Spatiotemporal Attentive Gated Recurrent Unit: A Novel Method to Forecast O_3

Authors: Yang Li[1], Xingguo Chen[1], Xiaoyan Xu[2], and Min Shao[3]*

Affiliations:

[1]Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

[2]School of the Environment, Nanjing University, Nanjing 210046, China

[3]School of Environment, Nanjing Normal University, Nanjing 210023, China

*Corresponding authors: Min Shao, mshao@masonlive.gmu.edu

Key points:

- A spatiotemporal attention mechanism is proposed to achieve accurate predictions.
- Interpretability of model shows short-term and periodical dependency in temporality and importance of wind direction in spatiality.

Abstract

Due to the limited understanding of the physical/chemical processes and large uncertainties in emissions, ozone prediction task becomes more difficult with numerical models. Deep learning provides an alternative way. However, most of the deep learning ozone prediction models only consider temporality and have limited capacity. Exist 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), which employs double attention mechanism and Gated Recurrent Unit (GRU) to capture spatiotemporal information. We compare STAGRU with Seq2Seq and their single attention version on nine monitoring stations in Nanjing. The results show that STAGRU outperforms other competitors in terms of RMSE, R^2, and SMAPE. In addition, we make an interpretability discussion for STAGRU. The discussion reveals that wind direction plays an important role in ozone transmission and temporality mainly involves short-term and periodical dependency.

Plain Language Summary: Ozone is considered as one of the most concerned primary pollutants in air pollution control. Ozone prediction is a promising way to provide assistance to the control process, which however is challengeable due to the complex and imperfectible understanding of ozone formation. We propose a spatiotemporal model, namely Spatiotemporal Attentive Gated Recurrent Unit, which learn the temporality from the past sequence of the target and capture the spatiality from the surrounding stations. Further, learned

model reveals that temporality information involves short-term and periodical dependency and spatiality information implies the importance of wind direction.

1. Introduction

Ground level Ozone (O_3) is the products of photochemical reactions of VOCs and NO_x(Atkinson, 2000), which has spatiotemporal disparity primarily resulted from different emission characteristics, synoptic conditions, topographic distribution, and land use types(Carvalho et al., 2006; Meng et al., 2022; Tu et al., 2007; Wang et al., 2017; Yu et al., 2021). Generally, emission source and meteorological characteristics are the fundamental and essential factors to the formation, transport and dispersion of O_3(Mousavinezhad et al., 2021). The source of O_3 precursors emission can be labeled as anthropogenic source and natural source(Lelieveld & Dentener, 2000), while the meteorological variables such as solar radiation, wind direction, wind speed, atmospheric pressure, high temperature and low humidity exhibit complicated relation with O_3(Camalier et al., 2007; Dueñas et al., 2002; Hu et al., 2021; Li et al., 2017; Pu et al., 2017). Recently, ozone pollution has gradually increased and become one of the primary pollutants of great concern in air pollution control(Lu et al., 2018; Sun et al., 2016; Wang et al., 2009) because of its detrimental impacts on both human health and agriculture field(Dimakopoulou et al., 2020; Keiser et al., 2018; Michaudel et al., 2018; Zhang et al., 2021). Considering the complexity of O_3 formation mechanism, the aggravated atmospheric combined pollution increases the difficulty in ozone control(Wennberg & Dabdub, 2008). One of the most important tasks affecting the assessment of efficient prevention and control strategies for O_3 pollution is O_3 prediction. Building precise O_3 prediction model can strongly support decision makers efficiently cutting heavy ozone pollution peaks, which is urgent and necessary.

Generally, ozone prediction approaches can be classified into two types: numerical and statistical approaches. Numerical approaches simulate real atmospheric environment based on an accurate anthropogenic emissions estimation and specific atmospheric physics and chemistry reaction. Some numerical approaches(Bey et al., 2001; Dennis et al., 1996; Grell et al., 2005) have been widely used in ozone prediction. Unfortunately, numerical models suffer by imperfectible understanding of complex ozone formation and sacrificing spatiotemporal resolutions. Therefore, the spatiotemporal representativeness, emission and model mechanisms still need to be perfected(Zhou et al., 2017). By contrast, statistical models do not take into account complicated reaction mechanisms, having higher flexibility and more computing advantages(Schlink et al., 2006). Classical statistical ozone prediction models mainly consist of basic regression models(Huang et al., 2017; Hubbard & Cobourn, 1998; Kumar & Jain, 2010; Pagowski et al., 2006; Wang et al., 2015), which limit the capacity of describing mostly non-linear and complex internal physicochemical processes. Therefore, they usually fail in meeting practical requirements(Comrie, 1997; Robeson & Steyn, 1990). Machine Learning (ML) methods, as a promising one, have inspired the ozone forecasting domain. Basic ML algorithms(Burrows et al., 1995;

Luna et al., 2014), advanced ensemble algorithms(Cai, 2018; Ding et al., 2018; Eslami et al., 2020; Requia et al., 2020) and Artificial Neural Network (ANN) (Al-Alawi et al., 2008; Arhami et al., 2013; Tsai et al., 2009) have been intensively studied in ozone forecasting. Nevertheless, these models cannot focus on capturing spatial and temporal features simultaneously, consequently, more powerful models are needed.

