

RESEARCH ARTICLE

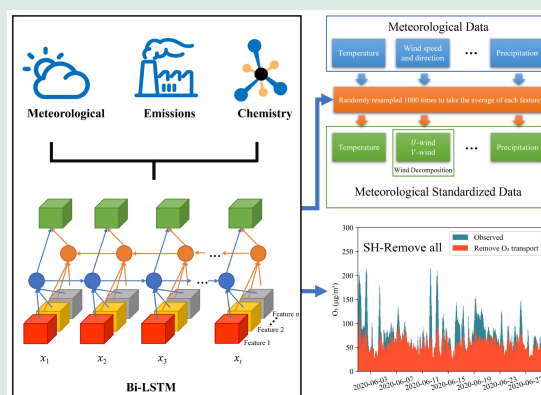
Data-driven machine learning quantifies ozone transport in the Hangzhou Bay urban cluster

Yuanxin Zhang¹, Shuwei Zhang², Song Gao ², Zhukai Ning², Zheng Jiao^{2,3}, Qing Hu ¹

1. State Key Laboratory of Soil Pollution Control and Safety, School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China
2. School of Environmental and Chemical Engineering, Shanghai University, Shanghai 200444, China
3. Shanghai Applied Radiation Institute, Shanghai University, Shanghai 200444, China


HIGHLIGHTS

- The Bi-LSTM model can identify and quantify the regional ozone transport.
- The meteorological removal method can identify regional ozone transport drivers.
- Ozone pollution characteristics are related to economics and geography.
- One-quarter of Shanghai's June ozone is transported from other cities.
- Regional ozone transport is evident on heavily polluted days.



ABSTRACT: Severe ozone (O_3) pollution has always been a serious problem faced by areas with rapid economic development, and the regional O_3 transport between cities is a major cause of this problem. Therefore, we used a bidirectional long short-term memory (Bi-LSTM) model to quantitatively identify the regional O_3 transport in Hangzhou Bay, China. Combined with the meteorological removal method, we were able to model O_3 concentrations that were not affected by transport. The contribution of regional transport to Shanghai's O_3 was quantified and validated using two different simulation schemes, which yielded highly consistent results of $18.41 \mu\text{g}/\text{m}^3$ (24% contribution) and $20.52 \mu\text{g}/\text{m}^3$ (27% contribution). According to the model simulation results, we found that approximately 24% of the O_3 pollution in Shanghai originates from other cities in the summer when the O_3 pollution is high. In addition, the regional O_3 transport was mainly concentrated during the high-value weather of O_3 pollution in Shanghai, and transport on non-pollution days was not apparent. Therefore, the regional O_3 transport from other cities is an important source of O_3 pollution in Shanghai. Overall, our study demonstrates the potential of machine-learning models coupled with meteorological removal for quantifying the inter-city influence of atmospheric pollutants.

KEYWORDS: Ozone transport, Ozone pollution, Machine learning, Meteorological removal, Bi-LSTM

 Corresponding authors. E-mails: njulegao@163.com (S. Gao); huq@sustech.edu.cn (Q. Hu)

Article history: Received 16 May 2025, Revised 30 August 2025, Accepted 31 August 2025, Available online 28 September 2025

© The Author(s) 2025. This article is published with open access at link.springer.com and journal.hep.com.cn

Special Issue—Atmospheric and Earth System Modeling towards Coordinated Pollution Control and Climate Change Mitigation (Guest Editors: Yuhang Wang, Hongliang Zhang, Yang Gao, Bin Zhao & Peng Wang)

1 Introduction

With the rapid development of China's economy, industrialization and urbanization have accelerated. As a result, the population and number of motor vehicles have shown a clear growth trend, resulting in a sharp increase in air pollutant concentrations (Song et al., 2017). Among them, near-ground O₃ concentrations have increased rapidly (Li et al., 2018a). Owing to the critical impact of O₃ on health (Cohen et al., 2017), ecology, and the climate, scholars and policymakers in China and abroad have made great efforts to reduce and control it (Guan et al., 2021). For example, after China formulated the Air Pollution Prevention and Control Action Plan (APPCAP) in 2013, the ozone (O₃) concentration gradually increased, so the APPCAP began to pay more attention to the coordinated control of O₃ and PM_{2.5} (Zhao et al., 2021). Europe is also taking action on O₃ pollution, with the rising O₃ levels posing a significant public health problem in EU-28 cities (Sicard et al., 2021). However, the current O₃ reduction and control efforts on a national scale are not significant because of the increase in background O₃ concentrations globally (Zhu et al., 2017). As a region with severe recent O₃ pollution (Lu et al., 2019), O₃ pollution research in China has attracted attention (Gao et al., 2016). Pollutant transport plays a vital role in O₃ formation (Wang et al., 2017) including in several heavily polluted regions in China such as the Beijing–Tianjin–Hebei urban zone, the Yangtze River Delta, and the Pearl River Delta (Zhu et al., 2022).

In recent years, signs of photochemical smog pollution have been found in different parts of China, and the average and maximum concentrations of O₃, frequency of limit exceedances, and duration of high values have been increasing annually in regions with rapid economic development (Qi et al., 2023). High concentrations of O₃ can produce photochemical smog pollution and endanger human health (Liu et al., 2021b). In addition, O₃ can be transported inter-regionally by air currents, thereby affecting other regions and cities and causing a series of environmental pollution problems (Xue et al., 2014). For example, long-range O₃ emissions contribute to California (USA)'s increased mortality burden from air pollution (Wang et al., 2019). Although O₃ pollution events usually occur in densely populated and industrialized regions, high concentrations of O₃ and its precursors migrate from one region to another via atmospheric transport (Foret et al., 2014), increasing surface-O₃ concentration in the Yangtze River Delta by 50 µg/m³ (Zhao et al., 2022). Therefore, quantifying inter-

regional O₃ influence is particularly important for combating O₃ pollution and providing strategic guidance for future atmospheric O₃ pollution control measures (Jiang et al., 2012).

However, to understand air quality and control O₃ pollution in real-time, traditional O₃ concentration monitoring technology mainly relies on traditional single-point instruments to obtain near-ground O₃ concentration trends (Gao et al., 2017), with meteorological data for backward trajectory analysis (Zhang et al., 2019) and the use of chemistry-transport models (CTMs) to simulate the process of O₃ transport and generation (Hu et al., 2018). Because CTMs are affected by emission inventories, complex chemical mechanisms, and other factors, obtaining more accurate results often requires considerable time to prepare and calculate. Researchers use machine learning to replace gas-phase chemistry solvers in CTMs can reach the highest acceleration of 85.2 times (Liu et al., 2021a). It is difficult to apply CTMs to some complex studies, particularly retrospective, multi-scenario analyses of inter-city transport. The primary challenges are their high computational cost and a strong dependence on emission inventories, which can have significant uncertainties. Therefore, we need to find new approaches to address the limitations of CTMs.

Recently, machine learning has attracted the attention of researchers owing to its powerful ability to resolve complex data patterns (Zhong et al., 2021), making it possible to overcome the limitations of traditional methods on pollutant transport effects and provide new opportunities for environmental pollution research (Liu et al., 2022). For example, a spatiotemporal weighted machine-learning method was used to obtain China's daily 1 km surface NO₂ concentrations (Wei et al., 2022). A machine-learning approach was applied to analyze potential causal relationships between COVID-19 severity and environmental factors (Kang et al., 2021). A data-driven machine-learning approach was used to predict emergent contaminants in the aquatic environment (Tong et al., 2022). Machine learning directly learn from the data and neural networks can learn almost all kinds of functions, which therefore can capture much more complicated relationships than pre-defined mechanisms. Compared to traditional methods, machine-learning methods reduce the reliance on emission inventories, relying only on large amounts of historical monitoring data to obtain complex nonlinear relationships between variables and O₃ (Pak et al., 2020). Additionally, compared to traditional models, machine-learning models save much computational time and computational resources (Wang et al., 2021) and are supported by big data models, showing better

performance and accuracy (Reichstein et al., 2019). Air quality data can be thought of as a time series dataset. Quantitatively identifying the regional O₃ transport relies on fitting the relationship between space–time and O₃ concentration. Commonly applied machine-learning models for time series include recurrent neural networks (RNNs) (Feng et al., 2019), long short-term memory (LSTM) networks (Kim et al., 2019), and gated recurrent unit (GRU) networks (Tao et al., 2019). In the above models, air quality data are only transmitted in one direction in the model, and there is a lack of interpretation of the relationship between future and past data. Therefore, their fitting results are often not satisfactory (Qu et al., 2022).

