Environmental Modelling and Software 190 (2025) 106466

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft

Interpolation of missing ozone data using graph machine learning and parameter analysis through eXplainable artificial intelligence comparison

Seonje Jung [a], Junsu Gil [a,b,*], Meehye Lee [a], Clara Betancourt [c], Martin Schultz [d], Yunsoo Choi [e], Taekyu Joo [a], Daigon Kim [f]

[a] Department of Earth and Environmental Sciences, Korea University, Seoul, 02841, South Korea
[b] Institute of Basic Science, Korea University, Seoul, 02841, Republic of Korea
[c] AXA Konzern AG, Colonia-Allee 10-20, 51067, Köln, Germany
[d] Jülich Supercomputing Centre, Jülich Research Centre, Wilhelm-Johnen-Straße, 52425, Jülich, Germany
[e] Department of Earth and Atmospheric Sciences, University of Houston, TX, 77004, USA
[f] Division of Atmospheric Environmental Research, National Institute of Environmental Research, Incheon, South Korea

ARTICLE INFO

Keywords:
Graph machine learning
Missing data interpolation
Spatiotemporal ozone characteristics
Climate change
Surface ozone
Explainable machine learning

ABSTRACT

Ozone (O_3), a short-lived climate pollutant, continues to increase despite policies aimed at suppressing its precursors in South Korea. The government operates approximately 500 observatories to monitor O_3 and trace gases. Researchers use these data to address the ongoing issue of increasing O_3 levels. However, challenges in data retrieval from observatories may introduce biases in O_3 studies. In this study, we developed a graph-based machine learning model to simulate missing O_3 concentrations for mitigate bias. The model incorporates spatiotemporal distribution characteristics using a merged observation dataset from South Korea in 2021. Regardless of region or length of missing data, the model effectively simulates O_3 variations with R^2 of up to 0.9 and RMSE of 3.6. To determine the influence of input parameters on O_3 interpolation, we used eXplainable AI methods. The results indicated that NO_2 is the most important factor in cities, while photochemical indicators are more influential in provinces.

1. Introduction

Ozone (O_3) is a major atmospheric oxidant that is toxic to vegetation and poses significant public health risks, including respiratory and cardiovascular diseases (Agathokleous et al., 2018; Mills et al., 2018; Nuvolone et al., 2018). To address the issues caused by O_3, each country has established its own criteria for issuing alerts regarding elevated O_3 levels. In South Korea, the national O_3 threshold is 120 ppbv, and an ozone alert is issued when predicted O_3 levels exceed this threshold. Since the 2010s, South Korea has experienced rising O_3 levels, driven by increasing temperatures, with O_3 levels in most regions exceeding the national standard, particularly during the spring and summer months (Schultz et al., 2017). To manage the risks associated with O_3, the Ministry of Environment of South Korea has established approximately 500 air quality monitoring stations, collectively known as the Korea Air Quality Network, since the 1970s. These stations continuously monitor O_3 levels, along with those of its photochemical precursors, NO_x and volatile organic compounds (VOCs) (Schultz et al., 2017). Although these monitoring stations typically provide high spatial-resolution, real-time air quality data, instances of missing data have increased owing to instrument malfunctions, data processing issues, and other factors. Such missing data can introduce bias into statistical analyses and reports, posing challenges for O_3-related studies (Davison and Hemphill, 1987; Schultz et al., 2017).

Specifically, missing data points pose a significant challenge for research examining the impact of atmospheric pollutants, such as variation of average O_3 on human health. Consequently, previous studies have attempted to estimate missing atmospheric pollutant data using statistical techniques, such as linear regression and machine learning (ML). Recently, the partial convolution neural network (PCNN), a method designed to simultaneously address spatial and temporal distributions, has been demonstrated to handle grid-like data structures effectively. In grid-like structures, the convolution operation processes information within localized windows, making it ideal for imputing missing values in structured data. However, PCNN has been deemed unsuitable for handling data with non-linear relationships to temporal

* Corresponding author. Department of Earth and Environmental Sciences, Korea University, Seoul, 02841, South Korea.
E-mail address: darkuncler@korea.ac.kr (J. Gil).

https://doi.org/10.1016/j.envsoft.2025.106466
Received 14 January 2025; Received in revised form 12 March 2025; Accepted 7 April 2025
Available online 11 April 2025

dependencies (Sayeed et al., 2022). Graph ML (GML), which is based on graph theory (GT) (Adam-Poupart et al., 2014; Barabási, 2016; Hamilton, 2020), has been explored as a potential alternative to PCNN, specifically for datasets with non-linear interconnections. It has proven particularly effective for estimating air pollutant concentrations when monitoring stations are sparsely distributed across a region (Betancourt et al., 2023; Ejurothu et al., 2023; Singh et al., 2024).

Among the various statistical methods used in GML, random forest (RF) methods have found extensive applications in previous research (Betancourt et al., 2023). For instance, previous studies focused on estimating air pollutant data have reported that tree-based ML methods and artificial neural networks outperform other techniques. This superior performance is attributed to the fact that tree-based ML methods integrate predictions from multiple decision trees, which reduces overfitting and effectively handles data with non-linear relationships to temporal dependencies (Sanjeev, 2021; Seraphim et al., 2024; Van et al., 2023; Zhao et al., 2023).

However, one major problem in interpolation studies using ML methods is that although numerous models demonstrate good performance, they fail to explain how they achieve these results. This lack of transparency can lead to errors in the estimation process and ultimately result in poor performance when ML models are applied to conditions different from those of the training dataset (Gil et al., 2023). Therefore, previous studies have sought to reveal the importance of features to

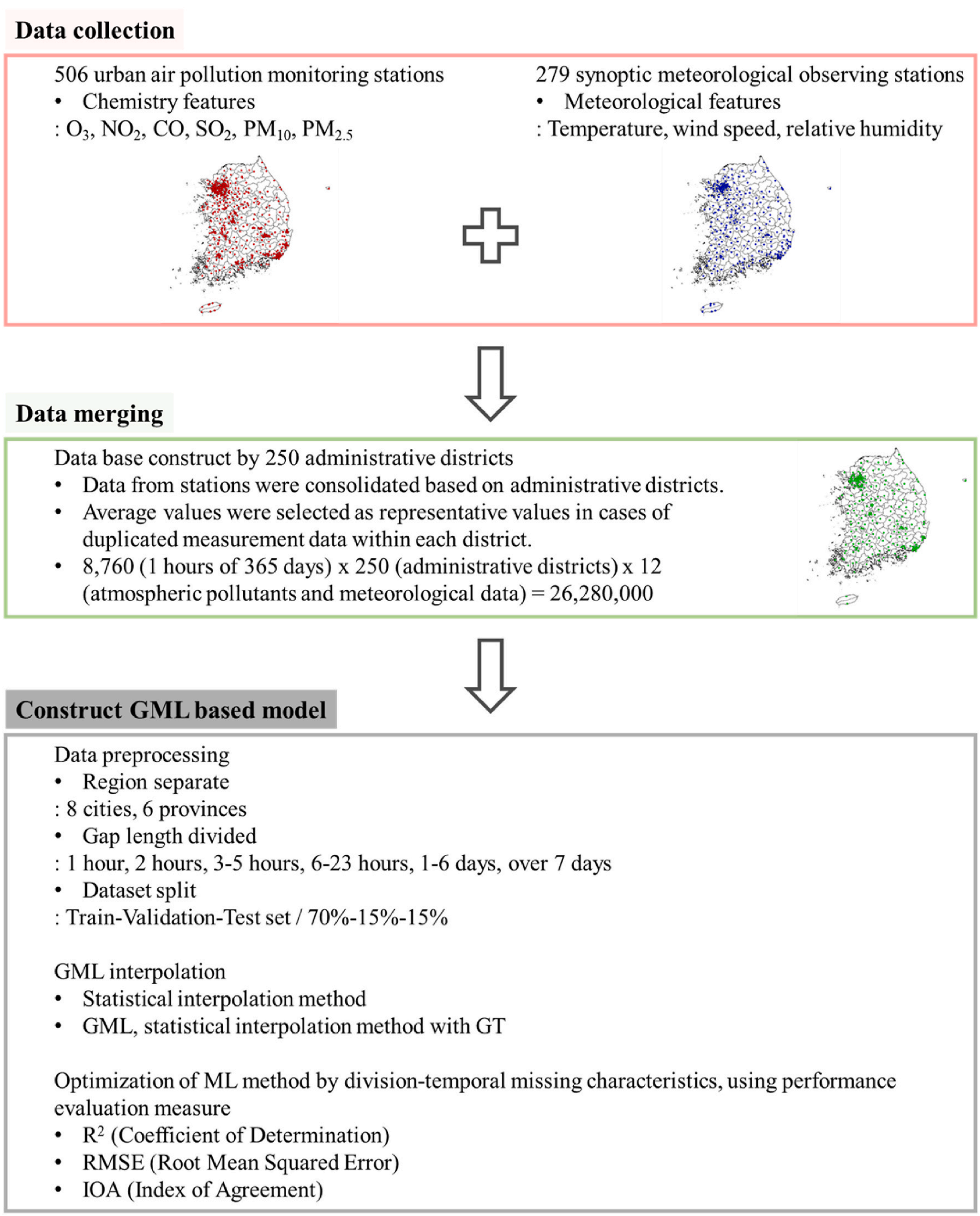

Fig. 1. Framework for constructing the GML model.

