A Novel Interpretable Deep Learning Model for Ozone Prediction

Xingguo Chen 1,†, Yang Li 1,†, Xiaoyan Xu 2 and Min Shao 3,*

1 Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; chenxg@njupt.edu.cn (X.C.); liyang211014@gmail.com (Y.L.)
2 School of the Environment, Nanjing University, Nanjing 210046, China; mf20250130@mail.nju.edu.cn
3 School of Environment, Nanjing Normal University, Nanjing 210023, China
* Correspondence: mshao@masonlive.gmu.edu; Tel.: +86-1529-538-6000
† These authors contributed equally to this work.

Abstract: Due to the limited understanding of the physical and chemical processes involved in ozone formation, as well as the large uncertainties surrounding its precursors, commonly used methods often result in biased predictions. Deep learning, as a powerful tool for fitting data, offers an alternative approach. However, most deep learning-based ozone-prediction models only take into account temporality and have limited capacity. Existing spatiotemporal deep learning models generally suffer from model complexity and inadequate spatiality learning. Thus, we propose a novel spatiotemporal model, namely the Spatiotemporal Attentive Gated Recurrent Unit (STAGRU). STAGRU uses a double attention mechanism, which includes temporal and spatial attention layers. It takes historical sequences from a target monitoring station and its neighboring stations as input to capture temporal and spatial information, respectively. This approach enables the achievement of more accurate results. The novel model was evaluated by comparing it to ozone observations in five major cities, Nanjing, Chengdu, Beijing, Guangzhou and Wuhan. All of these cities experience severe ozone pollution. The comparison involved Seq2Seq models, Seq2Seq+Attention models and our models. The experimental results show that our algorithm performs 14% better than Seq2Seq models and 4% better than Seq2Seq+Attention models. We also discuss the interpretability of our method, which reveals that temporality involves short-term dependency and long-term periodicity, while spatiality is mainly reflected in the transportation of ozone with the wind. This study emphasizes the significant impact of transportation on the implementation of ozone-pollution-control measures by the Chinese government.

Keywords: recurrent neural network; O₃ forecasting; attention mechanism; spatiotemporal information

1. Introduction

Ozone (O₃) is formed through photochemical reactions of compounds (VOCs) and NOₓ [1]. This disparity in spatiotemporal distribution is primarily caused by variations in emission characteristics, synoptic conditions, topographic distribution and land-use types [2–6]. Generally, the emission source and meteorological characteristics are the fundamental and essential factors for the formation, transport and dispersion of O₃ [7]. The source of O₃ precursor emissions can be categorized as anthropogenic sources and natural sources [8]. Additionally, meteorological variables such as solar radiation, wind direction, wind speed, atmospheric pressure, temperature and relative humidity have a complex relationship with O₃ [9–13]. Recently, ozone pollution has gradually increased and become a primary pollutant of great concern in air pollution control [14–16] due to its detrimental impacts on both human health and agriculture [17–20]. Considering the complexity of the O₃ formation mechanism, the exacerbated combined atmospheric pollution increases the difficulty in ozone control [21]. One of the most important tasks for assessing efficient prevention and control strategies for O₃ pollution is predicting O₃.
Building precise O₃ prediction models can strongly support decision-makers in efficiently reducing heavy ozone pollution peaks, which is an urgent and necessary task.

Generally, ozone-prediction approaches can be classified into two types: numerical and statistical approaches. Numerical approaches simulate the real atmospheric environment by utilizing accurate estimations of anthropogenic emissions and incorporating specific atmospheric physics and chemistry reactions. Some numerical approaches have been widely used in ozone prediction. Unfortunately, numerical models suffer from an imperfect understanding of complex ozone formation and thus sacrifice spatiotemporal resolution. Therefore, the spatiotemporal representativeness, emission and modeling mechanisms still need to be perfected. In contrast, statistical models do not take into account the complicated reaction mechanisms. However, they offer greater flexibility and more computing advantages. Classical statistical ozone-prediction models mainly consist of basic regression models, which can limit the model’s capacity to describe non-linear and complex internal physicochemical processes. Therefore, they often fail to meet practical requirements. Machine learning (ML) methods, as a promising approach, have inspired advancements in the field of ozone forecasting. Basic ML algorithms, advanced ensemble algorithms and artificial neural network (ANN) have been intensively studied in ozone forecasting. Nevertheless, these models fail to capture spatial and temporal information simultaneously and normally ignore the interaction between elements in the sequence due to their independent and identically distributed (i.i.d.) premise. Therefore, more powerful methods are needed.