In this study, a machine-learning model based on a bidirectional LSTM (Bi-LSTM) model was used to identify and quantify the regional transport effects of O₃. It is important to note that this framework is designed for efficient retrospective analysis, not real-time forecasting. As a data-driven approach, it offers a computationally efficient alternative to CTMs that does not rely on emission inventories for historical quantification. Unlike traditional RNN and LSTM models, which can only capture the relationship between variables in one direction, Bi-LSTM can better capture the nonlinear relationship between variables and pollutant concentrations in both forward and reverse directions (Siami-Namini et al., 2019). We used a trained model, changed the feature variables that are closely related to the transport effect, and evaluated the change in the model output to quantify the O₃ transport influence.

2 Study area and dataset

2.1 Study area and monitoring data

We chose to study the Hangzhou Bay city cluster throughout 2020. The Hangzhou Bay urban agglomeration includes the cities of Shanghai (SH), Hangzhou (HZ), Ningbo (NB), Shaoxing (SX), Jiaxing (JX), Huzhou (HuZ), and Zhoushan (ZS), with a land area of approximately 45400 km², total population of over 70 million, and GDP of over RMB 9 trillion yuan by 2021. Figure 1 shows the GDP and year-on-year growth of the Hangzhou Bay urban agglomeration in 2021. As one of the regions with the fastest economic growth in China, Hangzhou Bay has experienced increasingly severe air pollution, represented by O₃ pollution in recent years (Xue et al., 2023). As the most economically developed city, Shanghai is affected by O₃ pollution year-round. In addition to the local

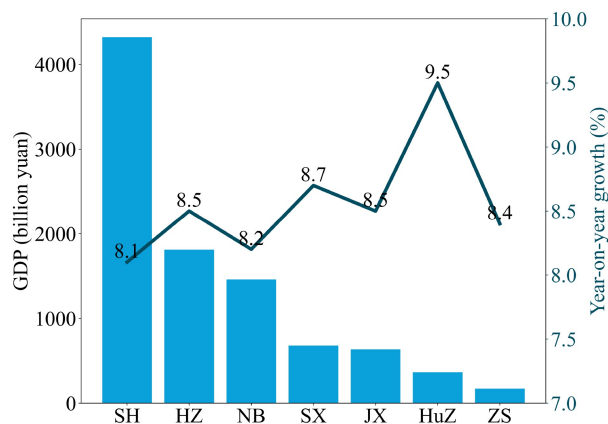


Fig. 1 GDP data of the Hangzhou Bay urban agglomeration. The histogram represents the total GDP, and the line chart represents the year-on-year growth rate.

generation of O₃ pollution, a considerable part of it is transported from surrounding areas, which is especially evident in economically developed areas (Suciu et al., 2017). Therefore, studying O₃ transport in rapidly developing economic areas can help identify prevention and control opportunities for O₃ pollution.

Hourly air quality data in the study area were obtained from the National Urban Air Quality Real-time Release Platform of the China National Environmental Monitoring Station (CNEMC). The data assurance, data quality, and instrument operation strictly followed the National Ambient Air Quality Standard (GB3095-2012). The air quality data content included the one-hour average concentration of PM_{2.5}, PM₁₀, SO₂, NO₂, O₃, and CO and the air quality index (AQI) at seven urban sites. The calculation method of AQI strictly followed the Technical Regulations on Ambient AQI (HJ633-2012). Meteorological data from the integrated surface database (ISD) of the National Climatic Data Center, including air temperature, dew point temperature, sea level pressure, wind direction, wind speed, cloud cover, and precipitation, were also used in this study. These data are provided as hourly averages and strictly follow NOAA Information Quality Guidelines in terms of data quality and data assurance.

2.2 Data preprocessing

We selected and reviewed meteorological and air quality data from the monitoring stations within each city. We eliminated abnormal data caused by instrument and sensor failure. However, when encountering routine maintenance activities such as calibration of monitoring instruments, communication failures, or power outages, some sites may experience

data loss. Therefore, we used data imputation to fill in the missing data to obtain the complete dataset. We chose the method of linear interpolation to fill in the missing data shown in Eq. (S1). When the station data are missing for a period (e.g., over 24 h), the time-series data filled by linear interpolation may not reflect the hourly variation of O₃. Therefore, we treat data missing for more than 24 h as anomalous data to be excluded. Shanghai has one of the largest economies and the most serious O₃ pollution in Hangzhou Bay. Consequently, the study of O₃ transport in Shanghai is significant for O₃ control throughout Hangzhou Bay and even the Yangtze River Delta region. We used the O₃ concentration in Shanghai as the target value, and all other data were used as characteristic variables for simulation.

To obtain a reliable and stable model, cross-validation (CV) was used to evaluate the performance of a machine-learning model (Roberts et al., 2017). To solve the problem that the data recorded by air pollution monitors exhibit space–time dependence, researchers proposed the leave-one-location-out (LOLO) CV method (Watson et al., 2019). We evaluated the predicted performance of the Bi-LSTM model using LOLO CV where the data of one of the cities are the validation set and the data of the other six cities are the training set. Each city was used once as a validation set, and seven CV datasets (CV 1–7) were constructed with 8760 datapoints each. Additionally, we chose the random forest (RF) model and the LSTM model to use the same dataset for validation to compare to the Bi-LSTM model. Additionally, we chose the RF model and the LSTM model to use the same dataset for validation to compare to the Bi-LSTM model. Furthermore, to rigorously assess the model's temporal generalization ability, we conducted an independent-validation (IV). For this, the model was trained on the entire 2020 dataset and then tested on the complete, unseen dataset from 2021.

3 Methods

3.1 Bi-LSTM model

As a specific RNN model, the LSTM network is modeled by capturing the nonlinear relationship between the long-term input and output data. Through the feedback mechanism, the LSTM model can learn from the data in the past and establish the expected model of the past and current data through the multiple gate-structured network architecture. Recently, an increasing number of researchers have used LSTM

models in air quality research. This is because they can learn from historical air quality data and have achieved excellent performance in air quality prediction and air pollution driver analysis. For example, Han et al. (2021) used an LSTM model to evaluate the effects of air pollution control regulations in Beijing, China. Zhang et al. (2022) used a modified LSTM model to accurately predict air pollutants. Wang et al. (2021) used an LSTM model to obtain a high-resolution PM_{2.5} distribution in China. However, we found that the traditional RNN and LSTM network that transfer information from front to back have limitations in many studies. For example, in the study of the driving factors behind air pollution, future air quality data must be considered because of the hysteresis effect among pollutants. To solve this problem, two LSTM networks (forward and reverse) were designed, which are designated as the Bi-LSTM model. The idea is to connect the same input sequence to the two forward and successive LSTMs and then connect the hidden layers of the two networks to access the output layer for prediction.

Typically, each LSTM is controlled by a control unit C_t at time t . Three gates control access to the control unit: a forget gate f_t , input gate i_t , and output gate o_t . The forget gate decides whether to retain the current learning information. The input gate specifies the addition of new information to the learning process. The output gate controls whether the current learning information benefits from the output. Eqs. (S2)–(S7) represent the state update of the LSTM.

As shown in Fig. 2(a) the Bi-LSTM model extends the LSTM model described above, which consists of a forward layer and backward layer as a two-layer LSTM network. Learning long-term dependencies and fully considering the impact of future data on current data can improve the accuracy of the model through forward and reverse contributions to the input data. It is crucial to clarify the conceptual role of the Bi-LSTM in this framework. The model's bidirectional processing of the time series is not intended to simulate the physical direction of transport (i.e., upwind vs. downwind). Instead, its purpose is to learn the most complete and accurate representation of the regional system's temporal dynamics by considering the full context of the historical data. A more accurate predictive model of the entire system provides a more reliable foundation for the subsequent counterfactual sensitivity analysis used to quantify inter-city influence.