output values using eXplainable AI (XAI) methods. For instance, a previous study applied SHapley Additive explanation (SHAP) as an XAI method (Ribeiro et al., 2016; Štrumbelj and Kononenko, 2014). Lundberg and Lee further proposed SHAP analysis using various tree models, including LightGBM, GBoost, CatBoost, and XGBoost (XGB) (Lundberg, 2017). Moreover, several studies have used SHAP analysis and tree-based models for predicting and estimating O_3 concentrations, demonstrating strong performance. Some of these studies have exclusively utilized XGB-based SHAP to identify features influencing O_3 (Kang et al., 2021), while others have combined SHAP with additional XAI methods to analyze key variables in depth (Han et al., 2022; Lundberg and Lee, 2017). In addition to SHAP, various XAI methods, such as the bootstrap test (BT) (Kleinert et al., 2021) and the exponential proportion method (Gil et al., 2021), have been used to examine the influence of features on output values. However, it remains unclear whether these methods produce consistent results, as they are based on different ML models and estimation techniques for feature influence (Adadi and Berrada, 2018; Guidotti et al., 2018; Lundberg et al., 2020).

Therefore, in this study, we developed a GML model to estimate missing O_3 data, considering the spatiotemporal characteristics of regions in South Korea using data from the 2021 national air quality monitoring network and meteorological data. The performance of the developed GML model was evaluated based on regional characteristics. Additionally, we conducted a comparative study of various XAI methods to evaluate their suitability for examining the influence of input parameters on O_3 formation.

2. Materials and methodology

2.1. Observation database construction

In this study, O_3 and associated trace gases (NO_2, CO, and SO_2), particulate matter ($PM_{2.5}$ and PM_{10}), ground-level synoptic meteorological data (temperature, relative humidity, and wind speed), and temporal variables (hour of the day [HOD], day of the week [DOW], and day of year [DOY]) were used as parameters to develop the O_3 estimation GML model (Fig. 1). These data span from January 1 to December 31, 2021. Air quality data were obtained from the observatory of the Ministry of Environment, while synoptic meteorological data were sourced from the Korea Meteorological Administration. Both datasets are publicly available for download at (https://www.airkorea.or.kr/, https://data.kma.go.kr/). The air quality data and meteorological data were originally recorded at 1 h and 1 min intervals, respectively. For database construction, the meteorological data were aggregated to 1 h intervals.

As of 2021, South Korea has 250 administrative districts, with 506 urban air pollution monitoring stations and 279 synoptic meteorological observing stations, some of which overlap in certain districts. Hence, in this study, data obtained from urban air pollution monitoring stations and synoptic meteorological observing stations were consolidated by administrative district. In cases of overlapping measurement data within a district, the average values were used as representative values. Consequently, the finalized database size was 8760 (hours per year) × 250 (administrative districts) × 12 (atmospheric pollutants and meteorological data), totaling 26,280,000 entries. The dataset was divided into two subsets: one for metropolitan areas and the other for provinces. Two GML models were trained for these subsets, designated as “City” and “Province,” respectively (Table S1).

A critical parameter in GML-based data estimation is the temporal length of missing values (Betancourt et al., 2023). Following previous studies, we grouped the temporal intervals of consecutive missing data into six categories: 1 h, 2 h, 3–5 h, 6–23 h, 1–6 days, and 7 days or more. The proportions of missing data were approximately 10 % in the City subset and 8 % in the Province subset. The 1-h interval group had the fewest missing data values, with approximately 3 % in the City subset and 2 % in the Province subset. In contrast, the 7-day-or-more interval group had the highest proportion of missing data, with 61 % in the City subset and 77 % in the Province subset (Table S2). The overall ratio for training, validation, and test data was set at 70:15:15 as randomly extracted from the whole period and consistently applied to each temporal group (Table S3).

2.2. GML for missing data estimation

Most statistical models for time series interpolation do not account for the spatial distribution of data. However, the air pollutant concentration data of nearby stations can exhibit strong correlations, while data from distant stations are generally independent. Therefore, GT, the foundation of GML, assigns a weight to represent the influence of spatial distance. This weight is calculated for connections (edges) between 1-h data points (nodes) that fall within the spatial characteristic shared distance (SCSD) (Table 1). This approach enhances the estimation accuracy of commonly used statistical models, particularly by accounting for spatiotemporal correlations. In this study, we developed a statistical baseline method (BM) for O_3 concentration estimation (Fig. 2). This method uses statistical interpolation techniques, including the spatial mean (SM) method, spatiotemporal mean method (STM), nearest neighbor hybrid (NNH), and RF (eqs. (1)–(4)), while incorporating spatial distribution weights derived from GT. NNH is composed of Linear Interpolation (LI) and Nearest Neighbor Interpolation (NNI), and it is determined by the blank length (L) and threshold length (L_t) (eqs. (5) and (6)). For a detailed explanation of the methodology employed in this study, refer to Text S1 in the supplementary materials and Betancourt's methodology (Betancourt et al., 2023).

$$\text{SM method } \left(\hat{y}_{SM}(t_j)\right) = \frac{1}{N(t_j)} \sum_{x_i, t_j} y(x_i, t_j) \tag{1}$$

$$\text{STM } \left(\hat{y}_{STM}\right) = \frac{1}{N} \sum_{x_i, t_j} y(x_i, t_j) \tag{2}$$

$$\text{NNH } \left(\hat{y}_{NNH}\left(x_i, t_j, \vec{f}_k\right)\right) = \begin{cases} \hat{y}_{LI}(x_i, t_j), \text{if } L < L_t \\ \hat{y}_{NNI}\left(\vec{f}_k\right), \text{if } L \geq L_t \end{cases} \tag{3}$$

$$\text{RF } \left(\hat{y}_{RF}\right) = RF\left(\vec{f}_k\right) \tag{4}$$

$$\hat{y}_{LI}(x_i, t_1) = y(x_i, t_1) + \frac{t_j - t_1}{t_2 - t_1}\left(y(x_i, t_2) - y(x_i, t_1)\right) \ (when, L < L_t) \tag{5}$$

$$\hat{y}_{NNI}\left(\vec{f}_k\right) = y\left(\vec{f}_{k'}\right)|_{d_{k,k'} = min} \text{ with } d_{k,k'} = \left|\left(\vec{f}_k - \vec{f}_{k'}\right)\right| \ (when, L \geq L_t) \tag{6}$$

Table 1
SCSD of each region in this study.