To capture spatiotemporal information, emerging deep learning (DL) models are a good choice because of their powerful representative capabilities. Theoretically, a deep learning neural network (DNN) is capable of fitting any form of function. However, training such a deep neural network can be extremely challenging. Considering the No Free Lunch theorem, the majority of sequence-oriented neural networks have been implemented using the recurrent neural network (RNN). RNNs are specifically designed for sequence forecasting by learning temporal patterns. Nonetheless, RNNs have issues with gradient explosion and gradient vanishing and they also lack long-term memory. Accordingly, the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) were proposed to address these issues by incorporating memory units and a gating mechanism. The LSTM and GRU are commonly used in ozone prediction in conjunction with the Encoder–Decoder framework. However, the performance of the Encoder–Decoder is limited by the fixed length of the hidden state. The presence of an attention mechanism breaks the bottleneck of the Encoder–Decoder. With the attention mechanism, certain methods can achieve more accurate results. Besides temporal features, spatial factors are also crucial. Ozone pollution is typically a regional air quality concern. Therefore, it is influenced not only by local emissions and meteorological conditions but also by the long-range transportation of ozone and its precursors. Several deep learning-based networks are widely used to capture spatial information. The CNN can learn shift-invariant features of data because the kernel remains constant during convolution operations. Some researchers have utilized CNN models to analyze images and forecast air pollutant concentrations. However, it is the shared parameters that make the CNN take all monitoring stations as identical within a single convolution. The attention mechanism serves the purpose of learning spatial relationships more directly by assigning different weights to each site. Some methods have been investigated, such as, that learn spatiotemporal information from a pre-organized feature matrix. It then applies LSTM and Multilayer Perceptron (MLP) to assign weights in the temporal and spatial dimensions to forecast PM₂.₅. Spatiality is an inherent characteristic of graph neural networks, which treats each monitoring station as a node in an undirected graph, with edges representing connections between the nodes. The value of an edge indicates the strength of the connection. Using GCNs to forecast air pollutants requires performing convolution in the spectral space, which demands substantial computational power. Furthermore,
the initial issue when predicting ozone with GCNs is defining the relationship between two randomly selected monitoring stations.

In this paper, we propose a novel method called the Spatiotemporal Attentive Gated Recurrent Unit (STAGRU) based on the Seq2Seq model and the attention mechanism. Our method aims to predict local ozone concentration in five mega cities across China suffering severe ozone pollution. We utilize spatiotemporal information obtained from in situ observations to improve the precision of our predictions. STAGRU applies two types of attention mechanisms. One captures temporal knowledge from historical data using an RNN-based network, while the other selects a significant moment from the historical data in the surrounding sites for the current prediction. Due to its ability to learn directly from historical data, the proposed model has a lower computational burden and higher interpretability. Details regarding model construction and the data used for training and experimental comparisons are presented in Section 2. Subsequently, we conducted experiments to compare STAGRU with Seq2Seq and Seq2Seq+Attention-based models. Finally, we will discuss the interpretability of STAGRU, which offers us insights into temporality and spatiality. Furthermore, a derivative model called STAGRU-Decoder is proposed, which predicts ozone concentration for multiple stations simultaneously.

2. Materials and Methods

2.1. Methods

The attention mechanism is commonly used in various domains and is effective in solving sequence-prediction problems. To learn temporality, we incorporated the attention mechanism to capture the temporal information from the previous sequence of the target monitoring station. Specifically, the past sequence is sent to the temporal attention layer and the corresponding attention weight vector quantifying the relation between each past moment and current prediction is obtained. After computing the dot product of the past sequence and the attention weight vector, temporal information, namely the temporal context vector, is derived. Intrinsically, by using temporal attention, we established a link between the present prediction and every previous moment. With this connection, the knowledge from previous sequences can assist in forecasting. This inspired us to employ the idea of using the attention mechanism to construct bridges that facilitate the flow of information between multiple stations in spatial learning.