3.2 Evaluation and analysis methods

To accurately identify and quantify the regional O₃

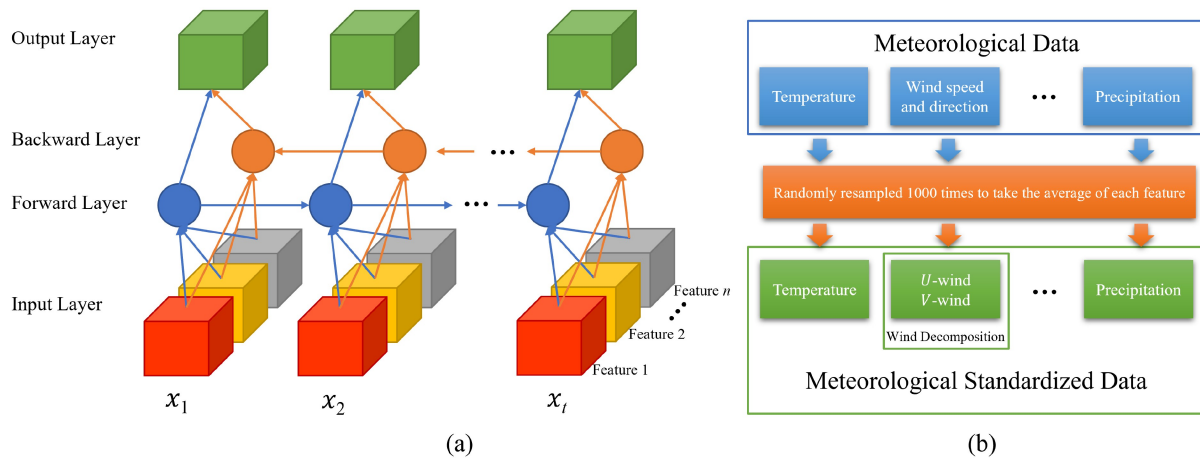


Fig. 2 (a) Bi-LSTM model. The $x_1, x_2, x_3, \dots, x_{\Delta t}$ are the data input in Δt period. The Features 1 to n are the feature variables input in each period. (b) Flowchart of the meteorological removal method.

transport pollution, we used three evaluation indicators: R^2 , mean absolute error (MAE), and root mean square error (RMSE) to evaluate the model's performance.

The MAE is a measure of the error between the predicted and actual values. The smaller the value, the higher is the accuracy of the model as shown in Eq. (S8).

The RMSE was used to measure the error between the target and predicted values as shown in Eq. (S9).

To gain a preliminary understanding of the O_3 transport relationship within the Hangzhou Bay urban agglomeration, we used the Pearson moment correlation coefficient (PMC) to analyze the O_3 concentration monitoring data to determine its correlation to other urban O_3 concentration monitoring data as described in Eq. (S10).

3.3 Meteorological removal

The sources of O_3 pollution are mainly composed of three factors: local emissions, chemical reactions, and regional transport (Zhang et al., 2022). Several studies have shown that extremely high O_3 concentrations are mainly due to regional O_3 transport in the residual layer (RL) and diurnal variations in the atmospheric boundary layer. Changes in meteorological conditions redistribute the entrained ozone-rich upper air to the ground evenly (Hu et al., 2018). Controlling meteorological conditions to observe changes in air quality over time can help to analyze the sources of air pollution (Grange et al., 2018). Removing the effects of meteorological changes from the simulations provides more certainty that the observed changes in O_3 concentration trends are due to regional transport. Therefore, we utilized a meteorological removal

method to eliminate its impact on O_3 pollution, thereby identifying and quantifying the regional O_3 pollution transport.

Machine-learning models can determine the impact of meteorological variables on O_3 pollution. Models can use meteorological removal methods to make new predictions by comparing them with observations to determine the transport concentration of O_3 pollution. Figure 2(b) shows a flow chart of the meteorological removal method. Our method's new dataset of input predictor features contains meteorological parameters (air temperature, dew point temperature, sea level pressure, wind direction, wind speed, cloud cover, and precipitation). First, parameters are randomly generated (i.e., resampled) from the original observation dataset. For example, for a specific hour of data (e.g., June 1, 2020, 0:00), the model randomly selects time series and corresponding meteorological parameters from the predicted features dataset throughout the study period. This is repeated 1000 times to provide a new input dataset for a specific time. The arithmetic mean of 1000 samples is then calculated as a new value for each feature, which is fed into the Bi-LSTM model to predict pollutant concentrations at a specific time. This sampling method forms the basis of a data-driven counterfactual analysis. The core principle is that by averaging the model's predictions over a large ensemble of randomly selected meteorological conditions, the influence of any single, specific weather pattern is effectively neutralized. This process generates a prediction for the O_3 concentration under climatologically normal weather conditions for the study period. It is important to note that this approach does not simulate the physical transport of air parcels. Instead, it quantifies the statistical influence of a source

city's atmospheric state (including pollutants and meteorology) on a target city's O_3 concentration. Therefore, the difference between the observed O_3 concentration and this meteorologically-neutral prediction is interpreted as the magnitude of the source city's influence, which arises from the impact of the actual, specific meteorological conditions that drive regional transport (Vu et al., 2019).

Because of uncertainty about how machine learning works, explaining the drivers of air pollution through machine learning is complex (Hou et al., 2022). Wind speed and direction have essential effects on the regional O_3 transport. Some researchers have noted that analyzing the wind direction components can facilitate locating the source of pollution (Liu et al., 2017). Therefore, we decomposed the wind direction and wind speed into U and V components, as the input features of the model.

3.4 Backward trajectory and potential source contribution function (PSCF) analysis

The hybrid single-particle Lagrangian integrated trajectory model (HYSPLIT) was used to compute the backward trajectory. The HYSPLIT calculation employs a hybrid Lagrangian method using a moving frame of reference for convection and diffusion calculations, as the trajectory, or the air package, moves from its initial position. However, the Euler method uses a fixed three-dimensional grid as a reference frame for calculating air concentrations of pollutants. Backward trajectories were calculated using meteorological field data (one-degree resolution, global) from the Global Data Assimilation System (GDAS), and the 24-h backward trajectory of the air mass was calculated as reaching 0.5 km above ground level in Shanghai.

PSCF is a gridded statistical analysis method based on the backward trajectory model, which is used to identify and calculate the source area of air pollution. This method can reflect the contribution of the source area to the receptor's degree of contamination according to the grid color. The calculation formula is shown in Eq. (1).

$$PSCF_{ij} = \frac{m_{ij}}{n_{ij}}, \quad (1)$$

where m_{ij} is the number of pollution trajectories passing through grid ij , and n_{ij} is the number of all trajectories in this grid. The higher the PSCF value, the greater the possibility of potential sources of pollutants at the receptor site.

To reduce the uncertainty of the grid PSCF results, a

weight function W was introduced, which is defined as shown in Eq. (2).

$$W_{ij} = \begin{cases} 1.0 & n_{ij} > 3.0n_{avg} \\ 0.71.5 & 0.71.5n_{avg} < n_{ij} \leq 3.0n_{ij} \\ 0.41.0 & 0.41.0n_{avg} < n_{ij} \leq 1.5n_{ij} \\ 0.2 & 0.2n_{ij} \leq 1.0n_{avg} \end{cases} \quad (2)$$

The resulting metric is the Weighted Potential Source Contribution Function (WPSCF), which is defined as shown in Eq. (3).

$$WPFC_{ij} = PFC_{ij} \times W_{ij}. \quad (3)$$

Cluster analysis has been demonstrated as effective in analyzing air pollution transport (Targino et al., 2019). Cluster analysis was performed using the K-means algorithm in MeteInfo software (v. 3.0.2) based on the Euclidean distance of the backward trajectories. Since there are 7 major cities in Hangzhou Bay, the "Cluster number" was set to 7, and the O_3 concentration bound to the trajectory was obtained from CNEMC. The results of the backward trajectory and cluster analysis were obtained according to Eqs. (1)–(3) to acquire the results of WPSCF. These results were then used to evaluate the accuracy of the Bi-LSTM model for quantitative identification of O_3 transport.