Division	Region	SCSD (km)
City	Busan	4
	Daegu	4
	Daejeon	4
	Gwangju	4
	Incheon	4
	Sejong	4
	Seoul	4
	Ulsan	4
Province	Chungcheong	11
	Gangwon	23
	Gyeonggi	9
	Gyeongsang	17
	Jeju	21
	Jeolla	17

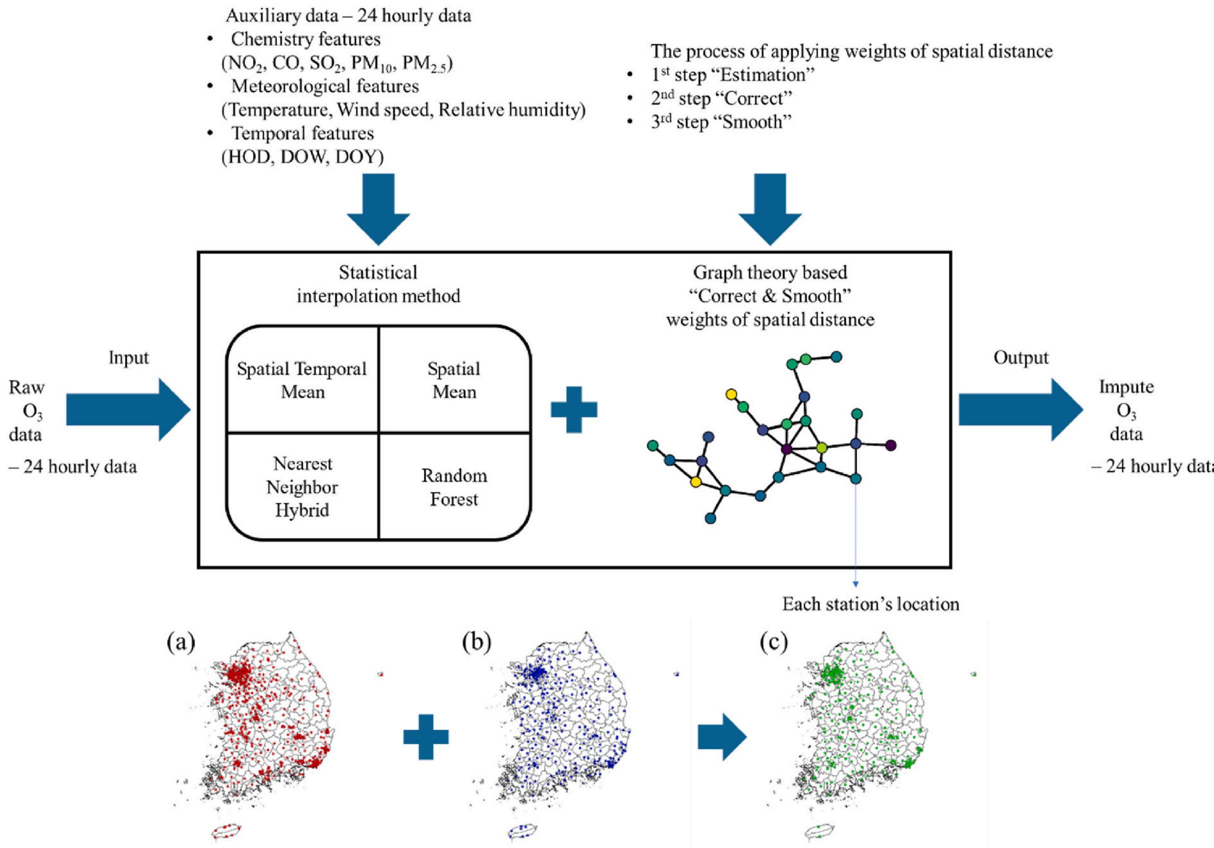

Fig. 2. Construction process of the database used for GML model development. The data were obtained from: (a) Urban air pollution monitoring stations. (b) Synoptic meteorological observing stations within a single administrative district of South Korea. These data were merged based on (c) administrative districts of South Korea in 2021 to represent concentrations.

In the above equations, y represents the actual observed values, $\hat{y}$ denotes predicted values, N represents the total number of observations, x_i denotes a measurement station with index i, and t_j represents a time step with index j. The, $\vec{f}$ represents spatialized input data excluding the data to be predicted for an arbitrary node k, which is every missing ozone value with index. For the NNH and RF methods, missing values were interpolated using the SM of the nearest neighboring observatories at the same time step to maintain data integrity.

After constructing the four BMs, we determined the influences of spatial distance as weights. The SCSD was defined as the 75th percentile of spatial distances (Δx), calculated as the distance between all measurement nodes within a city or province. The SCSD values were approximately 4 km for cities and ranged from 9 to 23 km for provinces. Connections between nodes within this radius were defined as edges, and the weights of these edges were calculated.

Specifically, the weights of edges were determined by the spatial distance (Δx) and temporal distance (Δt) between arbitrary nodes k and k'. Previous studies have reported that the criteria for O_3 transport effects determining Δt are ambiguous in metropolitan regions and provinces in South Korea (Ghim and Chang, 2000; Oh et al., 2010). Consequently, in this study, only spatial distance (Δx) was considered when assigning spatial distribution weights (eq. (7)).

$$w_{k\to k'} = \frac{x - |\Delta x_{I,i'}|}{x} \tag{7}$$

The process of applying these weights to the BM involves three steps, referred to as correct and smooth (CaS), with detailed procedures described in the supplementary materials (Text S1) (Han et al., 2023; Shi et al., 2020).

2.3. Evaluation matrix for GML and hyperparameter tuning based on grid search

The performance of GML was assessed using the coefficient of determination (R^2), root mean squared error (RMSE), and the index of agreement (IOA), as defined in eqs. (8)–(10).

$$R^2 = 1 - \frac{\sum_{k=1}^{N}(y_k - \hat{y}_k)^2}{\sum_{k=1}^{N}(y_k - \langle y\rangle)^2} \ \ with\ \langle y\rangle = \frac{1}{N}\sum_{k=1}^{N} y_k \tag{8}$$

$$RMSE = \sqrt{\sum_{k=1}^{N}\frac{(y_k - \hat{y}_k)^2}{N}} \tag{9}$$

$$IOA = 1 - \frac{\sum_{k=1}^{N}(y_k - \hat{y}_k)^2}{\sum_{k=1}^{N}(|y_k - \langle y\rangle| + |\hat{y}_k - \langle y\rangle|)} \tag{10}$$

Here, N denotes the total number of index nodes, y_k signifies the observed O_3, and $\hat{y}_k$ represents the estimated O_3.

A grid search was conducted to optimize hyperparameters, namely L_t for NNH, depth for RF, and adjustable parameters for CaS. The detailed optimization process and selected hyperparameter values are provided in the supplementary materials (text S2).

Input features for GML were selected using the forward feature selection (FFS) method (Table 2). The FFS method identifies the optimal combination of features by evaluating each feature individually and selecting the combination that produces the best results (Meyer et al., 2018). Our FFS analysis revealed that all of temporal feature (HOD, DOW, and DOY) were consistently selected across all regions, and NO_2 was similarly selected in chemistry feature. In meteorological feature,

Table 2
Selected features for each region based on the FFS method.

Division	Region	Temporal feature	Chemistry feature	Meteorological feature
City	Busan	HoD, DoW, DoY	NO_2	–
	Daegu	HoD, DoW, DoY	NO_2	Temp, RH
	Daejeon	HoD, DoW, DoY	NO_2	Temp
	Gwangju	HoD, DoW, DoY	NO_2, SO_2, $PM_{2.5}$	Temp, WS, RH
	Incheon	HoD, DoW, DoY	NO_2	Temp
	Sejong	HoD, DoW, DoY	NO_2	Temp
	Seoul	HoD, DoW, DoY	NO_2	Temp
	Ulsan	HoD, DoW, DoY	NO_2, SO_2	Temp, WS
Province	Chungcheong	HoD, DoW, DoY	NO_2, SO_2, PM_{10}, $PM_{2.5}$	Temp, RH
	Gangwon	HoD, DoW, DoY	NO_2, PM_{10}, $PM_{2.5}$	Temp, WS
	Gyeonggi	HoD, DoW, DoY	NO_2, PM_{10}, $PM_{2.5}$	Temp
	Gyeongsang	HoD, DoW, DoY	NO_2, $PM_{2.5}$	Temp, RH
	Jeju	HoD, DoW, DoY	NO_2, CO, SO_2	Temp
	Jeolla	HoD, DoW, DoY	NO_2, $PM_{2.5}$	Temp, WS, RH

the temperature was selected most regions, only excepted in Busan. As a result, these temporal features, NO_2, and other features were selected specific to each region.