To capture spatiotemporal information, the emerging Deep Learning (DL) is a good choice, which own powerful representative capability. Theoretically, DNN is able to fit any form of function, but it will be extremely hard to train. Considering the *No Free Lunch* Theorem(Wolpert, 1996; Wolpert & Macready, 1997), a majority of task-oriented neural network were conducted such as recurrent neural network (RNN)(Elman, 1990), which is designed for sequence forecasting. Nonetheless, RNN has the issues of gradient explosion and gradient vanishing, and also lack of long-term memory property. Accordingly, the Long-Short Term Memory (LSTM)(Lipton et al., 2015) and the Gated Recurrent Unit (GRU)(Chung et al., 2014) are proposed to avoid these problems by introducing memory units and gate mechanism(Pascanu et al., 2013). The LSTM and the GRU are normally applied in ozone prediction coupled with Encoder-Decoder framework(Sutskever et al., 2014). However, the performance of Encoder-Decoder is limited by the fixed length of hidden state. The presence of attention mechanism(Vaswani et al., 2017) breaks the bottleneck of Encoder-Decoder. With the attention mechanism, some methods(Liu et al., 2018, 2019) can achieve a more precise results. Besides temporal features, spatial factors are also crucial. Ozone pollution is usually a regional air quality issue, thus, it is not only affected by local emissions and meteorological conditions but also by long-range transport of ozone and precursors(Chung, 1977; Wild & Akimoto, 2001). Several DL based methods(Abirami & Chitra, 2021; Wang et al., 2019; Wang et al., 2020) have been proposed to learn spatiality, yet the application of these methods is usually hampered by model complexity.

In this paper, we propose a novel method called Spatiotemporal Attentive Gated Recurrent Unit (STAGRU) based on Seq2Seq model and double attention mechanism to predict local ozone levels using spatiotemporal information from in-situ observations. STAGRU applies two kinds of attention to capture both spatial and temporal information. Model construction details and data used for training and other experiments are introduced in section 2. Then, we conduct experiments that compare STAGRU with Seq2Seq and Seq2Seq+Attention based models. At last, we discuss about the interpretability of STAGRU, which provides us insights of temporality and spatiality. Furthermore, a derivative model, STAGRU-Decoding, is proposed, which predict ozone for multiple stations synchronously.

1. **Data and Methods**

2.1 Spatiotemporal attentive gated recurrent unit

$$w = Softmax\left(\text{score}\left([h]_{\text{past}}, h_{\text{current}}\right)\right),$$

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 learn spatial information from some specific moments in the past sequence of each surrounding station. Here, we propose to employ the particular moments not the whole sequence as it can reduce computation and alleviate the disturbance from unimportant moments. We utilize temporal attention layer to select the past moment that has the highest attention weight for each station and introduce another attention layer, spatial attention layer, to calculate spatial context vector. The purpose we use the temporal attention layer to make selections is that this layer contains knowledge about how to evaluate the importance of a certain past moment to the current prediction on the view of target station. After determining these specific moments, spatial attention layer calculates the attention score for each selected moment and obtain the spatial context vector. In this manner, spatial attention layer builds a connection between target station and its neighbors, and the knowledge in the past sequence of each station eventually affect the prediction of the target.

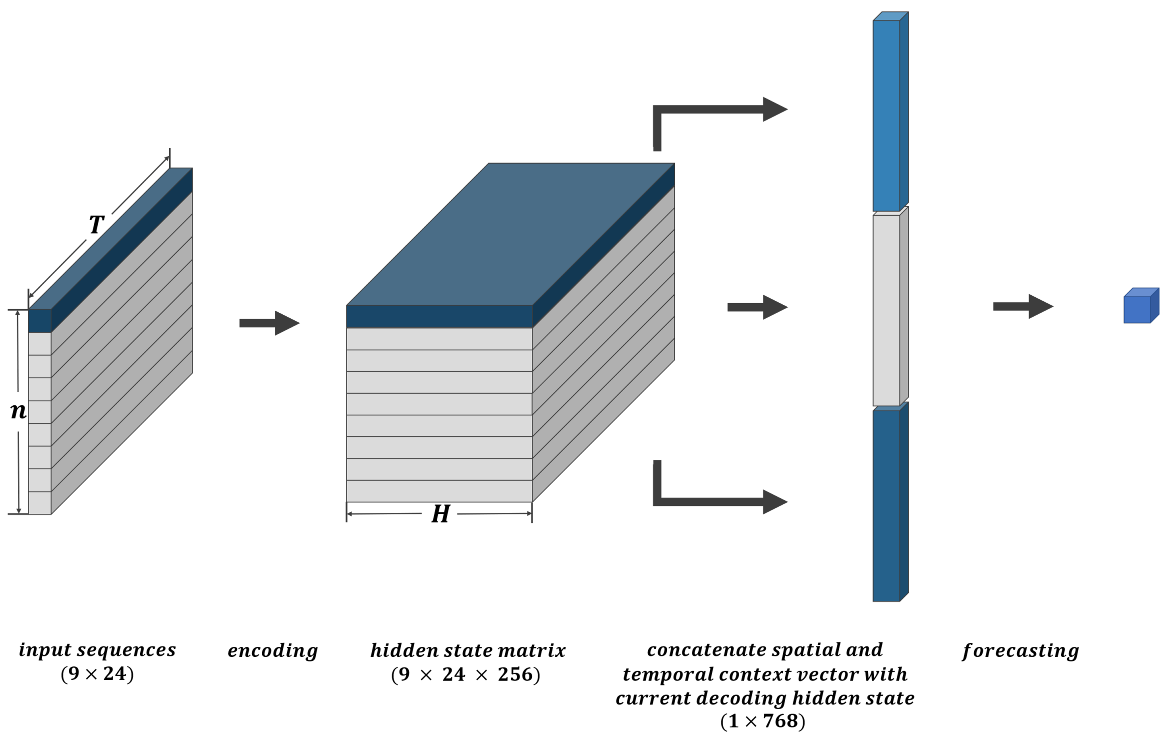

Figure 1. The conceptual diagram of STAGRU, where n is the number of stations, T is the length of past sequence and H is the size of hidden state. The

dark blue cuboid in the input sequence and the hidden state matrix represents the target station related information. Three vectors in the concatenate part are temporal context vector (upper), decoding hidden state (middle) and spatial context vector (lower). Both encoder and decoder are basing on GRU. (9×24) etc. represents the shape of data in that forward propagation step.

