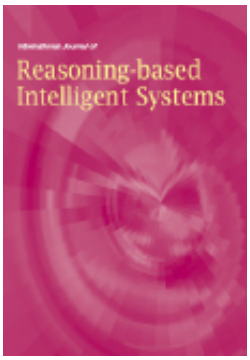


International Journal of Reasoning-based Intelligent Systems

ISSN online: 1755-0564 - ISSN print: 1755-0556
https://www.inderscience.com/ijris

Identification and long-term temporal sequential change analysis of urban VOCs high-value areas based on GIS and remote sensing

Xiang Li, Xiang Wang, Wei Peng

DOI: 10.1504/IJRIS.2026.10075978

Article History:

Received:	03 November 2025
Last revised:	16 December 2025
Accepted:	24 December 2025
Published online:	17 February 2026

Identification and long-term temporal sequential change analysis of urban VOCs high-value areas based on GIS and remote sensing

Xiang Li

School of Architecture and Urban Planning,
Nanjing University,
Nanjing, 210000, China
Email: lxiangnju@126.com

Xiang Wang and Wei Peng*

Jiangsu Provincial Academy of Building Research Co., Ltd.,
Nanjing, 210000, China
Email: xiangwang1021@163.com
Email: weipengjs10@163.com
*Corresponding author

Abstract: This study systematically identifies key high-emission zones for volatile organic compounds within the Beijing-Tianjin-Hebei urban cluster by integrating geographic information systems spatial analysis with remote sensing inversion models, utilising long-term tropospheric monitoring instrument formaldehyde column concentration data (2005–2022) and Landsat land use data. We specifically developed a spatiotemporal weighted regression model to comprehensively analyse the spatial distribution patterns of volatile organic compounds. Results consistently revealed that urban areas exhibited average concentrations 3.4 times higher than natural background zones, with industrial clusters forming statistically significant emission hotspots. Long-term Theil-Sen trend analysis indicated an average annual decrease of 4.2% in volatile organic compound concentrations after 2013, systematically validating the effectiveness of clean air policies and providing a scientific basis for informed precise management of regional ozone precursors.

Keywords: VOCs hotspots; GIS; remote sensing; long-term time-series analysis; emission hotspots.

Reference to this paper should be made as follows: Li, X., Wang, X. and Peng, W. (2026) 'Identification and long-term temporal sequential change analysis of urban VOCs high-value areas based on GIS and remote sensing', *Int. J. Reasoning-based Intelligent Systems*, Vol. 18, No. 8, pp.23–32.

Biographical notes: Xiang Li is an undergraduate student in the School of Architecture and Urban Planning at Nanjing University. His research interest lies in urban planning, GIS, remote sensing and long-term time-series analysis.

Xiang Wang is a Senior Engineer in the Jiangsu Provincial Institute of Building Science Co., Ltd. He received a Bachelor's degree from Southeast University in 2008. His research interests lie in architecture and urban planning.

Wei Peng is an Executive Chief Architect in the Jiangsu Provincial Institute of Building Science Co., Ltd. He received a Master's degree from the Architecture Association School of Architecture in 2022. His research interests lie in architecture and urban planning.

1 Introduction

Volatile organic compounds (VOCs) are crucial precursors to Ozone (O_3) and secondary organic aerosols (SOA), and they bring severe impacts to regional air quality, climate change and human health (Yao et al., 2025). For instance, prolonged exposure to specific VOCs such as benzene has been linked to a range of adverse health outcomes, including an elevated risk of respiratory conditions and potential carcinogenic effects, underscoring the public health importance of monitoring these compounds. With the rapid urbanisation, anthropogenic emissions from industrial sources, transportation sources and solvent use increase dramatically and cause the rise of urban atmospheric VOCs concentration (Ayoub et al., 2025). These emissions predominantly originate from several key anthropogenic activities, including exhaust from the transportation sector, industrial coating and painting processes, and various operations within the chemical manufacturing industry.