2.4. Intercomparison of feature importance using different XAI methods: BT and SHAP

We performed FFS to select the optimal combination of input features to improve model performance. These feature combinations represent the spatiotemporal characteristics for O_3 estimation. However, FFS only provides information regarding the influence of feature combinations on the output value but does not explain the details of each feature's impact. Therefore, in this study, we applied two different combinations of ML models and XAI methods to estimate individual feature importance: (1) RF, an ML model and one of the BMs used in this study, paired with the BT as the XAI method, and (2) XGB, an ML model, paired with SHAP as the XAI method. The BT method estimates feature importance by setting feature values to zero and monitoring changes in model performance (Gil et al., 2023; Kleinert et al., 2021). The SHAP method calculates the contribution of each feature to the output based on the Shapley value (Jabeur et al., 2024; Lundberg and Lee, 2017). The data used to construct the XGB model for SHAP analysis are identical to the data used to construct GML. Details of the XGB results in our dataset are provided in the supplementary materials (Text S3).

3. Results and discussion

3.1. Optimization of ML methods using division-temporal missing characteristics

The performance of the four BMs and their corresponding CaS models was evaluated using test datasets. Seoul was chosen to represent cities (Tables 3 and 4), while Gyeonggi was selected to represent provinces (Tables 5 and 6). Seoul and Gyeonggi have 25 and 109 observatories, respectively. Performance metrics for each station were averaged regionally for comparison.

The application of CaS, regardless of spatiotemporal conditions or BMs, resulted in improved model accuracy (Fig. 3). In Seoul, without CaS, the R^2 value ranged from 0.78 to 0.97, RMSE ranged from 3.37 ppbv to 8.90 ppbv, and IOA ranged from 0.94 to 0.99. However, after applying CaS, the performance improved, with the minimum R^2 increasing from 0.89 to 0.96 and the minimum IOA increasing from 0.97 to 0.99. Additionally, the maximum RMSE decreased significantly, from 8.90 ppbv to 5.88 ppbv. In Gyeonggi, the R^2 value ranged from 0.73 to 0.97, RMSE ranged from 3.54 ppbv to 10.95 ppbv, and IOA ranged from 0.93 to 0.99. The BM's performance in Gyeonggi was slightly lower or comparable to that in Seoul. Similarly, after applying CaS in Gyeonggi, the R^2 value increased from 0.84 to 0.94, IOA increased from 0.96 to 0.98, and RMSE decreased from 8.38 ppbv to 4.99 ppbv. For missing intervals of less than 1 h, the BM without CaS performed better. This suggests that estimating O_3 variation based on its own variability is sufficient for single-point predictions. The application of CaS

Table 3
Evaluation matrix for Seoul, representing city simulations for temporal intervals of missing length of less than 6 h.

Division	Region	Gap length	Model		R^2	RMSE (ppb)	IOA
City	Seoul	1 h	spatiotemporal mean		0.00	18.39	0.06
				+ correct and smooth	0.77	8.81	0.94
			spatial mên		0.90	5.93	0.97
				+ correct and smooth	0.91	5.41	0.98
			nearest neighbor hybrid		**0.97**	**3.37**	**0.99**
				+ correct and smooth	0.95	4.15	0.99
			random forest		0.91	5.42	0.98
				+ correct and smooth	0.93	4.89	0.98
		2 h	spatiotemporal mean		0.00	19.47	0.05
				+ correct and smooth	0.77	9.36	0.93
			spatial mean		0.91	5.88	0.98
				+ correct and smooth	0.92	5.33	0.98
			nearest neighbor hybrid		0.95	4.53	0.99
				+ correct and smooth	**0.95**	**4.45**	**0.99**
			random forest		0.92	5.56	0.98
				+ correct and smooth	0.94	4.86	0.98
		3–5 h	spatiotemporal mean		0.00	19.01	0.01
				+ correct and smooth	0.79	8.78	0.94
			spatial mean		0.90	5.88	0.97
				+ correct and smooth	0.92	5.31	0.98
			nearest neighbor hybrid		0.89	6.23	0.97
				+ correct and smooth	0.93	5.00	0.98
			random forest		0.91	5.57	0.98
				+ correct and smooth	**0.93**	**4.90**	**0.98**

Table 4
Evaluation matrix for Seoul, representing city simulations for temporal intervals of missing length of 6 h or more.

Division	Region	Gap length	Model		R^2	RMSE (*ppb*)	IOA
City	Seoul	6–23 h	spatiotemporal mean		−0.01	20.07	0.13
				+ correct and smooth	0.75	9.93	0.93
			spatial mean		0.91	5.89	0.98
				+ correct and smooth	0.92	5.54	0.98
			nearest neighbor hybrid		0.80	8.90	0.95
				+ correct and smooth	0.91	5.88	0.98
			random forest		0.92	5.51	0.98
				+ correct and smooth	**0.94**	**4.78**	**0.98**
		1–6 d	spatiotemporal mean		0.00	17.65	0.05
				+ correct and smooth	0.75	8.74	0.93
			spatial mean		0.89	5.87	0.97
				+ correct and smooth	0.91	5.36	0.98
			nearest neighbor hybrid		0.78	8.29	0.94
				+ correct and smooth	0.89	5.72	0.97
			random forest		0.89	5.74	0.97
				+ correct and smooth	**0.92**	**4.93**	**0.98**
		≥7d	spatiotemporal mean		−0.01	17.86	0.12
				+ correct and smooth	0.88	6.16	0.97
			spatial mean		0.90	5.75	0.97
				+ correct and smooth	0.93	4.84	0.98
			nearest neighbor hybrid		0.81	7.82	0.95
				+ correct and smooth	0.93	4.64	0.98
			random forest		0.93	4.75	0.98
				+ correct and smooth	**0.96**	**3.60**	**0.99**

Table 5
Evaluation matrix for Gyeonggi, representing province simulations for temporal intervals of missing length of less than 6 h.

Division	Region	Gap length	Model		R^2	RMSE (*ppb*)	IOA
Province	Gyeonggi	1 h	spatiotemporal mean		0.00	20.21	0.06
				+ correct and smooth	0.74	10.27	0.92
			spatial mean		0.84	8.18	0.95
				+ correct and smooth	0.89	6.82	0.97
			nearest neighbor hybrid		**0.97**	**3.54**	**0.99**
				+ correct and smooth	0.94	4.99	0.98
			random forest		0.89	6.69	0.97
				+ correct and smooth	0.92	5.72	0.98
		2 h	spatiotemporal mean		0.00	20.48	0.04
				+ correct and smooth	0.74	10.36	0.92
			spatial mean		0.84	8.18	0.96
				+ correct and smooth	0.88	7.05	0.97
			nearest neighbor hybrid		**0.95**	**4.74**	**0.99**
				+ correct and smooth	0.93	5.38	0.98
			random forest		0.89	6.79	0.97
				+ correct and smooth	0.92	5.87	0.98
		3–5 h	spatiotemporal mean		0.00	20.26	0.03
				+ correct and smooth	0.75	10.11	0.92
			spatial mean		0.84	8.00	0.96
				+ correct and smooth	0.88	6.96	0.97
			nearest neighbor hybrid		0.90	6.47	0.97
				+ correct and smooth	0.92	5.90	0.98
			random forest		0.88	6.98	0.97
				+ correct and smooth	**0.92**	**5.90**	**0.98**

demonstrated overall improvements in performance metrics across all groups regardless of the region, with the exception of 1 h missing intervals.

Previous studies employing GML for missing O_3 data estimation in other countries have reported R^2 values ranging from 0.00 to 0.97, RMSE values ranging from 2.43 ppbv to 16.25 ppbv, and IOA values ranging from 0.00 to 0.99 in the absence of CaS (Betancourt et al., 2023). After implementing CaS, R^2 values increased to a range of 0.52–0.97, RMSE values decreased to a range of 3.07 ppbv–10.62 ppbv, and IOA values increased to a range of 0.82–0.99. These findings suggest that applying CaS can significantly enhance overall performance metrics.

In previous studies focusing on South Korea, several ML models for estimating missing temporal O_3 data from the records of multiple ground observatories achieved R^2 values ranging from 0.74 to 0.93, IOA values ranging from 0.84 to 0.93, and RMSE values ranging from 5.50 ppbv to 10.70 ppbv (Eslami et al., 2020; Sayeed et al., 2022; Tang et al., 2024). Additionally, combining ML and Kriging methods to integrate spatiotemporal characteristics resulted in IOA values ranging from 0.20 to 0.61 (Eslami et al., 2020). These results indicate that using GML may achieve superior performance compared to other estimation methods by effectively accounting for complex spatial and temporal characteristics.