Along with the learning of temporal and spatial information, a model named STAGRU (Spatiotemporal Attentive Gated Recurrent Unit), which is spatiotemporal attentive based, is proposed. We employ the widely used recurrent neural network model, GRU, to be the time series related component, and compare the performance with another popular model LSTM in experiment. The process of STAGRU is shown in **Figure 1**. As shown in **Figure 1**, firstly, the past sequence of each station is fed into encoder. Encoder maps each moment into a hidden state, and a corresponding hidden state matrix is obtained. Then, temporal and spatial attention layer captures spatiotemporal information from the matrix, respectively, which produce temporal and spatial context vector. At last, two context vectors are concatenated with current decoding hidden state, which is sent to nonlinear layer to get current prediction. More details can be seen in **Figure S1**.

2.2 Dataset

In this paper, we utilized both air quality and meteorological data of nine monitoring stations in Nanjing from January 2017 to December 2020. The nine stations are ATZX (Aotizhongxin station), CCM (Caochangmen station), MGQ (Maigaoqiao station), PK (Pukou station), RJL (Ruijinlu station), SXL (Shanxilu station), XLDXC (Xianlindaxuecheng station), XWH (Xuanwuhu station) and ZHM (Zhonghuamen station). The longitude and latitude of each station is illustrated in **Table S1**, and their geographical locations is shown in **Figure 2a**. In **Figure 2a**, those 9 stations are distributed in different regions of Nanjing, like teaching area, downtown area and industrial area, etc.

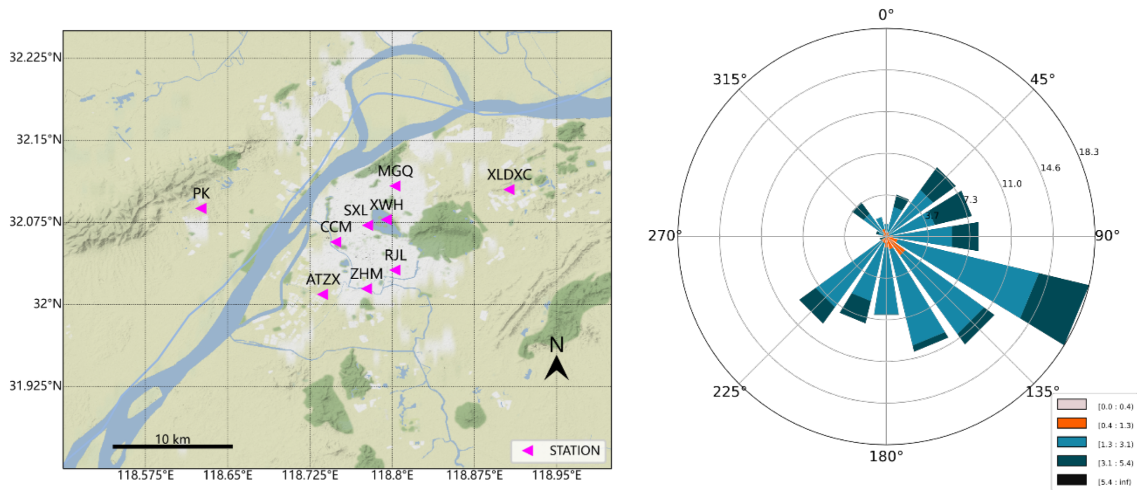

Figure 2. (a) the geographical locations of nine stations. (b) the wind rose map of the monitoring stations area in July 2019. (Unit: frequency). Hourly data contains wind direction and speed of nine stations.

Air quality data includes hourly $PM_{2.5}$, PM_{10}, NO_2, O_3, CO, SO_2 concentrations and the air quality index (AQI). The fifth generation of atmospheric reanalysis from European Centre for Medium-Range Weather Forecasts (ECMWF) (ERA5)(Hersbach et al., 2020) is used as meteorological inputs. ERA5 near-surface wind speed, wind direction, and temperature records are interpolated to the geolocations of the 9 stations in using ordinary kriging interpolation, where the variogram model is linear. We took the data from January 2017 to July 2019 as training data, August 2019 to March 2020 as validating data, and April 2020 to December 2020 as testing data. Validating data serves as an evaluator of the model to prevent over fitting in training. Testing data is used to evaluate the performance of final trained model.

Linear interpolation (Benesty et al., 2004) is used to fulfill the missing data. Normalization was applied to air quality data, wind speed, and temperature to achieve fast convergence, and one-hot encoding was applied to wind direction to gain corresponding binary code for training convenience. We transform raw dataset to supervised dataset which is consisted of input-output pairs based on sliding window. For the details of data flow, we took the past 24 hours air quality data and meteorological data of all stations as the input. 24-hour O_3 forecasts of the target station will be obtained via STAGRU.

2.3 Evaluation metric

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

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

where y_i represents the observation of item i; $\hat{y}_i$ represents the prediction of item i; n represents the number of items.