4 Results and discussion

4.1 Characteristics of O_3 pollution in Hangzhou Bay

We studied the characteristics of O_3 pollution in the Hangzhou Bay urban agglomeration. Fig. 3(a) shows the PMC of O_3 in the Hangzhou Bay area. All correlation coefficients of the 2020 annual O_3 concentration monitoring data for each city exceeded 0.57, with the highest value being close to 0.9. Among them, Zhoushan is an island city close to Ningbo and far from Hangzhou Bay. Therefore, the O_3 concentration in Zhoushan was less well correlated with that in other cities except for Ningbo. By contrast, Jiaying is close to the center of Hangzhou Bay. Its O_3 concentration was more than 0.8 correlated with other cities except for Zhoushan. As the two most economically developed cities in Hangzhou Bay, the O_3 concentrations in Shanghai and Hangzhou were highly correlated with other cities. Therefore, as some researchers have demonstrated, regional differences in O_3 pollution can be attributed to differences in economic size and geographic location (Cheng et al., 2018; Zheng et al., 2022).

According to the O_3 limit of Chinese National Ambient Air Quality Standards, an hourly

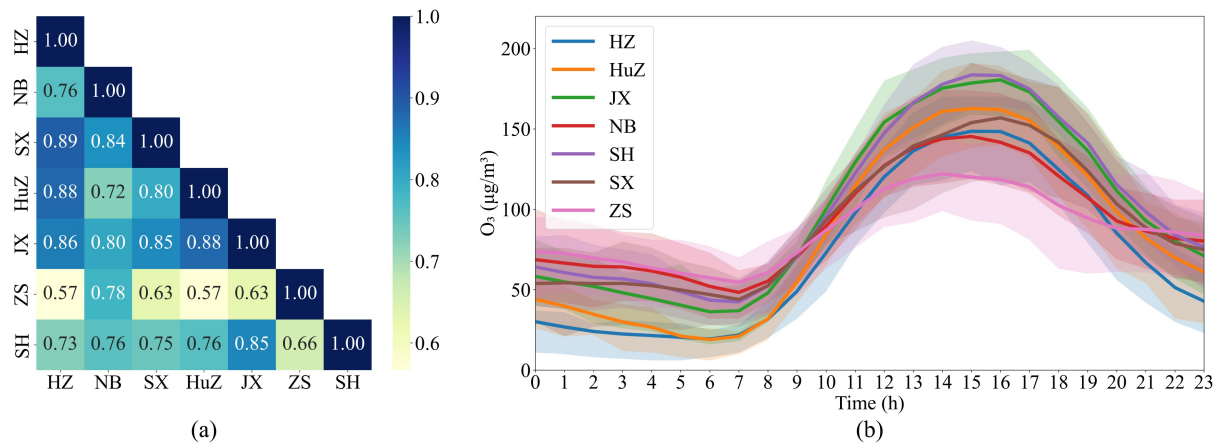


Fig. 3 Characteristics of O₃ pollution in the Hangzhou Bay urban agglomeration. (a) Heatmap showing the Pearson correlation coefficients of hourly O₃ concentrations between the seven cities in 2020. (b) Diurnal variation of O₃ concentrations on pollution days. An “O₃ pollution day” is defined as a day when the maximum 8-h average O₃ concentration exceeds the national standard of 160 µg/m³. The solid line represents the hourly mean concentration on these days, and the shaded area indicates the range between the 25th and 75th percentiles.

concentration exceeding 160 µg/m³ is regarded as nonattainment and the day is defined as an official Pollution Day. We selected the 2020 Hangzhou Bay Pollution Day data to study the characteristics of O₃ pollution in Hangzhou Bay. Figure 3(b) shows the 24-h concentration distribution of O₃ pollution days in the Hangzhou Bay area in 2020. The solid line represents the hourly mean O₃ concentration on pollution days, and shading indicates the O₃ concentration between the 25th and 75th percentiles. The O₃ concentration in the Hangzhou Bay area gradually increased after 7:00 am and reached its highest concentration from 2:00 to 5:00 p.m. Shanghai had the highest average O₃ concentration in one day, and individual pollution days exceeded 200 µg/m³. As Jiaxing is the closest to Shanghai, the O₃ concentration curve in Jiaxing was similar to that of Shanghai. Huzhou, Hangzhou, Shaoxing, and Ningbo had similar daily change curves. Zhoushan had the lowest GDP in Hangzhou Bay and the principle inland region, and it also had the lowest O₃ pollution level. Thus, the daily change in O₃ pollution concentration in Hangzhou Bay is affected by differences in many factors, showing different pollution characteristics.

Differences in air pollution emission levels, meteorological conditions, and photochemical reactions result from different economic and industrial contributions and geographic details. This is consistent with the conclusions of previous studies (Blanchard et al., 2014; Cheng et al., 2018). Although the economic output of Hangzhou Bay is enormous, the difference in GDP between different cities is also large, and the influence of different geographical factors among different cities has created a complex O₃ pollution

pattern in the Hangzhou Bay urban agglomeration. Regional transport is an essential source of O₃ pollution, and it is imperative to study the O₃ transport relationship between them.

4.2 Performance validation of the Bi-LSTM model

Using a model to predict outcomes is one way to identify whether a model meets expectations. Therefore, we validated the model by validating the accuracy of the Bi-LSTM model for O₃ pollution concentration prediction. Table 1 shows the performance of the Bi-LSTM model, RF and LSTM in CVs 1–7. The Bi-LSTM model outperforms RF and LSTM in all CV experiments in these results. In the 1-hour O₃ prediction performance of the Bi-LSTM model, the coefficient of determination (R^2) was 0.94, MAE was 6.71 µg/m³, and RMSE was 8.70 µg/m³. In the 6-h O₃ prediction performance, R^2 was 0.79, MAE was 12.79 µg/m³, and RMSE was 16.94 µg/m³. As other researchers have concluded, the Bi-LSTM model can learn long-term information in data from both positive and negative directions and accurately identify the impact of different feature variables on the results (Siami-Namini et al., 2019).

Figure 4 shows the scatter plot of the prediction results for different time periods. The forecast performance decreased as the forecast time increased, but the overall results were satisfactory. In addition, it can be seen from Fig. 4 that individually, the more significant errors in extremely high-value weather may be related to the fact that there are too few high-value weather samples. The model did not have enough

Table 1 Performance comparison of models using cross-validation and independent-validation

Model (metric)	CV 1	CV 2	CV 3	CV 4	CV 5	CV 6	CV 7	IV
Bi-LSTM (R^2)	0.92	0.92	0.93	0.95	0.94	0.94	0.95	0.93
RF (R^2)	0.83	0.79	0.86	0.86	0.86	0.84	0.81	0.80
LSTM (R^2)	0.88	0.91	0.90	0.87	0.89	0.86	0.90	0.90
Bi-LSTM (MAE)	6.18	7.31	7.07	6.20	7.88	7.99	6.68	7.22
RF (MAE)	10.19	11.37	9.24	9.43	9.34	9.69	10.73	11.06
LSTM (MAE)	9.30	8.39	8.76	9.46	8.86	8.18	8.50	8.61
Bi-LSTM (RMSE)	8.90	10.30	10.85	8.54	10.98	11.04	9.08	10.53
RF (RMSE)	14.71	16.29	13.63	13.81	13.72	14.17	15.45	15.38
LSTM (RMSE)	12.92	11.71	12.22	13.22	12.37	11.56	12.10	11.98

