM, encoder-decoder-based methods (i.e., Seq2seq, GeoMAN, MGED-Net and our model) achieve significant improvements in long-term predictions due to the decoder component. This justifies the adoption of an encoder-decoder architecture in our method to model the long-term temporal evolution. Moreover, it can be seen that our method's short-term prediction performance (e.g., 1–6 h and 7–12 h) is also much better than that of WGC-LSTM. We deem that it is due to the dispersion-driven dynamic spatial correlation modeling. In contrast, modeling the spatial influence by a static adjacency matrix in WGC-LSTM does not reflect the real air pollutant dispersion process well, and thus leads to less desirable prediction performance. In the following sections, we provide more experiments to study the effects of different spatial correlation modeling methods.

Benefits of Dynamic Spatial Graph. To demonstrate the benefits of modeling dynamic spatial influence with atmospheric prior, we conduct a set of experiments with different methods of modeling spatial correlation.

– **Without dynamic spatial** is a variant of our method that removes all dynamic spatial features. The rest is the same as the proposed method.
– **Fixed A** considers the geographic proximity (e.g., the Euclidean distance) between stations as edge weights of the spatial graph, which is set in advance before the training.
– **Shared A** is essentially the method in [19], where the elements in the adjacency matrix of the spatial graph are considered as learnable parameters. Note that the adjacency matrix here is shared among all time steps.

Table 2 shows the performance of different spatial correlation modeling methods. We can draw a few important observations. First, explicitly modeling the spatial correlations among monitoring stations, even only considering their Euclidean distance, is beneficial. Second, the spatial influence among different stations is indeed not simply determined by their geographic proximity. This explains why Fixed **A**'s performance is much worse than those of Shared **A** and



(a) Fixed **A** (b) Shared **A** (c) Our method

Fig. 3. Visualization of the spatial influence of different spatial correlation modeling methods on a target station

our method. Third, modeling dynamic spatial influence using well-established atmospheric prior is rewarding. It not only achieves much better performance, but also leads to a less complex model that is easier to train.

Dynamic Spatial Graph Visualization. Finally, to better understand how the dynamic spatial graph helps improve air quality prediction accuracy, we visualize the spatial influence of different stations on a target station (marked as a red star) at two representative time steps t_1 and t_2 in Fig. 3. The yellow dots represent the top-10 stations that have the most spatial influence on the target station. The size of a dot represents its pollution level. The larger a dot, the higher its PM2.5 value. The direction of an arrow indicates the wind direction at a station, and the length indicates the wind speed. The grey triangles denote other stations. Similarly, their sizes represent their air pollution level.

We have a few interesting observations. First, the most influential stations of our method at different time steps well align with the Gaussian plume model and one's intuition. The most influential stations at different time steps for a target station are also different, which are determined by multiple factors defined by the Gaussian plume model, such as the geographic distance, air quality, and meteorological conditions. This reflects the *dynamic* nature of the spatial correlation modeling in our method. Second, the most influential stations in both Fixed **A** and Shared **A** are fixed over time. For Fixed **A**, it is because the geographic proximity among stations does not change over time; for Shared **A**, it is due to the fact that the same adjacency matrix is shared among all time steps.

In particular, the most influential stations of Shared **A** are counter-intuitive. Third, while in general Shared **A** performs better than Fixed **A**, at time steps t_1 and t_2, its MAE values are worse than those of Fixed **A**. This is not difficult to understand—the adjacency matrix learned by minimizing the average error over all time steps cannot guarantee reasonable performance at every time step.

6 Conclusion and Future Work

In this paper, we took on a new perspective of air quality prediction, which models dynamic spatial influence among monitoring stations guided by atmospheric dispersion modeling. We proposed to construct a dynamic spatial graph based on the Gaussian plume model, generate graph embeddings by a GCN, and finally fuse spatial and temporal information seamlessly in an encoder-decoder LSTM network. Experiments on real-world benchmark datasets validate the superiority of the proposed model. In addition, we provided a case study to intuitively understand the benefits of dynamic spatial correlation modeling. In future work, we will investigate other possible factors to improve dynamic spatial correlation modeling, and explore more advanced prediction models to improve prediction accuracy (e.g., stacked LSTMs).