When predicting ozone for a certain station, we believe that its surrounding stations can provide assistance to improve prediction accuracy, because ozone pollution is usually a regional air quality issue. Thus, we learned spatial information from some specific moments in the past sequence of each surrounding station. Here, we propose to employ particular moments, not the whole sequence as it can reduce computation and alleviate disturbances from unimportant moments. We utilized the temporal attention layer to select the past moment with the highest attention weight for each station and introduced another attention layer, the spatial attention layer, to calculate the spatial context vector. The reason we used the temporal attention layer to make selections is that this layer contains knowledge about how to evaluate the importance of a past moment compared to the current prediction concerning the target station. Thus, the attention layer with learnable parameters is recommended. After determining these specific moments, the spatial attention layer calculates the attention score using decoding hidden states and the geographical location of that neighboring monitoring station for each selected moment to obtain the spatial context vector. In this manner, the spatial attention layer builds a connection between the target station and its neighbors and the knowledge of the past sequences of each station eventually affects the prediction.

Along with learning the temporal and spatial information, the spatiotemporal attention-based STAGRU (Spatiotemporal Attentive Gated Recurrent Unit) model was proposed. Details of the STAGRU are shown in Figure 1. As shown in Figure 1, there are two modeling parts in the STAGRU. Historical air quality data and meteorological data of the target station are fed into the Encoder and encoded into the hidden states. A temporal context
vector is derived from those hidden states by the temporal attention layer when decoding. For spatial modeling, the decoding hidden states with the highest attention weight are selected to learn the spatial knowledge and obtain the spatial context vector. After this, the temporal and spatial context vectors are concatenated with the current decoding hidden state and sent to the DNN to make a prediction.

Figure 1. The model structure of the Spatiotemporal Attentive Gated Recurrent Unit (STAGRU).

2.2. Data

2.2.1. Study Regions and Datasets

Ozone pollution is generally a regional issue, so we selected nine monitoring stations in several cities (Nanjing, Beijing, Chengdu, Guangzhou and Wuhan) that suffer from severe ozone pollution to represent the corresponding regions. The data used in our experiments comprised the air quality (including AQI, PM$_{2.5}$, PM$_{10}$, SO$_2$, NO$_2$, O$_3$ and CO) and meteorological (including radial wind, temperature, relative humidity and Residual Boundary Layer Height) data of those monitoring stations from January 2015 to December 2021. The longitude and latitude of each monitoring station in each city are shown in Table A1 and the geographical distribution of these monitoring stations is shown in Figure 2. These stations were distributed in different regions of their city, like the teaching area, downtown area, industrial area, etc.

The air quality data included the hourly PM$_{2.5}$, PM$_{10}$, NO$_2$, CO and SO$_2$ concentrations and the air quality index (AQI). The meteorological information included ERA5 near-surface wind speed, wind direction, temperature, relative humidity and planetary boundary layer height. In each monitoring station, we took the data from January 2015 to December 2020 as the training data and January 2021 to December 2021 as the testing data, which was used to evaluate the performance of the final trained model. Linear interpolation [58] was used to fill in the missing data, because linear interpolation is simple and able to mostly reserve statistical features of original data [59]. Normalization was applied to the air quality, wind speed and temperature data to achieve a faster convergence. We transformed the raw dataset into a supervised dataset, consisting of input–output pairs based on a sliding window. For details of the data flow, we took the past 24 h air quality
and meteorological data from all stations in a city as the input. Subsequently, 24 h O$_3$
forecasts of the target station were obtained via the STAGRU.

![Geographical topography and location of the studied cities and the distribution of the monitoring stations in each city.](image)

**Figure 2.** Geographical topography and location of the studied cities and the distribution of the monitoring stations in each city.

### 2.2.2. Evaluation Metrics

In this study, the Root Mean Squared Error (RMSE), $R^2$ and the Symmetric Mean Absolute Percentage Error (SMAPE) were used as the performance metrics. The formula of the RMSE is shown below

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

where $y_i$ represents the observation of item $i$; $\hat{y}_i$ represents the prediction of item $i$; and $n$ represents the number of items. $R^2$ measures the fitting degree between the model and the true state. The formula is defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y} = \sum_{i=1}^{n} y_i$.