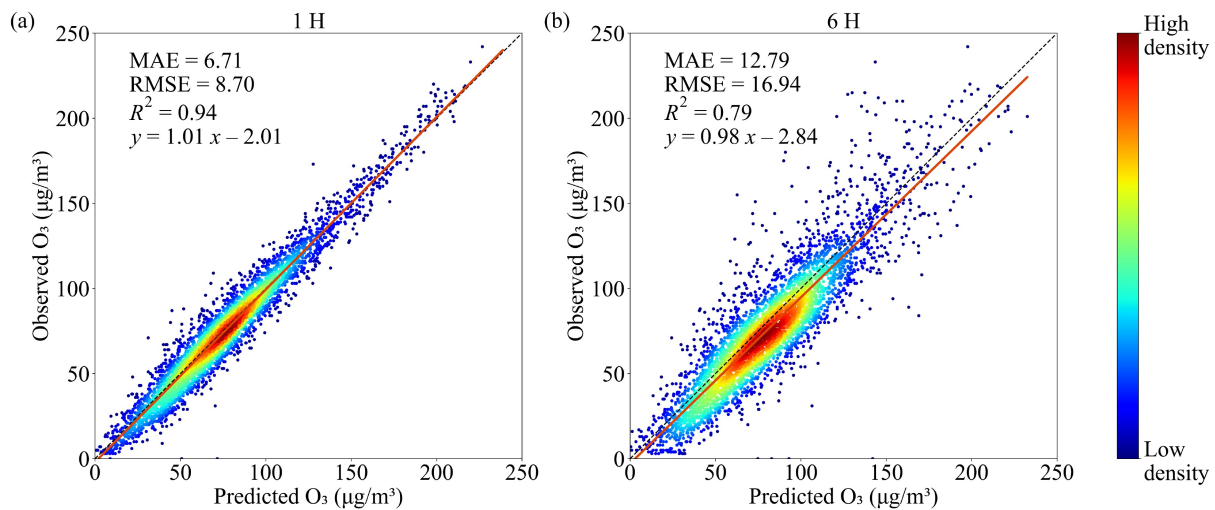


Fig. 4 Yearly validation results of the Bi-LSTM model. (a) 1-h O_3 prediction and (b) 6-h O_3 prediction scatter plots. The horizontal axis shows the O_3 concentrations predicted, and the vertical axis shows the O_3 concentrations observed. Dotted lines in scatter plots represent the 1:1 reference line, and red lines represent the observed and predicted regression line.

samples to understand the reasons for the high value of O_3 pollution, which is similar to the point made by some researchers (Wang et al., 2021). The intercept of the regression equation in the figure is negative, indicating that the model tends to predict a lower value than is observed, which is similar to the research conclusion of previous studies (Fang et al., 2016), but this does not affect the overall model prediction accuracy. If a model can predict accurately, it can identify the nonlinear relationship between the characteristic variables and the predicted results. This study represents the impact of changes in air pollutant concentrations and meteorological data on the final O_3 concentration prediction results over a period when the Bi-LSTM model was applied to different cities. In addition to cross-validation, the model's temporal robustness was evaluated on the completely independent 2021 dataset (Table 1). The Bi-LSTM model,

trained only on 2020 data, achieved an R^2 of 0.93, an MAE of $7.22 \mu\text{g}/\text{m}^3$, and an RMSE of $10.53 \mu\text{g}/\text{m}^3$ when tested on the 2021 data. These results, while showing a slight and expected decrease in performance compared to the validation on the training year, remain excellent. This demonstrates the model's strong ability to generalize to a different year, confirming that it has learned the underlying physical relationships rather than simply overfitting to the data of a single year.

4.3 Quantifying inter-city O_3 influence via sensitivity analysis

Before presenting the city-specific transport quantification, we first evaluated the overall impact and stability of our meteorological removal method. A comparison of the statistical distributions of the original observed O_3 concentrations and the final meteor-

ologically-neutral predictions for the entire month of June reveals a clear and systematic reduction in ozone levels (Fig. S1). This result demonstrates the robustness of the methodology in isolating the influence of meteorological conditions, providing confidence in the subsequent quantification of transport contributions. In general, relying on the excellent performance of the Bi-LSTM model, we can identify and quantify O_3 transport by observing the changes in O_3 concentration by changing the variables related to O_3 transport in the study. To better study O_3 transport, we simulated transport in June 2020, which is the first month of summer in Hangzhou Bay, when extremely high-value O_3 weather occurs but is not yet frequent. Therefore, the machine-learning model can better validate the identification and quantification of O_3 transport at different pollution levels. Shanghai has the highest GDP, most advanced industrial level, and largest population in the Hangzhou Bay urban agglomeration. It represents the most critical research value in Hangzhou Bay; therefore, we consider Shanghai as the city simulated by the Bi-LSTM model using other cities as model variables.

Figure 5 shows the simulation results for O_3 transport from Hangzhou and Ningbo to Shanghai. The blue-green shading represents the observed O_3 concentration in Shanghai, and the red shading represents the O_3 concentration in Shanghai after meteorological removal simulated by the machine-learning model. Fig. 5(a) shows that after removing the O_3 transport from Hangzhou to Shanghai, the O_3 concentration in Shanghai dropped by approximately $9.57 \mu\text{g}/\text{m}^3$. This was especially evident in high- O_3 weather, and the O_3 concentration on non-pollution days did not change significantly. This suggests that some O_3 pollution in Shanghai originates from the transport of Hangzhou. The simulation results from Ningbo to Shanghai, as presented in Fig. 5(b), are the opposite. Removing the O_3 transport from Ningbo to Shanghai, the O_3 concentration in Shanghai did not change significantly in high- O_3 weather. The O_3 concentration on non-polluted days increased slightly, indicating that Ningbo has a weak transport relationship with Shanghai. Some O_3 pollution in Ningbo comes from transport in Shanghai, which increases the O_3 concentration in Shanghai by approximately $11.79 \mu\text{g}/\text{m}^3$.

Figures S2–S5 show the simulation results of O_3 transport from Shaoxing to Shanghai, Huzhou to Shanghai, Jiaying to Shanghai, and Zhoushan to Shanghai, respectively. Because Shaoxing, Huzhou, Jiaying, and Hangzhou have similar geographical locations and are in the same wind direction, their transport relationships are similar to that of Hangzhou.

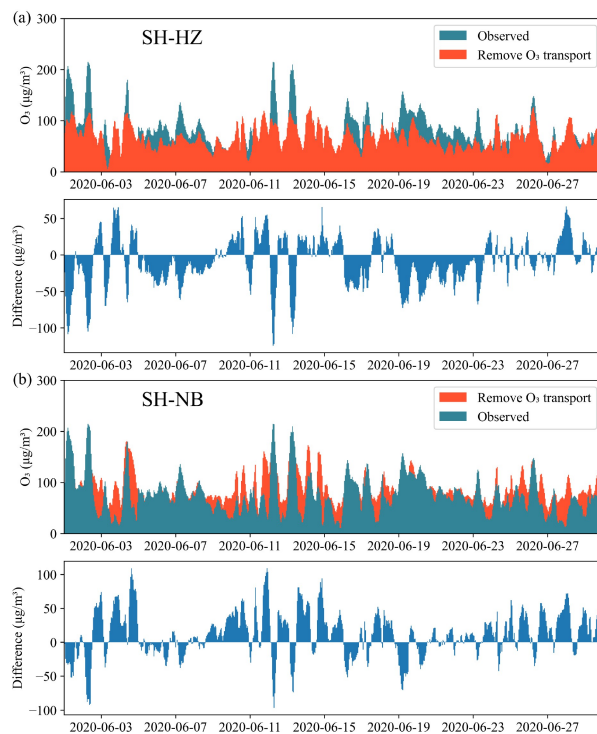


Fig. 5 Quantification of inter-city O_3 transport to/from Shanghai in June 2020. (a) Transport from Hangzhou to Shanghai. (b) Transport between Ningbo and Shanghai. The blue-green area represents the observed hourly O_3 concentrations in Shanghai. The red line represents the simulated O_3 concentrations after applying the meteorological removal method to the source city's data. The dark blue bars represent the calculated contribution of transport, which is the difference between the observed and simulated concentrations. A negative value indicates transport from the source city to Shanghai.

As a result, more transport occurs during high- O_3 weather and is not apparent on non-polluted days.