Based on the performance evaluation results, we selected the methods with the best performance metrics for each missing data time interval. Specifically, for intervals of 2 h or less, we selected NNH, which achieved R^2 values ranging from 0.95 to 0.97, RMSE values ranging from 3.37 ppbv to 4.74 ppbv, and an IOA value of 0.99. Meanwhile, for intervals exceeding 2 h, we selected RF, which achieved R^2 values ranging from 0.90 to 0.96, RMSE values ranging from 3.60 ppbv to 6.80 ppbv, and IOA values ranging from 0.97 to 0.99. After model selection, CaS was applied to estimate and interpolate O_3 concentrations for 250

Table 6
Evaluation matrix for Gyeonggi, representing province simulations for temporal intervals of missing length of 6 h or more.

Division	Region	Gap length	Model		R^2	RMSE (ppb)	IOA
Province	Gyeonggi	6–23 h	spatiotemporal mean		0.00	20.81	0.04
				+ correct and smooth	0.72	10.97	0.91
			spatial mean		0.84	8.34	0.96
				+ correct and smooth	0.88	7.13	0.97
			nearest neighbor hybrid		0.77	9.91	0.94
				+ correct and smooth	0.87	7.50	0.96
			random forest		0.89	6.86	0.97
				+ correct and smooth	**0.92**	**6.01**	**0.98**
		1–6 d	spatiotemporal mean		−0.01	20.27	0.12
				+ correct and smooth	0.71	10.91	0.91
			spatial mean		0.83	8.28	0.95
				+ correct and smooth	0.88	6.92	0.97
			nearest neighbor hybrid		0.78	9.55	0.94
				+ correct and smooth	0.87	7.25	0.96
			random forest		0.88	7.00	0.97
				+ correct and smooth	**0.92**	**5.88**	**0.98**
		≥7 d	spatiotemporal mean		−0.02	21.27	0.16
				+ correct and smooth	0.72	11.08	0.91
			spatial mean		0.81	9.12	0.95
				+ correct and smooth	0.86	7.93	0.96
			nearest neighbor hybrid		0.73	10.95	0.93
				+ correct and smooth	0.84	8.38	0.96
			random forest		0.86	7.97	0.96
				+ correct and smooth	**0.90**	**6.80**	**0.97**

administrative districts across the country (Fig. S2). The interpolated dataset resulted in an increase in average O_3 levels compared to the raw dataset due to differences in data distribution (Fig. S3). Specifically, the interpolated dataset led to annual O_3 increases of over 1.44 % in cities and more than 3.26 % in provinces (Fig. S4). Moreover, the discrepancy between the original and interpolated data was more pronounced in provinces than in cities. This discrepancy is likely due to differences in the number of observatories per area within the SCSD between cities and provinces. These results highlight the potential for bias in O_3 estimations when unobserved data are used and emphasize the importance of accurate missing data interpolation (Yeo and Kim, 2021).

3.2. Model-uncertainty-inducing parameters

Uncertainty in GML models for O_3 estimation is influenced by several factors. O_3 levels in Korea are subject to seasonal variations due to its mid-latitude position on the Korean Peninsula. The distribution of ozone concentrations exhibits temporal and spatial variability, influenced by the high population densities in specific cities and distinct industrial structures across cities and regions (Colombi et al., 2023; Jung et al., 2018; Lee and Park, 2022). Across the test dataset for each region, the monthly deviation, representing the difference between observed and estimated O_3 concentrations, increased significantly by up to 6.08 ppbv in June. This deviation also demonstrated the widest range, spanning from −7.61 to 18.01 ppbv (Fig. S5). This coincides with the season of highest O_3 concentration, highlighting the difficulty in accurately estimating extremely high O_3 levels due to the inherent limitations of ML models in capturing average variations. This underscores the need for specialized models to simulate the high O_3 concentrations observed during summer, which coincide with rising temperatures.

Furthermore, as the proportion of missing data increases, the discrepancy between average O_3 concentrations in the raw and interpolated datasets increases. In our study, the increase in mean O_3 concentration deviation relative to the missing data proportion was lower for cities (0.049) compared to that for provinces (0.087) (Fig. S6). This is likely due to the smaller average distance between observatories in cities (4.7 km) compared to that in provinces (10.9 km), as well as the higher average number of measurement stations per edge used to calculate spatial weights in cities (5.29) compared to that in provinces (1.38). These findings suggest that developing a customized model for each region is more appropriate for reflecting temporal characteristics of pollutant variations and hyperparameter settings. This approach may be more effective than constructing a universally optimized model for all spatiotemporal points.

3.3. Parameter analysis using BT and SHAP

As detailed previously, ML models are influenced by various underlying spatiotemporal characteristics. This makes it challenging to understand how the models operate and to ensure the reliability of their results. XAI is a method that provides insights into the processes by which models calculate output values from input parameters. However, different XAI methods may provide differing insights owing to their inherent characteristics, which can lead to confusion in result interpretation. Therefore, in this study, we used two different ML–XAI combinations (see Section 2.4 for details) and compared their results (Fig. 4). In the City subset, the R^2 value was 0.88 and the RMSE value was 6.30 ppbv when RF was applied without the CaS process. However, after incorporating spatial O_3 characteristics using the CaS process, the R^2 value improved to 0.90 and the RMSE decreased to 5.83 ppbv, reflecting improved performance metrics similar to those in Fig. 3. A similar pattern was observed in the Province subset. Without the CaS process, RF achieved an R^2 value of 0.82 and an RMSE value of 7.45 ppbv. However, after incorporating spatial O_3 characteristics using CaS, the R^2 value improved to 0.85 and the RMSE decreased to 6.73 ppbv. The results of XGBoost were similar to those of RF, with an R^2 value of 0.85 and an RMSE value of 7.14 ppbv for the City subset and an R^2 value of 0.81 and an RMSE value of 8.18 ppbv for the Province subset.

The BT results reveal that NO_2 and DOY were the most important features in both City and Province subsets (Fig. 5 and Table S4). These features indicate the significance of O_3 precursors (NO_2) and seasonal changes (DOY). NO_2 is the most important feature in the City subset but is less important than DOY in the Province subset. This suggests that direct NO_2 emissions are more significant for estimating O_3 levels in City regions, where observatories are located closer to emission sources (e.g., vehicles), compared to Province regions. In Province regions, the larger area weakens the influence of direct NO_2 emissions. Consequently, O_3 estimation is more strongly affected by seasonal changes in monthly averaged NO_2 levels and variations in sunlight, which drive photochemical reactivity changes. Furthermore, features with lower importance, such as HOD and temperature, share similar characteristics with DOY. These features reinforce photochemical reactivity during daytime