The SMAPE measures the accuracy of the predictions based on percentage error, which is defined as follows:

$$SMAPE = \frac{100\% \sum_{i=0}^{n} |\hat{y}_i - y_i|}{n \sum_{i=0}^{n} (|\hat{y}_i| + |y_i|)/2}$$

Note that each evaluation metric above has its advantages and disadvantages; thus, we integrated different approaches to measure the effectiveness of the model.

### 2.2.3. Experimental Design

In order to evaluate the capability of the STAGRU, we compared it with the Seq2Seq- and Seq2Seq+Attention-based models. Further, we also considered adopting LSTM to replace GRU in the STAGRU, producing an STALSTM, because these two RNN methods are frequently used in many time series prediction tasks [60,61]. The details of the experimental design are shown in Table 1. The number of hidden units was 256, the hidden layer was 1, the batch size was 48 and the optimizer was Adam [62] and the learning rate was 0.0001. Early stopping was applied to obtain an acceptable model, and schedule sampling was
also used. Further, all experiments were conducted on an NVIDIA GeForce GTX 1050Ti 4G GPU.

Table 1. Experimental design.

<table>
<thead>
<tr>
<th>Experiment Category</th>
<th>Experiment Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq2Seq</td>
<td>Seq2Seq_LSTM</td>
<td>Encoder–Decoder framework with recurrent component LSTM/GRU.</td>
</tr>
<tr>
<td>Seq2Seq+Attention</td>
<td>Seq2Seq_LSTM+Attention</td>
<td>Seq2Seq_LSTM/GRU with single attention mechanism.</td>
</tr>
<tr>
<td>Spatiotemporal attentive</td>
<td>STALSTM</td>
<td>Spatiotemporal attentive model with recurrent component LSTM/GRU</td>
</tr>
</tbody>
</table>

Note that Seq2Seq-based models are the combination of an Encoder–Decoder framework and LSTM or GRU and Seq2Seq+Attention applies a single attention mechanism based on Seq2Seq-based models. Spatiotemporal attentive-based methods include STAGRU and STALSTM. The settings of all models were consistent.

3. Results

We executed the six models mentioned above in each monitoring station in each city to predict the ozone levels in the subsequent 24 h. In each station, all models were trained and tested on the same training and testing datasets, respectively. The RMSE, R2 and SMAPE of each model from the nine stations are shown in Figures 3 and 4.

Figure 3. The performance of each model from the nine monitoring stations of the Nanjing and Beijing cities. The horizontal axis represents the prediction step and the vertical axis represents a specific metric. The solid line is the mean performance for each model from the nine stations. The shaded area represents the variation range in performance, where the upper bound is the maximum and the lower bound is the minimum.

Figure 3 shows the performance of each model in the Nanjing and Beijing cities, while Figure 4 shows the performance of each model in Guangzhou, Wuhan and Chengdu. The comparisons between the different types of models and between LSTM and GRU within each model were then analyzed. As shown in the two figures, the performance of all these models declines in the early stages and then gradually becomes stable after 12 h. As shown in the figures, the Seq2Seq-based model performed well in the early prediction stage, then it rapidly declined. Comparing LSTM and GRU, GRU generally performed better than LSTM. As for the Seq2Seq+Attention-based model, the performance was more
stable and errors accumulated more slowly than the Seq2Seq-based models. Considering our spatiotemporal-based model, the STAGRU performed best in the Nanjing and Beijing cities. The two spatiotemporal models performed similarly and made better predictions than the other models in the Chengdu, Guangzhou and Wuhan cities.

![Figure 4](image)

**Figure 4.** The performance of each model from the nine monitoring stations in the Guangzhou, Wuhan and Chengdu cities.