R^2 measures the fitting degree between the model and the true state. The formular is defined as follows

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

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

SMAPE measures the accuracy of predictions based on percentage errors, which is defined as follows

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

Note that each evaluating metric above have their own advantages and disadvantages, thus we integrate different perspectives to measure the effectiveness of the model.

2.4 Experiment design

In order to evaluate the capability of STAGRU, we compare it with Seq2Seq based and Seq2Seq+Attention based models. Further, we also consider that adopting LSTM to replace GRU in STAGRU, producing the STALSTM, because these two RNN methods are frequently used in many time series prediction tasks (Mehtab et al., 2020; Rajagukguk et al., 2020). The details of experiment design are shown in **Table S2**. The number of hidden units is 256, hidden layer is 1, batch size is 48, optimizer is Adam (Kingma & Ba, 2014), learning rate is 0.0001. Early stopping was applied to get a fine model, and schedule sampling was also used. Further, all experiments are conducted on NVIDIA GeForce GTX 1050Ti 4G for GPU and Intel i5-7300HQ 2.50GHz for CPU.

1. **Results**

We execute six models mentioned above on each monitoring station to predict future 24 hours ozone. In each station, all models are trained on the same training dataset and are tested on the same testing dataset. The RMSE, R^2, and SMAPE of each model in nine stations are shown in **Figure 3**. Model detailed results are illustrated in **Table S3** in Support information.

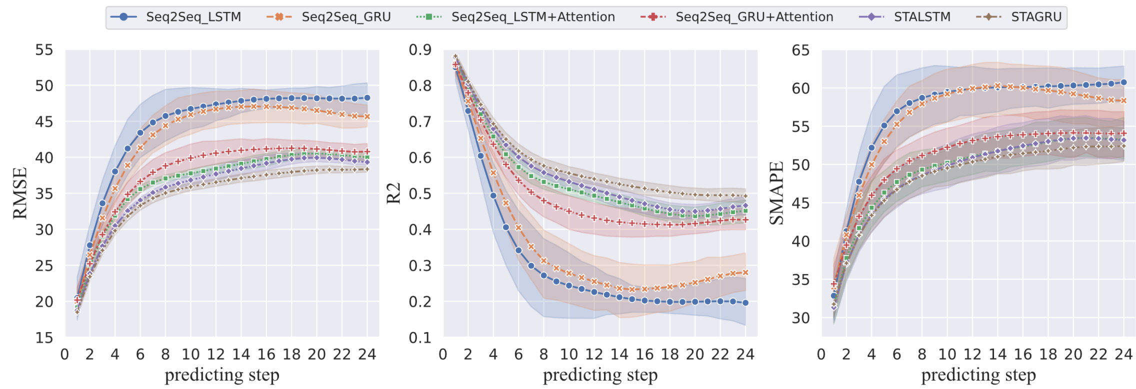

Figure 3. The performance of each model in nine monitoring stations. The horizontal axis represents predicting step, and the vertical axis represents specific metric. The solid line is the mean of performance for each model in nine stations. The undertint area represents the variation range of performance,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

As shown in **Figure 3**, the performance of all these models gets worse in the early stage, and then become stable after the 8^{th} hour. Seq2Seq based models perform worst and unstable. Seq2Seq+Attention based methods can achieve a better prediction than Seq2Seq. Concretely, there is a 27% improvement from Seq2Seq to Seq2Seq+Attention. Spatiotemporal attentive based methods are the best of all experiment categories in both accuracy and stability. There is a 7% improvement from Seq2Seq+Attention to spatiotemporal attentive. Further, STAGRU achieves the best performance and better than STALSTM.

Generally, the error of RNN based models, especially under Encoder-Decoder framework, accumulate as forecast lead time increase because the relevant between past observation values and predicting values is getting weak and the property of Encoder-Decoder framework, recursivity, makes error broadcast. The prediction of Seq2Seq based methods are generally adequate at the first forecast lead hour, but deteriorate at a much faster rate than the other methods as forecast lead time increases. The reason is that the only connection between 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 and make the error accumulate rapidly. Therefore, the fitting ability of the Seq2Seq based model is unsatisfied. Seq2Seq+Attention based methods introduce attention mechanism to relate each prediction with all past moments, which complement the shortage of Seq2Seq based methods and generate a better model with advanced 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. It is because the attention mechanism in Seq2Seq+Attention based models tends to learn periodicity from past sequence and the periodicity can be vulnerable to error accumulation that make the model vigorless. Spatiotemporal attentive based methods further bring in an extra attention layer comparing to Seq2Seq+Attention to make a connection between observations of spatially distributed stations.

In summary, Seq2Seq based models are only sufficient to solve short-term sequence prediction due to its limited representation capability and fast error accumulation. Attention mechanisms can reinforce Seq2Seq, but the information they capture is still insufficient. Spatiotemporal attentive based methods have the most satisfactory performance due to the spatiotemporal information learned, which improve the fitting ability and robustness. Moreover, to demonstrate the adaptability of GRU and LSTM to spatiotemporal attention mechanism, we compare the STAGRU and the STALSTM. The results showed that GRU is a better choice than LSTM for our proposed model.

1. **Discussion**

4.1 Interpretability discussion

Another important issue in O_3 prediction is how to interpret the results. For this purpose, July 2019 was chosen for interpretability discussion because of the generally consistent wind direction of the selected monitoring stations during this period. We formed the hourly data into supervised dataset and 697 samples were obtained. The wind speed (classified basing on Beaufort wind scale) and wind direction distribution in all stations during this period are shown in **Figure 2b**. From **Figure 2b**, it can be seen that the dominant wind direction in July 2019 is ESE and the wind speed is between 1.3m s^{-1} to 3.1m s^{-1}.