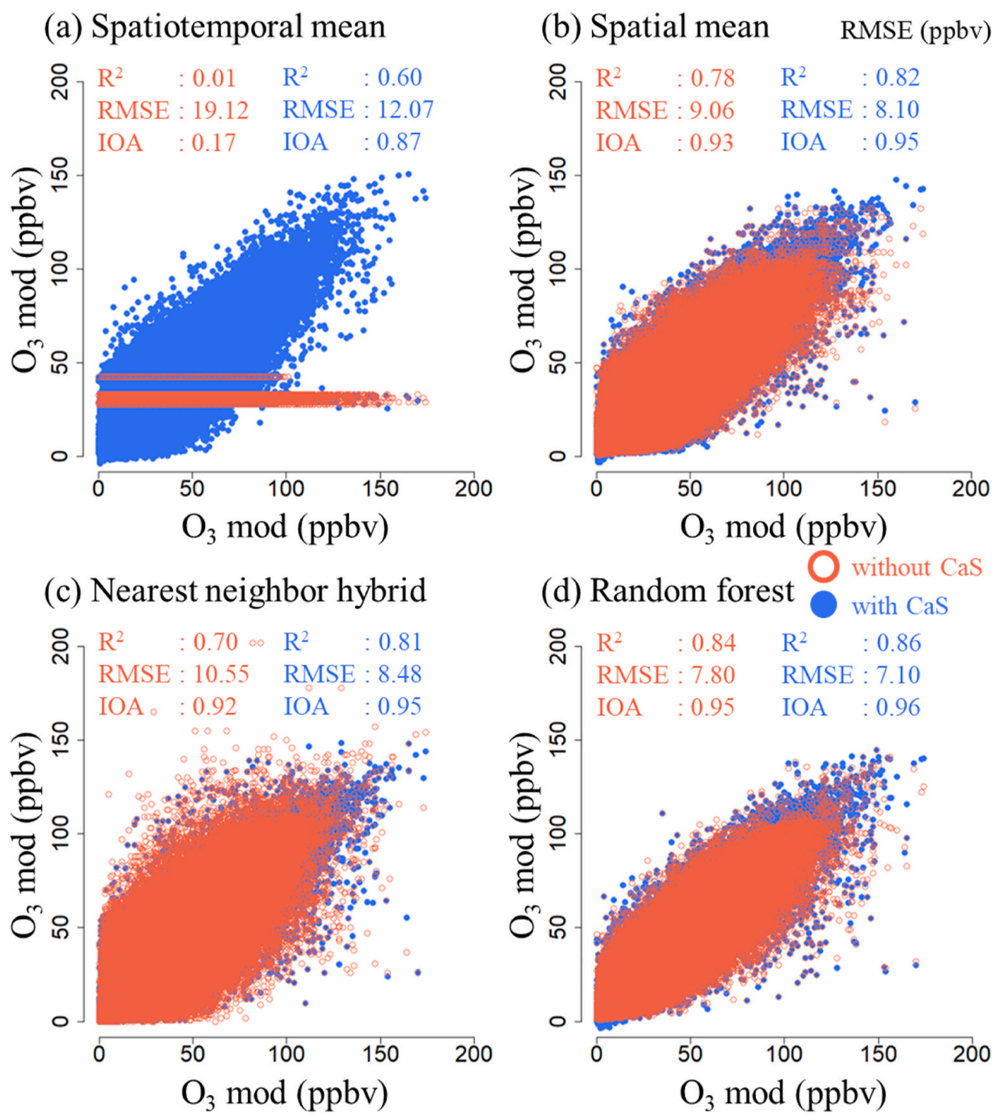

Fig. 3. Correlation between observed data and simulation data from four BMs, with CaS (blue closed circles) and without CaS (red open circles). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

(HOD) and summer periods (DOY) as temperatures increase. These results align with current knowledge, indicating that the model is well-constructed and accurately reflects natural atmospheric conditions. The BT results effectively explain the influence of each feature on O_3 output.

When CaS was applied to RF (closed boxes in Fig. 5), the deterioration in model performance was smaller compared to RF without CaS (open boxes in Fig. 5). This suggests that the CaS method reduces uncertainty from input variables, thereby improving model performance. Despite these changes, the major features (NO_2, DOY, HOD, and temperature) retained their rankings, indicating their consistent importance in estimating O_3 regardless of model performance. Additionally, when CaS was applied, the importance of DOY was reduced relative to that of NO_2 as CaS modulated the effects of spatial distribution. This suggests that, in XAI methods, CaS can serve as an optimal tool for interpreting complex spatiotemporal data.

Similar to the BT results, emission (NO_2) and reaction (HOD, DOY, and T) features in SHAP were also identified as the most influential factors in both the City and Province regions. In particular, the influence of NO_2 in City regions remains the most significant feature across all ML models (RF, XGB) and XAI methods (BT, SHAP). This indicates that changes in NO_2, already known as one of the major O_3 precursors (NO_x and VOCs), are the most important factor in estimating City O_3 levels. Additionally, the ML models effectively capture these atmospheric chemical conditions.

An interesting observation in Province regions is that the order of influence of HOD, DOY, and T differs between the BT and SHAP results. In BT, HOD and DOY are more influential than T, whereas in SHAP, T is the most important feature, followed by HOD and DOY. This difference arises from the shared characteristics of HOD, DOY, and T, which are linked to changes in photochemical reactivity influenced by sunlight levels, reaction speed, seasonality, and diurnality. This suggests that interpreting XAI results requires a background understanding of the relationships between input and output features, especially when features are interrelated.

Consequently, our results indicate that the choice of ML models and XAI methods is not critical for explaining the results. This suggests that when researchers apply multiple AI models with XAI methods and obtain similar results, the identified influential features are likely important not only for simulations but also in the actual atmosphere. Furthermore, these findings may provide clues for uncovering unknown mechanisms.

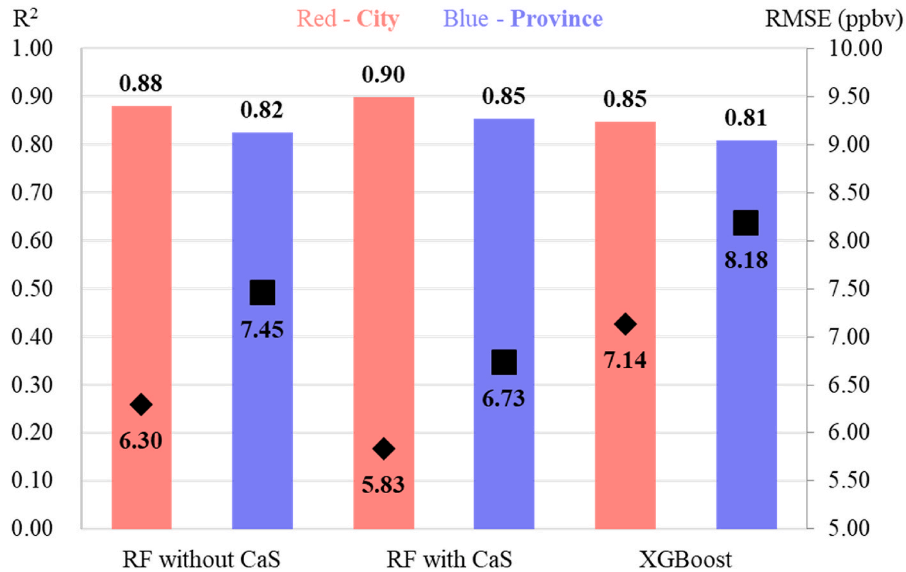

Fig. 4. Comparison of R^2 and RMSE results for RF and XGBoost models in City and Province regions: R^2 results: RF without CaS, RF with CaS, and XGBoost for City (red bars) and Province (blue bars). RMSE results: RF without CaS, RF with CaS, and XGBoost for City (black diamonds) and Province (black squares). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

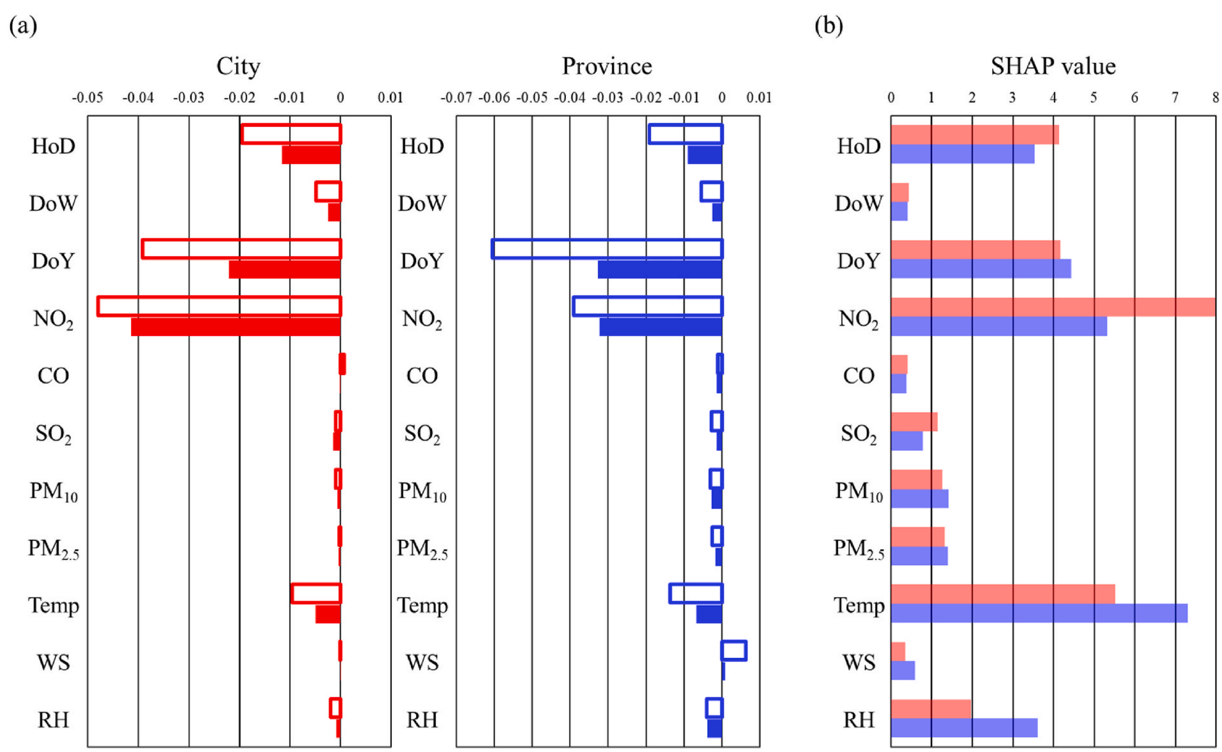

Fig. 5. Importance of input features estimated by (a) BT and (b) SHAP method for City (red) and Province (blue). BT results show the mean decrease in the determination coefficient (R^2). SHAP results indicate Shapley values. Open boxes represent results without CaS, and closed boxes represent results with CaS in panel (a). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3.4. Implications for human health and climate change