Generally, the error of the RNN-based models, especially under the Encoder–Decoder framework, accumulates as the forecast time increases; this is because the relevance between the past observation values and prediction values becomes weaker and thus the recursivity of the Encoder–Decoder framework makes further errors. The predictions of the Seq2Seq-based methods are generally adequate during the early forecasting hour; however, they begin to deteriorate at a much faster rate than the other methods as forecast time increases. The reason for this is that the only connection between the prediction and the past sequence is the last decoding hidden state, which cannot afford enough information to support forecasting due to the fixed length of the hidden state, thus making errors accumulate rapidly. Therefore, the fitting ability of the Seq2Seq-based model is unsatisfactory. The Seq2Seq+Attention-based methods introduce the attention mechanism to link each prediction with all past moments, which complements the shortage of Seq2Seq-based methods and generates a better model with improved performance. With the support of temporal information, Seq2Seq+Attention-based methods can significantly improve forecasting skills. However, Seq2Seq+Attention-based methods fluctuate as forecasting goes on. This is because the attention mechanism in the Seq2Seq+Attention-based models tends to learn periodicity from past sequences and the periodicity can be vulnerable to error accumulation which makes the model less dynamic. Spatiotemporal attentive-based methods bring an extra attention layer, similar to Seq2Seq+Attention, to make a connection between the observations of the spatially distributed stations.
In summary, Seq2Seq-based models are only sufficient to solve short-term sequence predictions due to their limited representation capabilities and fast error accumulation. Attention mechanisms can reinforce Seq2Seq, but the information they capture is still insufficient. Spatiotemporal attentive-based methods generally perform better than other models due to the spatiotemporal information learned, which improves the fitting ability and robustness of the models. Moreover, to demonstrate the adaptability of GRU and LSTM to the spatiotemporal attention mechanism, we compared the STAGRU and the STALSTM. The results showed that the GRU is a better choice than LSTM for our proposed model.

4. Discussion

4.1. Interpretability Discussion

Another important issue in O₃ prediction is how to interpret the results. For this purpose, we took the nine monitoring stations in Nanjing city as our analysis target. July 2019 was chosen for the interpretability discussion because of the generally consistent wind direction of the selected monitoring stations during this period. We collated the hourly data into a supervised dataset and 697 samples were obtained. The geographical distribution of the nine monitoring stations is shown in Figure 5a. The wind speed (classified based on the Beaufort wind scale) and wind direction distribution of all stations during this period are shown in Figure 5b. From Figure 5b, it can be seen that the dominant wind direction in July 2019 was ESE and the wind speed was between 1.3 m s⁻¹ and 3.1 m s⁻¹.

![Figure 5](image)

Figure 5. (a) The geographical locations of the nine monitoring stations in Nanjing. (b) The wind map of the monitoring stations in July 2019 (unit: frequency). Note, the hourly data contain wind direction and speed at the nine stations.

For the interpretability discussion of temporality, we reviewed how the temporal attention layer assigns weights to each past moment; some statistical procedures were then conducted. The statistical process was conducted in two steps as follows: (1) We computed the summation of the temporal attention weight using all samples in July 2019. A 24 × 24 matrix representing the sum of the attention weight for each forecast time and past moment was constructed. (2) The min-max normalization was applied to the matrix to highlight the relative importance of each past moment to the prediction of the target station at each forecast time (the closer the value is to 1, the more important the past moment is to the corresponding prediction step and vice versa).

As shown in Figure 6, the temporal attention layer tends to learn the short-term dependency in the prediction of the first several hours. Specifically, in the prediction of the 1st to 4th hour, the temporal attention layer assigns the largest weight to the −4th to −1st past moments and the Pearson correlation coefficient between the prediction steps and the most important moments was 0.94 (Figure 6). As the forecast lead time increases, the short-term dependency gradually shifts to periodical dependency. According to Figure 6, the periodical dependency dominates from the 8th to the last prediction step and the corresponding most important past moment is from the −18th to the −4th. The Pearson correlation coefficient was 0.99, which means a highly positive correlation (Figure 6). In
summary, the temporal information was learned in short-term and periodical dependency and the temporal attention finely captured the trade-off of these two kinds of dependency.

![Figure 6](image_url)

**Figure 6.** Weights that were assigned to each past moment by the temporal attention layer in PK. The $x$ axis represents the number of each past moment and the $y$ axis represents the prediction step (specifically, while the present moment is 0, +1 to +24 represents the forecasting moment and −1 to −24 represents the history to the current moment). Note that the weight 0 does not mean unrelated; it means relatively unimportant because of the min-max normalization.