For the interpretability discussion of temporality, we review that how the temporal-attention layer assigns weights to each past moment, some statistical procedures were conducted. The statistical process can be concluded into two steps: 1) We compute the summation of the temporal attention weight using all samples in July 2019. A 24 $\times$ 24 matrix representing the sum of attention weight for each forecast lead time and past moment is constructed; 2) The min-max normalization is applied to the matrix to highlight the relative importance of each past moment to the prediction of the target

station at each forecast lead time (the closer the value is to 1, the more important the past moment is to the corresponding predicting step and vice versa).

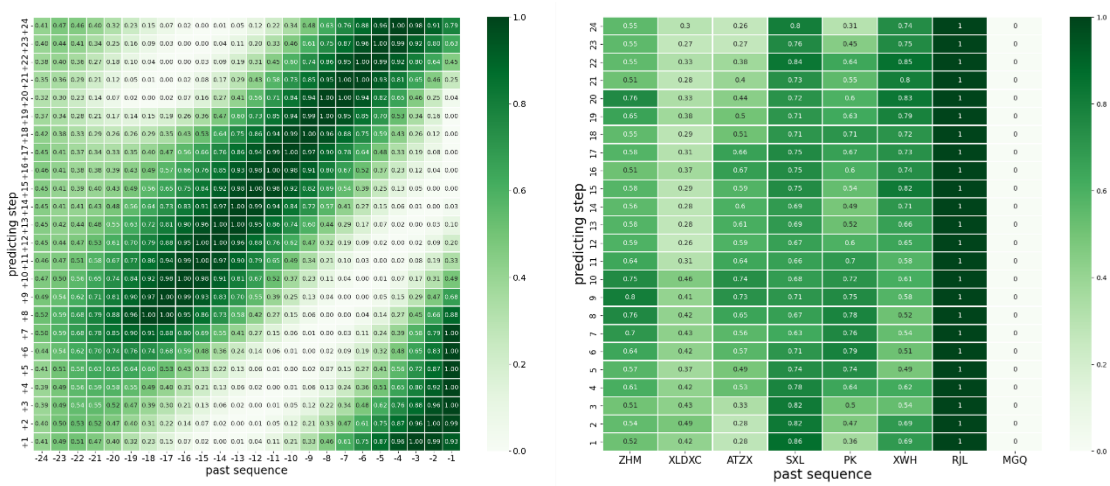

Figure 4. (a) Weights that were assigned to each past moment by temporal attention layer in PK. (b) Heatmap of the relative importance of other stations respect to CCM at each predicting step. The x axis represents the number of each past moment in (a) and represents surrounding stations in (b), and the y axis represents the predicting step (Specifically, while the present moment is 0, +1 to +24 represents forecasting moment and -1 to -24 represents the history to current moment). Note that the weight 0 does not mean totally unrelated which truly means relatively unimportant because of the min-max normalization.

As shown in **Figure 4a**, the temporal-attention layer tends to learn the short-term dependency in the prediction of first several hours. Specifically, in the prediction of the 1^{st} to the 4^{th} hour, the temporal-attention assigns the largest weight to the -4^{th} to -1^{st} past moment, and the Pearson correlation coefficient between predicting steps and the most important moments is 0.9439 (**Figure 4a**). As the forecast lead time increases, the short-term dependency shifts to periodical dependency gradually. According to **Figure 4a**, the periodical dependency dominates from the 8^{th} to the last prediction step, and the corresponding most important past moment is from the -18^{th} to the -4^{th}. Thire Pearson correlation coefficient is 0.9959, which means high positive correlation (**Figure 4a**). In summary, the temporal information learned involved in short-term and periodical dependency, and the temporal-attention finely captured the tradeoff of these two kinds of dependency.

For the interpretability discussion of how the predictions of the target station is affected by stations nearby in STAGRU, same statistical procedures were conducted. We compute the summation of the attention weight for each predicting step of each surrounding monitoring station in July 2019. A 24×8 matrix representing the relation between each forecast lead time and surrounding stations is created. The statistical results of the station CCM (located at the middle of the study area) are depicted in **Figure 4b**. According to the heat map, it is obvious that the RJL is the most important surrounding monitoring station to the CCM for the whole predicting, which is in consistence with their relative geolocations and the dominant wind direction. XWH, SXL, ZHM and ATZX are also important because there exist a certain number of NE, ENE, ESE and SW winds according to **Figure 2b**. According to the **Figure 2a**, it can be found that a mountain stands on the upwind position of the station MGQ, which weaken the airflow and reduce the wind blowing. Further, the relative humidity of the Xuanwu Lake is significant, which bring negative fluence to ozone formation (Dueñas et al., 2002). Meanwhile, MGQ is located at the downwind direction. Consequently, the station MGQ becomes the most unimportant neighbor. Empirically, we take off the data of station MGQ and evaluate the performance. Before eliminating station MGQ, the average RMSE, R^2 and SMAPE in station CCM are 34.83, 0.590 and 48.55 respectively. After removing station MGQ, the average RMSE, R^2 and SMAPE are 35.18, 0.593 and 49.09. The performance attenuation is slight.

In summary, on the one hand, the interpretable results are consistent with the actual cognition, which investigate the reliability of our proposed model, on the other hand, analyzing the inner mechanism of model continuously deepen and strengthen the acknowledgement of spatiality. Concretely, the distance is not the only important factor influencing pollutant transmission in here, the results show that the wind direction matters either. Thus, future works about spatial prediction can pay more attention to the utilizing of other spatial information, like wind direction etc.