Notably, Jiaying, the city closest to Shanghai, had the highest O_3 concentration throughout Hangzhou Bay area to Shanghai, which increased the O_3 concentration in Shanghai by approximately $11.56 \mu\text{g}/\text{m}^3$. As Huzhou and Shaoxing are far from Shanghai, they increased the O_3 in Shanghai by 5.13 and $0.89 \mu\text{g}/\text{m}^3$, respectively. This shows that distance also affects the O_3 transport. Zhoushan is in the sea east of Shanghai, China. Therefore, it is affected by the East Asian summer monsoon and will also affect the O_3 concentration in Shanghai (Li et al., 2018b). Our simulation results showed that Zhoushan delivered an average of $5.16 \mu\text{g}/\text{m}^3$ of O_3 concentration to Shanghai. Table 2 shows the average impact and percentage of O_3 transport in Shanghai by the six cities in Hangzhou Bay and the change in the values at 90% and 75% (P90 and P75). P90 can evaluate the pollution level of the weather with a high O_3 value, and P75 can evaluate the pollution

Table 2 Simulation results of the model

City	Average change ($\mu\text{g}/\text{m}^3$)	Average change (%)	P90 ($\mu\text{g}/\text{m}^3$)	P75 ($\mu\text{g}/\text{m}^3$)
HZ	-9.57	-12.43	-25.47	-10.80
Huz	-5.13	-6.67	-24.94	-7.75
JX	-11.56	-15.02	-27.76	-15.34
NB	11.79	15.32	0.37	6.54
SX	-0.89	-1.16	-15.67	-1.72
ZS	-5.16	-6.71	-23.72	-9.29

Notes: City abbreviations are as follows: HZ (Hangzhou), Huz (Huzhou), JX (Jiaxing), NB (Ningbo), SX (Shaoxing), ZS (Zhoushan). A negative average change indicates transport from the specified city to Shanghai, while a positive change indicates transport from Shanghai to the specified city.

level of the weather with a high O_3 value. The changes in P90 in Hangzhou, Huzhou, Jiaxing, Shaoxing, and Zhoushan were significantly greater than those in P75. This shows that Shanghai is more obviously affected by transport from other regions during high O_3 value weather. While the model's predictions carry higher uncertainty during these extreme pollution events (Fig. 4), this conclusion is robust. The large difference between P90 and P75 contributions is a strong signal, and it is corroborated by both the independent WPSCF analysis and our comparative analysis for a low-pollution month (Fig. S6), which shows minimal transport effects. Together, these results strongly support the conclusion that regional transport is a dominant factor on high-pollution days. Ningbo is the only city in the Hangzhou Bay area that has O_3 transported from Shanghai. Its P90 was significantly lower than that of P75, indicating that the high-value weather in Shanghai was not transported from Ningbo. In general, Shanghai's high- O_3 weather is not entirely due to local emissions and chemical reactions, and regional transport contributes considerably.

Studying the regional O_3 transport is a way to both study the relationship between the two cities as well as the entire region. Therefore, we studied the effect throughout Hangzhou Bay on O_3 concentration in Shanghai by removing all meteorological data. Figure 6 shows the results of using a machine-learning model to simulate the effect of eliminating O_3 transport throughout the Hangzhou Bay area in Shanghai. Shanghai's O_3 would be relatively stable if it were not transported from the Hangzhou Bay urban agglomeration. Specifically, for high O_3 values, a significant proportion of O_3 pollution in Shanghai originates from transport from other regions. It is evident that Shanghai's O_3 would be relatively stable if it were not transported from the Hangzhou Bay urban agglomeration. This is consistent with the results obtained from a single-city simulation. To verify the robustness of our quantification, we compared two distinct simulation schemes to calculate the total regional

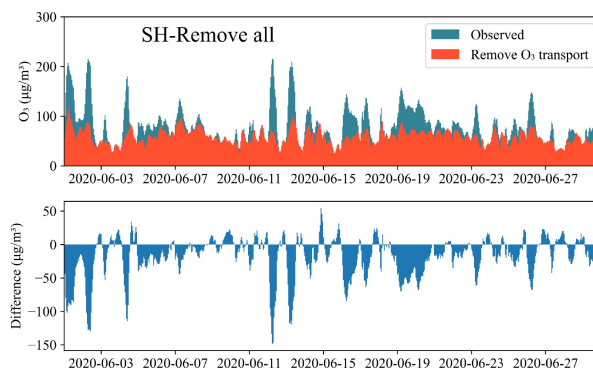


Fig. 6 Quantification of total O_3 transport from the entire Hangzhou Bay urban agglomeration to Shanghai in June 2020. The blue-green area represents the observed hourly O_3 concentrations in Shanghai. The red line represents the simulated O_3 concentrations after applying the meteorological removal method to all other cities in the region simultaneously. The dark blue bars represent the total transport contribution.

influence on Shanghai (Table 3). The first, a “holistic approach”, involved simultaneously removing the meteorological influences from all six neighboring cities, which resulted in a predicted O_3 reduction of $18.41 \mu\text{g}/\text{m}^3$ (24% contribution). The second, an “additive approach”, involved summing the individually calculated influences from the five contributing cities (excluding Ningbo, which was a net recipient), yielding a total calculated influence of $20.52 \mu\text{g}/\text{m}^3$ (27% contribution). The remarkable consistency between these two independent schemes, with a difference of only 3%, strongly validates our model's ability to reliably quantify inter-city O_3 influence.

In summary, the simulation results suggest that approximately 24% of Shanghai's O_3 pollution comes from regional transport, which is similar to the conclusions of some scholars using the chemical weather research and forecasting model (Hu et al., 2018). Through two simulation methods, we demonstrated that the Bi-LSTM model combined with meteorological removal can identify and quantify the regional O_3 transport. It can accurately capture O_3

Table 3 Comparison of the results of two-way verification

Method	O ₃ change	
	μg/m ³	%
Remove all meteorological	-18.41	-24
Add up individual results	-20.52	-27
Error	2.11	3

pollution drivers by learning from air quality data and meteorological data and is consistent with the conclusions of traditional models. This shows the potential of the machine-learning model combined with the meteorological elimination method in atmospheric pollutant transport research. The primary practical application of this model's computational efficiency lies in this retrospective analytical capability. While the Bi-LSTM's reliance on the full time series precludes its use for real-time forecasting, its speed enables the rapid execution of numerous sensitivity analyses—such as quantifying the influence of multiple cities individually and holistically—over long time periods. Performing such extensive scenario analysis with traditional CTMs would be computationally prohibitive. Therefore, our framework serves as a valuable diagnostic tool for policymakers to quantitatively assess the drivers of past pollution events and to provide data-driven support for the development of regional joint-control strategies.

The findings have significant implications for air quality management. The quantification of substantial O₃ transport into Shanghai, particularly during high-pollution episodes, underscores the limitations of city-level emission control strategies implemented in isolation. Our results provide data-driven evidence supporting the need for coordinated, regional joint-control policies across the entire Hangzhou Bay urban agglomeration to effectively mitigate severe O₃ pollution.

Furthermore, the methodological framework presented here has broad generalizability. The Bi-LSTM with meteorological removal approach is not specific to O₃ and can be readily adapted to quantify the transport of other pollutants like PM_{2.5} or NO₂, provided that long-term, high-frequency monitoring data are available. It is also geographically transferable to other urban clusters worldwide. However, it is crucial to distinguish the generalizability of the method from that of the results. The specific transport pathways and contributions quantified in this study are inherently context-specific to the Hangzhou Bay region, which is characterized by unique meteorological patterns and a dense, complex network of emission sources. Applying this method to other regions would likely yield different quantitative outcomes reflecting their local geography

and meteorology.

4.4 Analysis of WPSCF results

To better understand the potential sources of near-ground O₃, we used the backward trajectory model of the air mass throughout June 2020. Figure S7 shows the 24-h backward trajectory calculated every 6 h in Shanghai during the observation day. We found that part of the air mass in Shanghai comes from the East China Sea, where Zhoushan City is located, and part of the air mass comes from Hangzhou, Shaoxing, Huzhou, and other cities southwest of Shanghai. This suggests that these cities were affected by the backward trajectory of the air mass, which transported O₃ to Shanghai.

Figure 7 shows the results of the WPSCF, which reveals the spatial probability distribution of potential O₃ transport sources. Hangzhou, Huzhou, Shaoxing, Zhoushan, and Jiaxing are in areas with high spatial probabilities. This indicates that these cities contributed more to the O₃ concentration in Shanghai. Crucially, the results of the WPSCF, a method based on physical air mass trajectories, are in striking agreement with the statistical influence patterns identified by our data-driven Bi-LSTM model. This agreement provides important physical validation for our statistical findings. For instance, the WPSCF shows no significant potential source contribution from Ningbo, which independently corroborates the Bi-LSTM model's finding that Ningbo was a net recipient of O₃ from Shanghai rather than a source.