For the interpretability discussion of how the predictions of the target station are affected by stations nearby in the STAGRU, the same statistical procedures were conducted. We computed the summation of the attention weight for each prediction step of each surrounding monitoring station in July 2019. A $24 \times 8$ matrix representing the relation between each forecast lead time and surrounding stations was created. The statistical results of the station CCM (located in the middle of the study area) are depicted in Figure 7. According to the heat map, it is obvious that the RJL is the most important surrounding monitoring station to the CCM for the whole prediction, which is consistent with their relative geolocation and the dominant wind direction. XWH, SXL, ZHM and ATZX are also important; a certain number of NE, ENE, ESE and SW winds were present according to Figure 5b. According to Figure 5a, it can be seen that a mountain stands upwind of the station MGQ, which weakens the airflow and reduces the wind. Further, the relative humidity of the Xuanwu Lake is significant, which negatively influences ozone formation [10]. Meanwhile, MGQ is located downwind. Consequently, the station MGQ becomes the least important neighbor. Empirically, we remove the data of station MGQ and then evaluate the performance. Before eliminating station MGQ, the average RMSE, $R^2$ and SMAPE in station CCM were 34.83, 0.59 and 48.55, respectively. After removing station MGQ, the average RMSE, $R^2$ and SMAPE were 35.18, 0.59 and 49.09, respectively. Thus, the performance attenuation was slight.
Figure 7. Heatmap of the relative importance of other stations concerning the CCM at each prediction step. Note, 0 means less important compared to the other stations.

In summary, on the one hand, the interpretable results are consistent with the actual findings, which investigated the reliability of our proposed model; on the other hand, analyzing the inner mechanism of the model continuously deepened and strengthened the presence of spatiality. Concretely, distance is not the only important factor influencing pollutant transmission; the results also show that the wind direction matters. Thus, future work concerning spatial prediction should pay more attention to utilizing additional spatial information, like wind direction, etc.

4.2. Derivative Model Discussion

In this study, we designed a model that learns spatial information from the encoding hidden states (Figure A1). However, spatial information can be also captured during decoding. Based on the STAGRU, we moved the spatial information learning process from the encoding hidden states of past sequences of each monitoring station to the decoding hidden states of each station that are in the same prediction step (namely STAGRU-Decoder). The details of this model are shown in Figure A2. Firstly, it is clear that the spatial attention layer of the STAGRU-Decoder receives the decoding hidden states of all monitoring stations in the same prediction step. Then, all stations make predictions simultaneously during decoding. In this manner, the STAGRU-Decoder can achieve synchronous prediction for multiple monitoring stations, which reduces the model training overhead.

We investigated the effectiveness of the STAGRU-Decoder by comparing it with the STAGRU in five cities. The results are shown in Figure A3. According to the results, the mean performance of the STAGRU-Decoder is better than the STAGRU in Beijing and Guangzhou and it performed similarly in other cities. However, the STAGRU-Decoder becomes unstable as the forecast lead time increases, after the 8th hour specifically, according to the undertint area. We considered that the predictions made by the STAGRU-Decoder for the monitoring stations are built on the predictions of others, which causes error superposition. Thus, the stability of the STAGRU-Decoder deteriorates as forecasting continues. Furthermore, we note that the applicable scope of the spatial information learning in the STAGRU and STAGRU-Decoder is limited by the wind force, where the air pollutant is transported more widely when the wind becomes stronger.
5. Conclusions

In this paper, we propose a novel model called the Spatiotemporal Gate Recurrent Unit (STAGRU), which captures spatiotemporal information using two types of attention mechanisms: temporal attention and spatial attention. Temporal attention captures information from the past sequence, while spatial attention captures information from the surrounding monitoring stations. We demonstrated the effectiveness of the STAGRU model compared to Seq2Seq and Seq2Seq+Attention models in five major cities: Nanjing, Beijing, Wuhan, Guangzhou and Chengdu. Statistically, our proposed method is 14% better than Seq2Seq-based methods and 4% better than Seq2Seq+Attention-based methods. Furthermore, we proposed another model that captures spatial information during decoding. This model was able to apply forecasting to multiple stations simultaneously, but it came at the cost of stability. In addition, it provided insight into our proposed model, enabling a discussion on interpretability from the perspective of statistical temporality and spatiality. The analysis shows that the temporality of ozone variation involves short-term dependency and long-term periodicity and that the spatiality of ozone transportation is mainly affected by wind, including speed and direction. By utilizing our model, policy decision-makers can make accurate ozone predictions in advance for specific regions of interest. If it is predicted that there will be ozone pollution in this area, decision-makers can provide early warnings and take appropriate measures to control the area. Current results presented in this manuscript are considered preliminary. For future work, the two major objectives are to extend the domain by including more observational stations (e.g., mainland China) and to further improve accuracy. It is worth noting that the amount of CPU time required for the calculation is directly proportional to the number of monitoring stations being considered. This implies that larger domain tests necessitate significant computational resources.