4.2 Derivative model discussion

In this study, we designed the model that learns spatial information from the encoding hidden states (**Figure 1**). However, spatial information can be also captured during decoding. Basing on the STAGRU, we move the spatial information learning process from the encoding hidden states of past sequence of each monitoring station to the decoding hidden states of each station that in the same predicting step (namely STAGRU-Decoding). The details of this model are shown in **Figure S2**. Firstly, it is clear that the spatial attention layer of STAGRU-Decoding receives the decoding hidden states of all monitoring stations in the same predicting step. Then, all stations make predictions simultaneously in decoding. In this manner, STAGRU-Decoding is able to achieve synchronous prediction for multiple monitoring stations, which reduce model training overhead.

We investigate the effectiveness of STAGRU-Decoding by compare it with STAGRU in nine monitoring stations. The results are shown in **Figure S3**. According to the results, the mean performance of STAGRU-Decoding is similar to STAGRU, but it becomes unstable as the forecast lead time increases, after 8^{th} hour in specific, according to the undertint area. We consider that the predictions STAGRU-Decoding made for a monitoring station are built on the prediction of others, which causes error superposition. Thus, the stability of STAGRU-Decoding deteriorate as forecasting goes on. Also, we note that the applicable scope of the spatial information learning in STAGRU and STAGRU-

Decoding is limited by the wind-force, while the air pollutant transmits more widely along with the wind becoming stronger.

1. **Conclusion**

In this paper, we proposed a novel model, Spatiotemporal Gate Recurrent Unit (STAGRU), which capture spatiotemporal information using two kinds of attention mechanisms: temporal-attention and spatial-attention. Temporal-attention captures temporal information from the past sequence and spatial-attention captures spatial information from the surrounding monitoring stations. We illustrate the effectiveness of STAGRU in comparing with Seq2Seq based and Seq2Seq+Attention based models. Furthermore, we proposed another model that captures spatial information in decoding. This model is able to apply forecast to multi-stations synchronously at the expense of stability. In addition, to give an insight to our proposed model, we make interpretability discussion in the perspective of temporality and spatiality statistically. For the future works, we are going to extend the prediction to the whole of China and further improve the accuracy of our model.

Acknowledgements

This work was supported by the Natural Science Foundation of Jiangsu Province (No. BK20210574)

Data Availability

Data related to the meteorological inputs, including wind speed, wind angle and temperature records, are from the fifth generation of atmospheric reanalysis from European Centre for Medium-Range Weather Forecasts (ECMWF) (ERA5) (Hersbach et al., 2020), available at https://doi.org/10.24381/cds.adbb2d47.

Data related to the air pollution, including hourly $PM_{2.5}$, PM_{10}, NO_2, O_3, CO, SO_2 concentrations and the air quality index (AQI), are available at https://doi.org/10.17632/kvgwcrbjm3.1 .

Reference

https://doi.org/10.1016/j.jclepro.2020.125341

https://doi.org/10.1029/2001JD000807

https://doi.org/10.1016/j.atmosenv.2006.02.014

https://doi.org/10.1175/1520-0450(1977)016

https://doi.org/https://doi.org/10.1016/0364-0213(90)90002-E

https://doi.org/10.1016/j.envpol.2022.118865

https://doi.org/10.1016/j.atmosres.2007.02.003

https://doi.org/10.1038/s41598-017-03929-w

https://doi.org/10.3390/atmos12121675
Abirami, S., & Chitra, P. (2021). Regional air quality forecasting using spatiotemporal deep learning. *Journal of Cleaner Production*, *283*. Al-Alawi, S. M., Abdul-Wahab, S. A., & Bakheit, C. S. (2008). Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone. *Environmental Modelling & Software*, *23*(4), 396-403. Arhami, M., Kamali, N., & Rajabi, M. M. (2013). Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations. *Environmental Science and Pollution Research*, *20*(7), 4777-4789. Atkinson, R. (2000). Atmospheric chemistry of VOCs and NOx. *Atmospheric Environment*, *34*(12-14), 2063-2101. Benesty, J., Chen, J., & Huang, Y. (2004). Time-delay estimation via linear interpolation and cross correlation. *IEEE Transactions on speech and audio processing*, *12*(5), 509-519. Bey, I., Jacob, D. J., Yantosca, R. M., Logan, J. A., Field, B. D., Fiore, A. M., Li, Q. B., Liu, H. G. Y., Mickley, L. J., & Schultz, M. G. (2001). Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation. *Journal of Geophysical Research: Atmospheres*, *106*(D19), 23073-23095. Burrows, W. R., Benjamin, M., Beauchamp, S., Lord, E. R., McCollor, D., & Thomson, B. (1995). CART decision-tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver, Montreal, and Atlantic regions of Canada. *Journal of Applied Meteorology and Climatology*, *34*(8), 1848-1862. Cai, W. (2018). Using Machine Learning Method for Predicting the Concentration of Ozone in the Air. Camalier, L., Cox, W., & Dolwick, P. (2007). The effects of meteorology on ozone in urban areas and their use in assessing ozone trends. *Atmospheric Environment*, *41*(33), 7127-7137. Carvalho, A. C., Carvalho, A., Gelpi, I., Barreiro, M., Borrego, C., Miranda, A. I., & Perez-Munuzuri, V. (2006). Influence of topography and land use on pollutants dispersion in the Atlantic coast of Iberian Peninsula. *Atmospheric Environment*, *40*(21), 3969-3982. Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*. Chung, Y. S. (1977). GROUND-LEVEL OZONE AND REGIONAL TRANSPORT OF AIR-POLLUTANTS [Article]. *JOURNAL OF APPLIED METEOROLOGY*, *16*(11), 1127-1136. <1127:GLOART>2.0.CO;2 Comrie, A. C. (1997). Comparing neural networks and regression models for ozone forecasting. *Journal of the Air & Waste Management Association*, *47*(6), 653-663. Dennis, R. L., Byun, D. W., Novak, J. H., Galluppi, K. J., Coats, C. J., & Vouk, M. A. (1996). The next generation of integrated air quality modeling: EPA's Models-3. *Atmospheric Environment*, *30*(12), 1925-1938. Dimakopoulou, K., Douros, J., Samoli, E., Karakatsani, A., Rodopoulou, S., Papakosta, D., Grivas, G., Tsilingiridis, G., Mudway, I., & Moussiopoulos, N. (2020). Long-term exposure to ozone and children's respiratory health: Results from the RESPOZE study. *Environmental research*, *182*, 109002. Ding, S., Chen, B., Wang, J., Chen, L., Zhang, C., Sun, S., & Huang, C. (2018). An applied research of decision-tree based statistical model in forecasting the spatial-temporal distribution of O3.