4.5 Limitations

This study demonstrates a robust data-driven

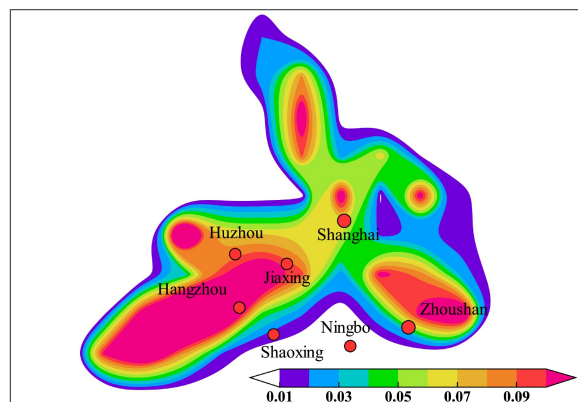


Fig. 7 Spatial distribution of WPSCF values for O₃ in Shanghai during June 1–30, 2020.

framework; however, its limitations highlight key areas for future research. The model's performance is inherently dependent on the training data, meaning its generalizability across different years or regions with varying emission and meteorological regimes would benefit from rigorous testing on temporally independent datasets and the establishment of recalibration protocols. Additionally, we acknowledge that the model's predictive accuracy is lower for extreme high-concentration events, likely due to the relative scarcity of such data. While our qualitative conclusions about the importance of transport on these days are robust, this introduces greater uncertainty into the precise quantification of transport contributions during peak pollution episodes.

From a methodological perspective, our approach also presents opportunities for future refinement. The meteorological removal method operates on the useful simplification that weather's influence can be disentangled from pollutant concentrations by resampling; future work could explore more advanced causal inference frameworks to better address these confounding effects. Furthermore, to move beyond the "black box" nature of the model and gain deeper physical insights, future research could integrate explainable artificial intelligence techniques. Methods such as SHapley Additive exPlanations (SHAP) can quantify the contribution of each input feature, providing a powerful, feature-level understanding of the drivers of O₃ transport and strengthening the role of data-driven models in air quality research.

5 Conclusions

The overall goal of this study was to validate the potential of machine-learning models in quantifying the inter-city influence of atmospheric pollutants. We used a Bi-LSTM model combined with meteorological removal to quantify the inter-city statistical influence of O₃. We used a two-way validation approach to simulate O₃ transport in the Hangzhou Bay urban agglomeration in both forward and reverse directions. The results showed that approximately 24% of the O₃ pollution in Shanghai came from other cities, including Hangzhou, Jiaxing, Shaoxing, Huzhou, and Zhoushan. In addition, the O₃ transport was mainly concentrated in the high-value weather of O₃ pollution in Shanghai, and transport on non-pollution days was not apparent. Therefore, the O₃ transport from other cities is of great significance for the formation of severe O₃ pollution in Shanghai. In addition, the transport relationship

obtained by the Bi-LSTM model is consistent with the potential O₃ transport source obtained by WPSCF. This study demonstrated the potential of machine-learning models combined with meteorological removal methods for quantifying the inter-city influence of atmospheric pollutants.

Conflict of Interests The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this study.

Acknowledgements This study was supported by the National Key R&D Program of China (No. 2022YFC3700400), the Shanghai Science and Technology Plan Project, China (No. 22SQBS01000), and the High-level University Special Fund, China (No. G03050K001).

Electronic Supplementary Material Supplementary material is available in the online version of this article at <https://dx.doi.org/10.1007/s11783-025-2089-1> and is accessible for authorized users.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Blanchard C L, Tanenbaum S, Hidy G M (2014). Spatial and temporal variability of air pollution in Birmingham, Alabama. *Atmospheric Environment*, 89: 382–391
- Cheng L J, Wang S, Gong Z Y, Li H, Yang Q, Wang Y Y (2018). Regionalization based on spatial and seasonal variation in ground-level ozone concentrations across China. *Journal of Environmental Sciences*, 67: 179–190
- Cohen A J, Brauer M, Burnett R, Anderson H R, Frostad J, Estep K, Balakrishnan K, Brunekreef B, Dandona L, Dandona R, et al. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *The Lancet*, 389(10082): 1907–1918
- Fang X, Zou B, Liu X P, Sternberg T, Zhai L (2016). Satellite-based ground PM_{2.5} estimation using timely structure adaptive modeling. *Remote Sensing of Environment*, 186: 152–163
- Feng R, Zheng H J, Gao H, Zhang A R, Huang C, Zhang J X, Luo K, Fan J R (2019). Recurrent Neural Network and random forest for analysis and accurate forecast of atmospheric pollutants: a case study in Hangzhou, China. *Journal of Cleaner Production*, 231:

- 1005–1015
- Foret G, Eremenko M, Cuesta J, Sellitto P, Barré J, Gaubert B, Coman A, Dufour G, Liu X, Joly M, et al. (2014). Ozone pollution: what can we see from space? A case study. *Journal of Geophysical Research: Atmospheres*, 119(13): 8476–8499
- Gao J H, Zhu B, Xiao H, Kang H Q, Hou X W, Shao P (2016). A case study of surface ozone source apportionment during a high concentration episode, under frequent shifting wind conditions over the Yangtze River Delta, China. *Science of the Total Environment*, 544: 853–863
- Gao W, Tie X X, Xu J M, Huang R J, Mao X Q, Zhou G Q, Chang L Y (2017). Long-term trend of O₃ in a Mega City (Shanghai), China: characteristics, causes, and interactions with precursors. *Science of the Total Environment*, 603–604: 425–433
- Grange S K, Carslaw D C, Lewis A C, Boleti E, Hueglin C (2018). Random forest meteorological normalisation models for Swiss PM₁₀ trend analysis. *Atmospheric Chemistry and Physics*, 18(9): 6223–6239
- Guan Y, Xiao Y, Wang Y M, Zhang N N, Chu C J (2021). Assessing the health impacts attributable to PM_{2.5} and ozone pollution in 338 Chinese cities from 2015 to 2020. *Environmental Pollution*, 287: 117623
- Hou L L, Dai Q L, Song C B, Liu B W, Guo F Z, Dai T J, Li L X, Liu B S, Bi X H, Zhang Y F, Feng Y C (2022). Revealing drivers of haze pollution by explainable machine learning. *Environmental Science & Technology Letters*, 9(2): 112–119
- Han Y, Lam J C K, Li V O K, Reiner D (2021). A Bayesian LSTM model to evaluate the effects of air pollution control regulations in Beijing, China. *Environmental Science & Policy*, 115: 26–34
- Hu J, Li Y C, Zhao T L, Liu J, Hu X M, Liu D Y, Jiang Y C, Xu J M, Chang L Y (2018). An important mechanism of regional O₃ transport for summer smog over the Yangtze River Delta in eastern China. *Atmospheric Chemistry and Physics*, 18(22): 16239–16251
- Jiang X Y, Wiedinmyer C, Carlton A G (2012). Aerosols from fires: an examination of the effects on ozone photochemistry in the western United States. *Environmental Science & Technology*, 46(21): 11878–11886
- Kang Q, Song X, Xin X Y, Chen B, Chen Y Z, Ye X D, Zhang B Y (2021). Machine learning-aided causal inference framework for environmental data analysis: a COVID-19 case study. *Environmental Science & Technology*, 55(19): 13400–13410
- Kim H S, Park I, Song C H, Lee K, Yun J W, Kim H K, Jeon M, Lee J, Han K M (2019). Development of a daily PM₁₀ and PM_{2.5} prediction system using a deep long short-term memory neural network model. *Atmospheric Chemistry and Physics*, 19(20): 12935–12951
- Li K, Jacob D J, Liao H, Shen L, Zhang Q, Bates K H (2018a). Anthropogenic drivers of 2013–2017 trends in summer surface ozone in China. *Proceedings of the National Academy of Sciences of the United States of America*, 116(2): 422–427
- Li S, Wang T J, Huang X, Pu X, Li M M, Chen P L, Yang X Q, Wang M H (2018b). Impact of East Asian summer monsoon on surface ozone pattern in China. *Journal of Geophysical Research: Atmospheres*, 123(2): 1401–1411
- Liu C, Zhang H R, Cheng Z, Shen J Y, Zhao J H, Wang Y C, Wang S, Cheng Y (2021a). Emulation of an atmospheric gas-phase chemistry solver through deep learning: case study of Chinese Mainland. *Atmospheric Pollution Research*, 12(6): 101079
- Liu X, Lu D W, Zhang A Q, Liu Q, Jiang G B (2022). Data-driven machine learning in environmental pollution: gains and problems. *Environmental Science & Technology*, 56(4): 2124–2133
- Liu X F, Guo H, Zeng L W, Lyu X P, Wang Y, Zeren Y Z, Yang J, Zhang L Y, Zhao S Z, Li J, Zhang G (2021b). Photochemical ozone pollution in five Chinese megacities in summer 2018. *Science of the Total Environment*, 801: 149603
- Liu X L, Huang W M, Gill E W (2017). Wind direction estimation from rain-contaminated marine radar data using the ensemble empirical mode decomposition method. *IEEE Transactions on Geoscience and Remote Sensing*, 55(3): 1833–1841
- Lu X, Zhang L, Chen Y F, Zhou M, Zheng B, Li K, Liu Y M, Lin J T, Fu T M, Zhang Q (2019). Exploring 2016–2017 surface ozone pollution over China: source contributions and meteorological influences. *Atmospheric Chemistry and Physics*, 19(12): 8339–8361
- Pak U, Ma J, Ryu U, Ryom K, Juhyok U, Pak K, Pak C (2020). Deep learning-based PM_{2.5} prediction considering the spatiotemporal correlations: a case study of Beijing, China. *Science of the Total Environment*, 699: 133561
- Qi L, Tian Z G, Jiang N, Zheng F Y, Zhao Y C, Geng Y S, Duan X L (2023). Collaborative control of fine particles and ozone required in China for health benefit. *Frontiers of Environmental Science & Engineering*, 17(8): 92
- Qu Z W, Li H T, Li Z H, Zhong T (2022). Short-term traffic flow forecasting method with M-B-LSTM hybrid network. *IEEE Transactions on Intelligent Transportation Systems*, 23(1): 225–235
- Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743): 195–204
- Roberts D R, Bahn V, Ciuti S, Boyce M S, Elith J, Guillera-Aroita G, Hauenstein S, Lahoz-Monfort J J, Schröder B, Thuiller W, et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8): 913–929
- Siami-Namini S, Tavakoli N, Namin A S (2019). The performance of LSTM and BiLSTM in forecasting time series. In: *Proceedings of 2019 IEEE International Conference on Big Data (Big Data)*. Los Angeles: IEEE, 3285–3292
- Sicard P, Agathokleous E, De Marco A, Paoletti E, Calatayud V (2021). Urban population exposure to air pollution in Europe over the last decades. *Environmental Sciences Europe*, 33(1): 28
- Song C B, Wu L, Xie Y C, He J J, Chen X, Wang T, Lin Y C, Jin T S, Wang A X, Liu Y, et al. (2017). Air pollution in China: status and spatiotemporal variations. *Environmental Pollution*, 227:

334–347

- Suciu L G, Griffin R J, Masiello C A (2017). Regional background O₃ and NO_x in the Houston-Galveston-Brazoria (TX) region: a decadal-scale perspective. *Atmospheric Chemistry and Physics*, 17(11): 6565–6581
- Tao Q, Liu F, Li Y, Sidorov D (2019). Air pollution forecasting using a deep learning model based on 1D convnets and bidirectional GRU. *IEEE Access*, 7: 76690–76698
- Targino A C, Harrison R M, Krecl P, Glantz P, de Lima C H, Beddows D (2019). Surface ozone climatology of South Eastern Brazil and the impact of biomass burning events. *Journal of Environmental Management*, 252: 109645
- Tong X N, You L H, Zhang J J, He Y L, Gin K Y H (2022). Advancing prediction of emerging contaminants in a tropical reservoir with general water quality indicators based on a hybrid process and data-driven approach. *Journal of Hazardous Materials*, 430: 128492
- Vu T V, Shi Z B, Cheng J, Zhang Q, He K B, Wang S X, Harrison R M (2019). Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique. *Atmospheric Chemistry and Physics*, 19(17): 11303–11314
- Wang B, Yuan Q Q, Yang Q Q, Zhu L Y, Li T W, Zhang L P (2021). Estimate hourly PM_{2.5} concentrations from Himawari-8 TOA reflectance directly using geo-intelligent long short-term memory network. *Environmental Pollution*, 271: 116327
- Wang T, Xue L K, Brimblecombe P, Lam Y F, Li L, Zhang L (2017). Ozone pollution in China: a review of concentrations, meteorological influences, chemical precursors, and effects. *Science of the Total Environment*, 575: 1582–1596
- Wang T Y, Zhao B, Liou K N, Gu Y, Jiang Z, Song K, Su H, Jerrett M, Zhu Y F (2019). Mortality burdens in California due to air pollution attributable to local and nonlocal emissions. *Environment International*, 133: 105232
- Watson G L, Telesca D, Reid C E, Pfister G G, Jerrett M (2019). Machine learning models accurately predict ozone exposure during wildfire events. *Environmental Pollution*, 254: 112792
- Wei J, Liu S, Li Z Q, Liu C, Qin K, Liu X, Pinker R T, Dickerson R R, Lin J T, Boersma K F, et al. (2022). Ground-Level NO₂ surveillance from space across China for high resolution using interpretable spatiotemporally weighted artificial intelligence. *Environmental Science & Technology*, 56(14): 9988–9998
- Xue J, Wang F T, Zhang K, Zhai H H, Jin D, Duan Y S, Yaluk E, Wang Y J, Huang L, Li Y W, et al. (2023). Elucidate long-term changes of ozone in Shanghai based on an integrated machine learning method. *Frontiers of Environmental Science & Engineering*, 17(11): 138
- Xue L K, Wang T, Gao J, Ding A J, Zhou X H, Blake D R, Wang X F, Saunders S M, Fan S J, Zuo H C, et al. (2014). Ground-level ozone in four Chinese cities: precursors, regional transport and heterogeneous processes. *Atmospheric Chemistry and Physics*, 14(23): 13175–13188
- Zhang K, Zhou L, Fu Q Y, Yan L, Bian Q G, Wang D F, Xiu G L (2019). Vertical distribution of ozone over Shanghai during late spring: a balloon-borne observation. *Atmospheric Environment*, 208: 48–60
- Zhang Y B, Yu S C, Chen X, Li Z, Li M Y, Song Z, Liu W P, Li P F, Zhang X Y, Lichtfouse E, et al. (2022). Local production, downward and regional transport aggravated surface ozone pollution during the historical orange-alert large-scale ozone episode in eastern China. *Environmental Chemistry Letters*, 20(3): 1577–1588
- Zhao D D, Xin J Y, Wang W F, Jia D J, Wang Z F, Xiao H, Liu C, Zhou J, Tong L, Ma Y J, et al. (2022). Effects of the sea-land breeze on coastal ozone pollution in the Yangtze River Delta, China. *Science of the Total Environment*, 807: 150306
- Zhao H, Chen K Y, Liu Z, Zhang Y X, Shao T, Zhang H L (2021). Coordinated control of PM_{2.5} and O₃ is urgently needed in China after implementation of the “Air pollution prevention and control action plan”. *Chemosphere*, 270: 129441
- Zheng D Y, Huang X J, Guo Y H (2022). Spatiotemporal variation of ozone pollution and health effects in China. *Environmental Science and Pollution Research*, 29(38): 57808–57822
- Zhong S F, Zhang K, Bagheri M, Burken J G, Gu A, Li B K, Ma X M, Marrone B L, Ren Z J, Schrier J, et al. (2021). Machine learning: new ideas and tools in environmental science and engineering. *Environmental Science & Technology*, 55(19): 12741–12754
- Zhu Q Y, Bi J Z, Liu X, Li S S, Wang W H, Zhao Y, Liu Y (2022). Satellite-based long-term spatiotemporal patterns of surface ozone concentrations in China: 2005–2019. *Environmental Health Perspectives*, 130(2): 027004
- Zhu Y, Liu J, Wang T J, Zhuang B L, Han H, Wang H M, Chang Y, Ding K (2017). The impacts of meteorology on the seasonal and interannual variabilities of ozone transport from North America to East Asia. *Journal of Geophysical Research: Atmospheres*, 122(20): 10612–10636