Author Contributions: X.C.: Conceptualization, Writing—review and editing and supervision. Y.L.: conceptualization, methodology, validation, investigation, data curation, writing—original draft and visualization. X.X.: methodology, investigation, writing—original draft, visualization and data curation. M.S.: writing—review and editing, resources and supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was partially supported by the National Natural Science Foundation of China (No.62276142, 62206133, 62202240, 62192783), the Special Science and Technology Innovation Program for Carbon Peak and Carbon Neutralization of Jiangsu Province (No. BE2022612), and the Natural Science Foundation of Jiangsu Province (No. BK20210574).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in FigShare accessed on 18 September 2023. Li, Yang (2023). Air pollutant and meteorological data of five major cities in China. figshare. Dataset.

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

Appendix A

This appendix contains three parts as follows: a detailed neural network structure of the proposed model in two versions; the performance comparison between the two version of the model; the geolocation of all the monitoring stations in each city.
Figure A1. The structure of the STAGRU. The target represents the past sequence of the target station, while $X_i$ represents the past sequence of the surrounding station $i$. Sending data to the Encoder, the GRU component encodes each moment into a hidden state and a hidden state matrix is produced, named $H_e(k,n_{in})$. Note, the row and column of $H_e(k,n_{in})$ depends on the number of stations and the length of each past sequence. The last encoding hidden state of the target sequence is fed into the Decoder. With this hidden state, the Decoder produces the decoding hidden state for each prediction step. In each prediction step, the temporal context vector, derived from the encoding hidden states of the target station and the spatial context vector, derived from the encoding hidden states with the highest attention weight in each monitor station, are applied to make a prediction. Specifically, the temporal context vector, $c_{temp}$, and the spatial context vector, $c_{spa}$, are concatenated with the current decoding hidden state, $h^d_t$, then the concatenation is sent to a linear layer to forecast $Y$.

Figure A2. The structure of the STAGRU-Decoder. The main difference is in the spatial information learning. The STAGRU-Decoder learns the spatiality from the decoding hidden states of the other surrounding stations. This operation is applied for the prediction step of each station. In this manner, there is no target station; all stations are forecasted synchronously.
Figure A3. Performance comparison between the Seq2Seq (GRU and LSTM), Seq2Seq+Attention (GRU_attention and LSTM_attention) and spatiotemporal attentive models (STAGRU_Encoder, STALSTM_Encoder, STAGRU_Decoder and STALSTM_Decoder) in Beijing, Chengdu, Guangzhou and Wuhan. The x axis represents the prediction step and the y axis represents the three criteria (RMSE, R2 and SMAPE). Solid line denotes the mean value over the nine monitoring stations in the prediction step and the shaded area is the performance variation region of a certain model.
Table A1. Longitude and latitude of each selected monitoring station in Nanjing, Beijing, Chengdu, Guangzhou and Wuhan.