Acta Scientiae Circumstantiae, *38*(8), 3229-3242. Dueñas, C., Fernández, M., Cañete, S., Carretero, J., & Liger, E. (2002). Assessment of ozone variations and meteorological effects in an urban area in the Mediterranean Coast. *Science of the Total Environment*, *299*(1-3), 97-113. Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179-211. Eslami, E., Salman, A. K., Choi, Y., Sayeed, A., & Lops, Y. (2020). A data ensemble approach for real-time air quality forecasting using extremely randomized trees and deep neural networks. *Neural Computing and Applications*, *32*(11), 7563-7579. Grell, G. A., Peckham, S. E., Schmitz, R., McKeen, S. A., Frost, G., Skamarock, W. C., & Eder, B. (2005). Fully coupled "online" chemistry within the WRF model. *Atmospheric Environment*, *39*(37), 6957-6975. Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., & Schepers, D. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999-2049. Hu, C., Kang, P., Jaffe, D. A., Li, C., Zhang, X., Wu, K., & Zhou, M. (2021). Understanding the impact of meteorology on ozone in 334 cities of China. *Atmospheric Environment*, *248*, 118221. Huang, L., Zhang, C., & Bi, J. (2017). Development of land use regression models for PM2. 5, SO2, NO2 and O3 in Nanjing, China. *Environmental research*, *158*, 542-552. Hubbard, M. C., & Cobourn, W. G. (1998). Development of a regression model to forecast ground-level ozone concentration in Louisville, KY. *Atmospheric Environment*, *32*(14-15), 2637-2647. Keiser, D., Lade, G., & Rudik, I. (2018). Air pollution and visitation at US national parks. *Science advances*, *4*(7), eaat1613. Kingma, D., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *Computer Science*. Kumar, U., & Jain, V. (2010). ARIMA forecasting of ambient air pollutants (O 3, NO, NO 2 and CO). *Stochastic Environmental Research and Risk Assessment*, *24*(5), 751-760. Lelieveld, J., & Dentener, F. J. (2000). What controls tropospheric ozone? *Journal of Geophysical Research: Atmospheres*, *105*(D3), 3531-3551. Li, K., Chen, L., Ying, F., White, S. J., Jang, C., Wu, X., Gao, X., Hong, S., Shen, J., & Azzi, M. (2017). Meteorological and chemical impacts on ozone formation: A case study in Hangzhou, China. *Atmospheric Research*, *196*, 40-52. Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*. Liu, B., Yan, S., Li, J., Qu, G., Li, Y., Lang, J., & Gu, R. (2018). An attention-based air quality forecasting method. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Liu, B., Yan, S., Li, J., Qu, G., Li, Y., Lang, J., & Gu, R. (2019). A sequence-to-sequence air quality predictor based on the n-step recurrent prediction. *IEEE Access*, *7*, 43331-43345. Lu, X., Hong, J., Zhang, L., Cooper, O. R., Schultz, M. G., Xu, X., Wang, T., Gao, M., Zhao, Y., & Zhang, Y. (2018). Severe surface ozone pollution in China: a global perspective. *Environmental Science & Technology Letters*, *5*(8), 487-494. Luna, A., Paredes, M., De Oliveira, G., & Corrêa, S. (2014). Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil. *Atmospheric Environment*, *98*, 98-104. Mehtab, S., Sen, J., & Dasgupta, S. (2020). Robust analysis of stock price time series using CNN and LSTM-based deep learning models. 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Meng, X., Wang, W. D., Shi, S., Zhu, S. Q., Wang, P., Chen, R. J., Xiao, Q. Y., Xue, T., Geng, G. N., Zhang, Q., Kan, H. D., & Zhang, H. L. (2022). Evaluating the spatiotemporal ozone characteristics with high-resolution predictions in mainland China, 2013-2019. *Environmental Pollution*, *299*, Article 118865. Michaudel, C., Mackowiak, C., Maillet, I., Fauconnier, L., Akdis, C. A., Sokolowska, M., Dreher, A., Tan, H.-T. T., Quesniaux, V. F., & Ryffel, B. (2018). Ozone exposure induces respiratory barrier biphasic injury and inflammation controlled by IL-33. *Journal of Allergy and Clinical Immunology*, *142*(3), 942-958. Mousavinezhad, S., Choi, Y., Pouyaei, A., Ghahremanloo, M., & Nelson, D. L. (2021). A comprehensive investigation of surface ozone pollution in China, 2015–2019: Separating the contributions from meteorology and precursor emissions. *Atmospheric Research*, *257*, 105599. Pagowski, M., Grell, G., Devenyi, D., Peckham, S., McKeen, S., Gong, W., Delle Monache, L., McHenry, J., McQueen, J., & Lee, P. (2006). Application of dynamic linear regression to improve the skill of ensemble-based deterministic ozone forecasts. *Atmospheric Environment*, *40*(18), 3240-3250. Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. International conference on machine learning, Pu, X., Wang, T., Huang, X., Melas, D., Zanis, P., Papanastasiou, D., & Poupkou, A. (2017). Enhanced surface ozone during the heat wave of 2013 in Yangtze River Delta region, China. *Science of the Total Environment*, *603*, 807-816. Rajagukguk, R. A., Ramadhan, R. A., & Lee, H.-J. (2020). A review on deep learning models for forecasting time series data of solar irradiance and photovoltaic power. *Energies*, *13*(24), 6623. Requia, W. J., Di, Q., Silvern, R., Kelly, J. T., Koutrakis, P., Mickley, L. J., Sulprizio, M. P., Amini, H., Shi, L., & Schwartz, J. (2020). An ensemble learning approach for estimating high spatiotemporal resolution of ground-level ozone in the contiguous United States. *Environmental science & technology*, *54*(18), 11037-11047. Robeson, S., & Steyn, D. (1990). Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations. *Atmospheric Environment. Part B. Urban Atmosphere*, *24*(2), 303-312. Schlink, U., Herbarth, O., Richter, M., Dorling, S., Nunnari, G., Cawley, G., & Pelikan, E. (2006). Statistical models to assess the health effects and to forecast ground-level ozone. *Environmental Modelling & Software*, *21*(4), 547-558. Sun, L., Xue, L., Wang, T., Gao, J., Ding, A., Cooper, O. R., Lin, M., Xu, P., Wang, Z., & Wang, X. (2016). Significant increase of summertime ozone at Mount Tai in Central Eastern China. *Atmospheric Chemistry and Physics*, *16*(16), 10637-10650. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in neural information processing systems, Tsai, C.-h., Chang, L.-c., & Chiang, H.-c. (2009). Forecasting of ozone episode days by cost-sensitive neural network methods. *Science of the Total Environment*, *407*(6), 2124-2135. Tu, J., Xia, Z.-G., Wang, H., & Li, W. (2007). Temporal variations in surface ozone and its precursors and meteorological effects at an urban site in China. *Atmospheric Research*, *85*(3-4), 310-337. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you* *need* Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA. Wang, H., Li, X. B., Wang, D., Zhao, J., & Peng, Z. (2019). Regional prediction of ground-level ozone using a hybrid sequence-to-sequence deep learning approach. *Journal of Cleaner Production*, *253*, 119841. Wang, M., Keller, J. P., Adar, S. D., Kim, S.-Y., Larson, T. V., Olives, C., Sampson, P. D., Sheppard, L., Szpiro, A. A., & Vedal, S. (2015). Development of long-term spatiotemporal models for ambient ozone in six metropolitan regions of the United States: the MESA Air study. *Atmospheric Environment*, *123*, 79-87. Wang, S., Li, Y., Zhang, J., Meng, Q., Meng, L., & Gao, F. (2020). PM2.5-GNN: A Domain Knowledge Enhanced Graph Neural Network For PM2.5 Forecasting. Wang, T., Wei, X., Ding, A., Poon, C., Lam, K. S., Li, Y. S., Chan, L., & Anson, M. (2009). Increasing surface ozone concentrations in the background atmosphere of Southern China, 1994–2007. *Atmospheric Chemistry and Physics*, *9*(16), 6217-6227. Wang, W. N., Cheng, T. H., Gu, X. F., Chen, H., Guo, H., Wang, Y., Bao, F. W., Shi, S. Y., Xu, B. R., Zuo, X., Meng, C., & Zhang, X. C. (2017). Assessing Spatial and Temporal Patterns of Observed Ground-level Ozone in China. *Scientific Reports*, *7*, Article 3651. Wennberg, P. O., & Dabdub, D. (2008). Rethinking ozone production. *Science*, *319*(5870), 1624-1625. Wild, O., & Akimoto, H. (2001). Intercontinental transport of ozone and its precursors in a three-dimensional global CTM. *Journal of Geophysical Research: Atmospheres*, *106*(D21), 27729-27744. Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural computation*, *8*(7), 1341-1390. Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, *1*(1), 67-82. Yu, R. L., Lin, Y. L., Zou, J. H., Dan, Y. B., & Cheng, C. (2021). Review on Atmospheric Ozone Pollution in China: Formation, Spatiotemporal Distribution, Precursors and Affecting Factors. *Atmosphere*, *12*(12), Article 1675. Zhang, Y., Ma, Y., Feng, F., Cheng, B., Shen, J., Wang, H., Jiao, H., & Li, M. (2021). Respiratory mortality associated with ozone in China: A systematic review and meta-analysis. *Environmental Pollution*, 116957. Zhou, G., Xu, J., Xie, Y., Chang, L., Gao, W., Gu, Y., & Zhou, J. (2017). Numerical air quality forecasting over eastern China: An operational application of WRF-Chem. *Atmospheric Environment*, *153*, 94-108.