In this study, we developed a graph-based ML model by integrating GT into a statistical interpolation method to estimate missing O_3 concentrations. The model utilized atmospheric pollution data and surface meteorological observations recorded by South Korea's urban air monitoring network in 2021. The O_3 estimation results revealed that, except for short-term missing data lasting less than 1 h, the graph-based ML model consistently outperformed conventional statistical interpolation methods and other ML techniques across all regions.

The discrepancy between the raw data and the interpolation results obtained using GML models varies depending on spatiotemporal factors. These factors include high O_3 concentrations during seasonal transitions and variations in the density of regional observatories. These findings align with those of previous studies on $PM_{2.5}$ prediction using GML (Ejurothu et al., 2023), which emphasize the need to account for regional and temporal variations in O_3 when developing ML-based air pollutant prediction models. It is therefore imperative to conduct further research into not only the regional chemistry of O_3 but also national trends in O_3 increases, including transboundary transport and its impacts.

In particular, as summer heatwaves increase due to global climate change, the frequency of high O_3 concentration cases is expected to rise in the future. Our parameter analysis results, based on the XAI method, identified the potential for O_3 rise caused by temperature changes. This O_3 increase presents a critical issue in terms of its impact on vegetation and human health. A previous study demonstrated that elevated O_3 levels are associated with an increased risk of cardiovascular and respiratory diseases. This included a 0.3 % increase in the risk of death and a 6.3 % increase in the risk of death from out-of-hospital coronary events for every 5 ppbv increase in the daily maximum O_3 concentration (Nuvolone et al., 2018). Our results demonstrate that the interpolated O_3 dataset is more sensitive to temperature changes, especially during the hot season (Table S5). This implies that the health effects of O_3 under global warming may have been underestimated in previous studies due to the absence of O_3 data. These findings indicate that missing O_3 data may introduce bias in the estimation of annual average O_3 concentrations, which could have significant implications for research on O_3's health effects. Consequently, it is essential to interpolate missing data using appropriate methodologies for O_3 influence studies. The GML model will be particularly helpful in addressing gaps in understanding O_3 enhancement, and comparison of different ML-XAI combination results can be applied as useful method for validating the developed model.

CRediT authorship contribution statement

Seonje Jung: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Data curation, Conceptualization. **Junsu Gil:** Writing – review & editing, Methodology, Investigation, Conceptualization. **Meehye Lee:** Writing – review & editing, Resources, Investigation. **Clara Betancourt:** Writing – review & editing, Methodology, Investigation. **Martin Schultz:** Writing – review & editing, Investigation. **Yunsoo Choi:** Writing – review & editing, Investigation. **Taekyu Joo:** Writing – review & editing, Investigation. **Daigon Kim:** Data curation.

Software and data availability section 1

- Name of data: TOAR Data Portal
- Contact: Jülich Supercomputing Centre, Forschungszentrum Jülich GmbH, Wilhelm-Johnen-Straße, 52428, Jülich, Germany, info@toar-data.org
- Database at: https://toar-data.org/

Software and data availability section 2

- Name of software: Ozone-interpolation-in-South-Korea
- Developers: Seonje Jung
- Contact: sjjung99@korea.ac.kr
- Date first available: January 13, 2025
- Program language: Python
- Documentation: Detailed documentation for application installation, testing, and deployment can be found at https://github.com/Merkyarhhh/Ozone-interpolation-in-South-Korea/blob/main/README.md
- Source code at: https://github.com/Merkyarhhh/Ozone-interpolation-in-South-Korea

Software and data availability section 3

- Name of software: SHAP
- Developers: Scott Lundberg
- Date first available: May 22, 2018
- Program language: Python
- Source code at: https://shap.readthedocs.io/en/latest/

Software and data availability section 4

The Ozone-interpolation-in-South-Korea for the ozone interpolation was implemented in python (version 3.8) using PyTorch library. Authors' experimental environment was as follows.

- OS: Window 10 (version – 22H2, OS build – 19045.4046)
- CPU: Intel(R) Core(TM) i9-9900K CPU @ 3.60 GHz (12 threads)
- RAM: 64 GB
- GPU: NVIDIA GeForce RTX 2080ti

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00573406).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envsoft.2025.106466.

Data availability

Data will be made available on request.

References

Adadi, A., Berrada, M., 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052.

Adam-Poupart, A., Brand, A., Fournier, M., Jerrett, M., Smargiassi, A., 2014. Spatiotemporal modeling of ozone levels in Quebec (Canada): a comparison of kriging, land-use regression (LUR), and combined Bayesian maximum entropy–LUR approaches. Environ. Health Perspect. 122, 970–976. https://doi.org/10.1289/ehp.1306566.

Agathokleous, E., Kitao, M., Kinose, Y., 2018. A review study on ozone phytotoxicity metrics for setting critical levels in Asia. Asian J. Atmos. Environ. 12, 1–11. https://doi.org/10.5572/ajae.2018.12.1.001.

Barabási, A.-L., 2016. Network Science. Cambridge Univ. Press.

Betancourt, C., Li, C.W.Y., Kleinert, F., Schultz, M.G., 2023. Graph machine learning for improved imputation of missing tropospheric ozone data. Environ. Sci. Technol. 57, 18246–18258. https://doi.org/10.1021/acs.est.3c05104.

Colombi, N.K., Jacob, D.J., Yang, L.H., Zhai, S., Shah, V., Grange, S.K., Yantosca, R.M., Kim, S., Liao, H., 2023. Why is ozone in South Korea and the Seoul metropolitan area so high and increasing? Atmos. Chem. Phys. 23, 4031–4044. https://doi.org/10.5194/acp-23-4031-2023.

Davison, A., Hemphill, M., 1987. On the statistical analysis of ambient ozone data when measurements are missing. Atmos. Environ. 21, 629–638. https://doi.org/10.1016/0004-6981(87)90045-X.

Ejurothu, P.S.S., Mandal, S., Thakur, M., 2023. Forecasting $PM_{2.5}$ concentration in India using a cluster-based hybrid graph neural network approach. Asia-Pac. J. Atmos. Sci. 59, 545–561. https://doi.org/10.1007/s13143-022-00291-4.

Eslami, E., Choi, Y., Lops, Y., Sayeed, A., 2020. A real-time hourly ozone prediction system using deep convolutional neural network. Neural Comput. Appl. 32, 8783–8797. https://doi.org/10.48550/arXiv.1901.11079.

Ghim, Y.S., Chang, Y.-S., 2000. Characteristics of ground-level ozone distributions in Korea for the period of 1990–1995. J. Geophys. Res. Atmos. 105, 8877–8890. https://doi.org/10.1029/1999JD901179.

Gil, J., Kim, J., Lee, M., Lee, G., Ahn, J., Lee, D.S., Jung, J., Cho, S., Whitehill, A., Szykman, J., Lee, J., 2021. Characteristics of HONO and its impact on O_3 formation in the Seoul metropolitan area during the korea-US air quality study. Atmos. Environ. 246, 118182. https://doi.org/10.1016/j.atmosenv.2020.118182.