<table>
<thead>
<tr>
<th>City</th>
<th>Station</th>
<th>Longitude</th>
<th>Latitude</th>
<th>Station</th>
<th>Longitude</th>
<th>Latitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nanjing</td>
<td>ATZX</td>
<td>118.737</td>
<td>32.009</td>
<td>RJL</td>
<td>118.803</td>
<td>32.031</td>
</tr>
<tr>
<td></td>
<td>CCM</td>
<td>118.749</td>
<td>32.057</td>
<td>SXL</td>
<td>118.778</td>
<td>32.072</td>
</tr>
<tr>
<td></td>
<td>MGQ</td>
<td>118.803</td>
<td>32.108</td>
<td>XLDXC</td>
<td>118.907</td>
<td>32.105</td>
</tr>
<tr>
<td></td>
<td>PK</td>
<td>118.626</td>
<td>32.088</td>
<td>XWH</td>
<td>118.795</td>
<td>32.078</td>
</tr>
<tr>
<td></td>
<td>ZHM</td>
<td>118.777</td>
<td>32.014</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beijing</td>
<td>WSXG</td>
<td>116.3621</td>
<td>39.8784</td>
<td>DL</td>
<td>116.2202</td>
<td>40.2915</td>
</tr>
<tr>
<td></td>
<td>DS</td>
<td>116.4174</td>
<td>39.9289</td>
<td>TT</td>
<td>116.4072</td>
<td>39.8863</td>
</tr>
<tr>
<td></td>
<td>HDWL</td>
<td>116.2878</td>
<td>39.9611</td>
<td>SYXC</td>
<td>116.6636</td>
<td>40.135</td>
</tr>
<tr>
<td></td>
<td>HRZ</td>
<td>116.6275</td>
<td>60.3275</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chengdu</td>
<td>JQLH</td>
<td>103.9728</td>
<td>30.7236</td>
<td>SLD</td>
<td>103.3832</td>
<td>23.0201</td>
</tr>
<tr>
<td></td>
<td>SWY</td>
<td>104.0594</td>
<td>30.5767</td>
<td>SHP</td>
<td>104.1122</td>
<td>30.6306</td>
</tr>
<tr>
<td></td>
<td>JP</td>
<td>104.0431</td>
<td>30.6556</td>
<td>LYS</td>
<td>103.6202</td>
<td>31.0201</td>
</tr>
<tr>
<td></td>
<td>DSXL</td>
<td>104.0219</td>
<td>30.6558</td>
<td>LQXQ</td>
<td>104.2725</td>
<td>30.5589</td>
</tr>
<tr>
<td></td>
<td>LJJ</td>
<td>103.8458</td>
<td>30.6994</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guangzhou</td>
<td>GYZX</td>
<td>113.2347</td>
<td>23.1423</td>
<td>SWZ</td>
<td>113.2612</td>
<td>23.105</td>
</tr>
<tr>
<td></td>
<td>GDSXY</td>
<td>113.3478</td>
<td>23.0916</td>
<td>SBSLZ</td>
<td>113.4332</td>
<td>23.1047</td>
</tr>
<tr>
<td></td>
<td>FYZX</td>
<td>113.3505</td>
<td>22.9483</td>
<td>HDSF</td>
<td>113.2146</td>
<td>23.3916</td>
</tr>
<tr>
<td></td>
<td>SJCZ</td>
<td>113.2597</td>
<td>23.1331</td>
<td>JLZZL</td>
<td>113.5618</td>
<td>23.312</td>
</tr>
<tr>
<td></td>
<td>LH</td>
<td>113.2765</td>
<td>23.1544</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wuhan</td>
<td>DHLY</td>
<td>114.3677</td>
<td>30.5584</td>
<td>HYYH</td>
<td>114.2529</td>
<td>30.558</td>
</tr>
<tr>
<td></td>
<td>HKHQ</td>
<td>114.282</td>
<td>30.6189</td>
<td>WCZY</td>
<td>114.3025</td>
<td>30.5332</td>
</tr>
<tr>
<td></td>
<td>QSGH</td>
<td>114.3646</td>
<td>30.6217</td>
<td>DKXQ</td>
<td>114.1566</td>
<td>30.4825</td>
</tr>
<tr>
<td></td>
<td>HKJT</td>
<td>114.3014</td>
<td>30.5944</td>
<td>WJS</td>
<td>114.135</td>
<td>30.6319</td>
</tr>
<tr>
<td></td>
<td>CHQH</td>
<td>113.8454</td>
<td>30.2917</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

References


27. Huang, L.; Zhang, C.; Bi, J. Development of land use regression models for PM$_{2.5}$, SO$_2$, NO$_2$ and O$_3$ in Nanjing, China. *Environ. Res.* 2017, 158, 542–552. [CrossRef]


35. Luna, A.; Paredes, M.; De Oliveira, G.; Corrêa, S. Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil. *Atmos. Environ.* 2014, 98, 98–104. [CrossRef]


52. Liu, B.; Yan, S.; Li, J.; Qu, G.; Li, Y.; Lang, J.; Gu, R. A sequence-to-sequence air quality predictor based on the n-step recurrent prediction. IEEE Access 2019, 7, 43331–43345. [CrossRef]
60. Rajagukguk, R.A.; Ramadhani, R.A.; Lee, H.J. A review on deep learning models for forecasting time series data of solar irradiance and photovoltaic power. Energies 2020, 13, 6623. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.