References

1. Abdel-Rahman, A.A.: On the dispersion models and atmospheric dispersion. Int. J. Glob. Warming **3**(4), 257–273 (2011)
2. Arystanbekova, N.K.: Application of gaussian plume models for air pollution simulation at instantaneous emissions. Math. Comput. Simul. **67**(4), 451–458 (2004)
3. Bergin, M.S., Noblet, G.S., Petrini, K., Dhieux, J.R., Milford, J.B., Harley, R.A.: Formal uncertainty analysis of a Lagrangian photochemical air pollution model. Environ. Sci. Technol. **33**(7), 1116–1126 (1999)
4. Chen, L., Cai, Y., Ding, Y., Lv, M., Yuan, C., Chen, G.: Spatially fine-grained urban air quality estimation using ensemble semi-supervised learning and pruning. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp), pp. 1076–1087 (2016)
5. Cheng, W., Shen, Y., Zhu, Y., Huang, L.: A neural attention model for urban air quality inference: learning the weights of monitoring stations. In: Proceedings of the 32th AAAI Conference on Artificial Intelligence (AAAI), pp. 2151–2158 (2018)
6. Guizilini, V., Ramos, F.: A nonparametric online model for air quality prediction. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI), pp. 651–657 (2015)
7. Hsieh, H., Lin, S., Zheng, Y.: Inferring air quality for station location recommendation based on urban big data. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 437–446 (2015)
8. Jiang, Y., Sun, X., Wang, W., Young, S.D.: Enhancing air quality prediction with social media and natural language processing. In: Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL), pp. 2627–2632 (2019)

9. Jin, B.J., Bu, P.S., Jin, K.J.: Urban flow and dispersion simulation using a CFD model coupled to a mesoscale model. J. Appl. Meteorol. Climatol. **48**(8), 1667–1681 (2009)
10. Jutzeler, A., Li, J.J., Faltings, B.: A region-based model for estimating urban air pollution. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI), pp. 424–430 (2014)
11. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: Proceedings of the 5th International Conference on Learning Representations (ICLR) (2017)
12. Li, X., et al.: Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation. Environ. Pollut. **231**, 997–1004 (2017)
13. Liang, Y., Ke, S., Zhang, J., Yi, X., Zheng, Y.: GeoMAN: multi-level attention networks for geo-sensory time series prediction. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), pp. 3428–3434 (2018)
14. Luo, Z., Huang, J., Hu, K., Li, X., Zhang, P.: AccuAir: winning solution to air quality prediction for KDD cup 2018. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 1842–1850 (2019)
15. Paolo, Z.: Gaussian models. In: Air Pollution Modeling, pp. 141–183. Springer, Boston (1990). https://doi.org/10.1007/978-1-4757-4465-1_7
16. Pramanik, P., Mondal, T., Nandi, S., Saha, M.: AirCalypse: can Twitter help in urban air quality measurement and who are the influential users? In: Proceedings of the 29th International World Wide Web Conferences (WWW), pp. 540–545 (2020)
17. Rakowska, A., et al.: Impact of traffic volume and composition on the air quality and pedestrian exposure in urban street canyon. Atmos. Environ. **98**, 260–270 (2014)
18. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Proceedings of the 28th Conference on Neural Information Processing Systems (NIPS), pp. 3104–3112 (2014)
19. Wilson, T., Tan, P., Luo, L.: A low rank weighted graph convolutional approach to weather prediction. In: Proceeding of the 18th IEEE International Conference on Data Mining (ICDM), pp. 627–636 (2018)
20. Yi, X., Zhang, J., Wang, Z., Li, T., Zheng, Y.: Deep distributed fusion network for air quality prediction. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 965–973 (2018)
21. Zhang, Y., et al.: Multi-group encoder-decoder networks to fuse heterogeneous data for next-day air quality prediction. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), pp. 4341–4347 (2019)
22. Zhao, X., Xu, T., Fu, Y., Chen, E., Guo, H.: Incorporating spatio-temporal smoothness for air quality inference. In: Proceeding of the 17th IEEE International Conference on Data Mining (ICDM), pp. 1177–1182 (2017)
23. Zheng, Y., Liu, F., Hsieh, H.: U-air: when urban air quality inference meets big data. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 1436–1444 (2013)
24. Zheng, Y., et al.: Forecasting fine-grained air quality based on big data. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 2267–2276 (2015)