Gil, J., Lee, M., Kim, J., Lee, G., Ahn, J., Kim, C.-H., 2023. Simulation model of reactive nitrogen species in an urban atmosphere using a deep neural network: RNDv1.0. Geosci. Model Dev. (GMD) 16, 5251–5263. https://doi.org/10.5194/gmd-16-5251-2023.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D., 2018. A survey of methods for explaining black box models. ACM Comput. Surv. 51, 1–42. https://doi.org/10.1145/3236009.

Hamilton, W.L., 2020. Graph Representation Learning, Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer, Cham. https://doi.org/10.1007/978-3-031-01588-5.

Han, J., Liu, H., Xiong, H., Yang, J., 2023. Semi-supervised air quality forecasting via self-supervised hierarchical graph neural network. IEEE Trans. Knowl. Data Eng. 35, 5230–5241. https://doi.org/10.1109/TKDE.2022.3149815.

Han, L., Zhao, J., Gao, Y., Gu, Z., 2022. Prediction and evaluation of spatial distributions of ozone and urban heat island using a machine learning modified land use regression method. Sustain. Cities Soc. 78, 103643. https://doi.org/10.1016/j.scs.2021.103643.

Jabeur, S.B., Mefteh-Wali, S., Viviani, J.-L., 2024. Forecasting gold price with the XGBoost algorithm and SHAP interaction values. Ann. Oper. Res. 334, 679–699. https://doi.org/10.1007/s10479-021-04187-w.

Jung, H.-C., Moon, B.-K., Wie, J., 2018. Seasonal changes in surface ozone over South Korea. Heliyon 4, e00515. https://doi.org/10.1016/j.heliyon.2018.e00515.

Kang, Y., Choi, H., Im, J., Park, S., Shin, M., Song, C.-K., Kim, S., 2021. Estimation of surface-level NO_2 and O_3 concentrations using TROPOMI data and machine learning over East Asia. Environ. Pollut. 288, 117711. https://doi.org/10.1016/j.envpol.2021.117711.

Kleinert, F., Leufen, L.H., Schultz, M.G., 2021. IntelliO3-ts v1.0: a neural network approach to predict near-surface ozone concentrations in Germany. Geosci. Model Dev. (GMD) 14, 1–25. https://doi.org/10.5194/gmd-14-1-2021.

Lee, H.-M., Park, R.J., 2022. Factors determining the seasonal variation of ozone air quality in South Korea: regional background versus domestic emission contributions. Environ. Pollut. 308, 119645. https://doi.org/10.1016/j.envpol.2022.119645.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2, 56–67. https://doi.org/10.1038/s42256-019-0138-9.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., Nauss, T., 2018. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environ. Model. Software 101, 1–9. https://doi.org/10.1016/j.envsoft.2017.12.001.

Mills, G., Pleijel, H., Malley, C.S., Sinha, B., Cooper, O.R., Schultz, M.G., Neufeld, H.S., Simpson, D., Sharps, K., Feng, Z., 2018. Tropospheric Ozone Assessment Report: present-day tropospheric ozone distribution and trends relevant to vegetation. Elem. Sci. Anth. 6, 47. https://doi.org/10.1525/elementa.302.

Nuvolone, D., Petri, D., Voller, F., 2018. The effects of ozone on human health. Environ. Sci. Pollut. Res. 25, 8074–8088. https://doi.org/10.1007/s11356-017-9239-3.

Oh, I.-B., Kim, Y.-K., Hwang, M.-K., Kim, C.-H., Kim, S., Song, S.-K., 2010. Elevated ozone layers over the Seoul Metropolitan Region in Korea: evidence for long-range ozone transport from eastern China and its contribution to surface concentrations. J. Appl. Meteorol. Climatol. 49, 203–220. https://doi.org/10.1175/2009JAMC2213.1.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778.

Sanjeev, D., 2021. Implementation of machine learning algorithms for analysis and prediction of air quality. Int. J. Eng. Res. Technol. 10, 533–538.

Sayeed, A., Choi, Y., Pouyaei, A., Lops, Y., Jung, J., Salman, A.K., 2022. CNN-based model for the spatial imputation (CMSI version 1.0) of in-situ ozone and $PM_{2.5}$ measurements. Atmos. Environ. 289, 119348. https://doi.org/10.1016/j.atmosenv.2022.119348.

Schultz, M.G., Schröder, S., Lyapina, O., Cooper, O.R., Galbally, I., Petropavlovskikh, I., von Schneidemesser, E., Tanimoto, H., Elshorbany, Y., Naja, M., Seguel, R.J., Dauert, U., Eckhardt, P., Feigenspan, S., Fiebig, M., Hjellbrekke, A.-G., Hong, Y.-D., Kjeld, P.C., Koide, H., Lear, G., Tarasick, D., Ueno, M., Wallasch, M., Baumgardner, D., Chuang, M.-T., Gillett, R., Lee, M., Molloy, S., Moolla, R., Wang, T., Sharps, K., Adame, J.A., Ancellet, G., Apadula, F., Artaxo, P., Barlasina, M. E., Bogucka, M., Bonasoni, P., Chang, L., Colomb, A., Cuevas-Agulló, E., Cupeiro, M., Degorska, A., Ding, A., Fröhlich, M., Frolova, M., Gadhavi, H., Gheusi, F., Gilge, S., Gonzalez, M.Y., Gros, V., Hamad, S.H., Helmig, D., Henriques, D., Hermansen, O., Holla, R., Hueber, J., Im, U., Jaffe, D.A., Komala, N., Kubistin, D., Lam, K.-S., Laurila, T., Lee, H., Levy, I., Mazzoleni, C., Mazzoleni, L.R., McClure-Begley, A., Mohamad, M., Murovec, M., Navarro-Comas, M., Nicodim, F., Parrish, D., Read, K. A., Reid, N., Ries, L., Saxena, P., Schwab, J.J., Scorgie, Y., Senik, I., Simmonds, P., Sinha, V., Skorokhod, A.I., Spain, G., Spangl, W., Spoor, R., Springston, S.R., Steer, K., Steinbacher, M., Suharguniyawan, E., Torre, P., Trickl, T., Weili, L., Weller, R., Xiaobin, X., Xue, L., Zhiqiang, M., 2017. Tropospheric Ozone Assessment Report: database and metrics data of global surface ozone observations. Elem. Sci. Anth. 5, 58. https://doi.org/10.1525/elementa.244.

Seraphim, B.I., Nagarajan, E., Pathak, S., 2024. Predicting air quality: ensemble learning approach. In: 2024 Second International Conference on Data Science and Information System (ICDSIS). IEEE. https://doi.org/10.1109/ICDSIS61070.2024.10594204.

Shi, Y., Huang, Z., Feng, S., Zhong, H., Wang, W., Sun, Y., 2020. Masked label prediction: unified message passing model for semi-supervised classification. https://doi.org/10.48550/arXiv.2009.03509.

Singh, D., Choi, Y., Park, J., Salman, A.K., Sayeed, A., Song, C.H., 2024. Deep-BCSI: a deep learning-based framework for bias correction and spatial imputation of $PM_{2.5}$ concentrations in South Korea. Atmos. Res. 301, 107283. https://doi.org/10.1016/j.atmosres.2024.107283.

Štrumbelj, E., Kononenko, I., 2014. Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. 41, 647–665. https://doi.org/10.1007/s10115-013-0679-x.

Tang, B., Stanier, C.O., Carmichael, G.R., Gao, M., 2024. Ozone, nitrogen dioxide, and $PM_{2.5}$ estimation from observation-model machine learning fusion over S. Korea: influence of observation density, chemical transport model resolution, and geostationary remotely sensed AOD. Atmos. Environ. 331, 120603. https://doi.org/10.1016/j.atmosenv.2024.120603.

Van, N., Van Thanh, P., Tran, D., Tran, D.-T., 2023. A new model of air quality prediction using lightweight machine learning. Int. J. Environ. Sci. Technol. 20, 2983–2994. https://doi.org/10.1007/s13762-022-04185-w.

Yeo, M.J., Kim, Y.P., 2021. Long-term trends of surface ozone in Korea. J. Clean. Prod. 294, 125352. https://doi.org/10.1016/j.jclepro.2020.125352.

Zhao, X., Xia, K., Fan, S., Feng, X., 2023. Research on atmospheric pollution gas concentration detection system based on hybrid learning model. In: 2023 3rd International Conference on Electronic Information Engineering and Computer Communication (EIECC), pp. 1–5. https://doi.org/10.1109/EIECC60864.2023.10456665.