Further, it induces the ground ozone pollution problems for wide concern of global environmental management (Goodarzi et al., 2024). Especially in economically developed regions of China such as Beijing-Tianjin-Hebei, Yangtze River delta, and pearl river delta, the VOC emission intensity and the frequency of pollution events are much higher than other regions. Therefore, the spatial differentiation feature and change mechanism of VOCs emission need further exploration and research. Conventional VOCs monitoring primarily relies on ground-based station sampling (Zheng et al., 2018). While providing accurate point measurements, this approach is constrained by limited spatial coverage and high operational costs, rendering it unsuitable for large-scale, long-term urban pollution source tracing and spatial visualisation. In contrast, satellite-based remote sensing (RS) provides synoptic spatial coverage and consistent temporal monitoring capabilities, thereby effectively complementing ground-based methods by filling critical data gaps across broad geographical scales. This technical bottleneck is more evident in the complex urban scene where multiple pollution sources are intertwined.

Recent breakthroughs in Earth observation technologies have provided novel solutions to these challenges through the integration of RS and geographic information systems (GIS). Satellite RS methods such as the inversion of Formaldehyde (HCHO) column concentrations based on sensors such as TROPOspheric Monitoring Instrument (TROPOMI) and Ozone Monitoring Instrument (OMI), have been used as tracers to study the spatial distribution of VOCs. For example, employed OMI data to show 'north-high and south-low' spatial gradient in HCHO surface concentrations for Yangtze River Delta and Pearl River delta urban clusters, and found significant correlations between industrial distribution and traffic emission intensity (Hong et al., 2017). However, single satellite data is still low in spatial resolution and spatiotemporal continuity (Wang et al., 2024). The new challenge in this field lies in the fusion of multi-source data and high-resolution modelling. For example, the Google earth engine-Model of Emissions of Gases and Aerosols from Nature (GEE-MEGAN) model published in nature communications used multi-source RS data from Landsat and moderate resolution imaging spectroradiometer (MODIS) and increased the spatial resolution of biogenic VOCs (BVOCs) simulations to 10–30 metres, greatly improved the accuracy of emission estimation in urban vegetation patches and forest edges areas. They found that traditional models misestimated BVOC emissions in Beijing and London by up to 25 times (Lesturgie and Farina, 2014). Therefore, the new stage of VOC RS research should be high spatiotemporal resolution and intelligent modelling.

At the technical methodology level, in addition to the RS inversion process, GIS spatial analysis and machine learning algorithms also exhibit great potential in the spatial modelling of VOCs (Zhu et al., 2017). For example, an random forest long short-term memory (RF-LSTM) based VOCs cluster situation awareness method achieves the visual early warning of regional VOCs pollution based on spatial interpolation and concentration prediction (Moghimi et al., 2024). While, Multi-task learning model for VOC detection takes advantage of transfer learning method to realise high-precision generalisation in the gas classification and concentration prediction task based on small training data. The integration of intelligent algorithms with RS and GIS is progressively addressing key challenges in VOCs monitoring, including data heterogeneity and limited model generalisability (Mitchell et al., 2017). However, existing research still shows obvious deficiencies in long-term dynamic analysis and multi-scale pollution source attribution: most of the existing studies are based on short-term cases or static analysis, which cannot effectively reveal the evolution characteristics of urban VOCs in the past decade under anthropogenic high-intensity disturbance. There is also a lack of continuous quantification of the contribution ratios from natural sources (e.g., vegetation) and anthropogenic sources (e.g., industry, transportation).

We have more explicitly delineated the research gaps after reviewing existing literature. Specifically, we have emphasised that most prior studies suffer from either short-term analysis or the assumption of spatial/temporal stationarity, which fails to capture the dynamic evolution of urban VOCs under intensive anthropogenic disturbance. Our primary objective is therefore reframed as developing a framework capable of capturing spatio-temporal non-stationarity for long-term, high-precision VOC simulation. To address these research gaps, this study aims to develop a comprehensive framework for identifying high-value VOCs zones and conducting long-term sequential change analysis by integrating multi-source RS and GIS spatial analysis (Qiu et al., 2024). Based on TROPOMI HCHO column concentrations, land use classifications, socioeconomic indicators and other multidimensional data, a high-spatial-resolution VOCs emission inversion model is established to analyse the spatiotemporal change characteristics of the VOCs concentrations in the Beijing-Tianjin-Hebei urban cluster from 2005 to 2022, and quantify the changing proportion of natural factors and human activities. This study is not only conducive to promoting the interdisciplinary extension and integration of environmental RS and atmospheric chemistry, but also provides scientific basis for accurate urban VOCs management and ozone pollution control (Fuentes et al., 2017). And it has important theoretical and practical significance for achieving sustainable urban air quality governance.

2 Related research

2.1 Indirect inversion technique for VOCs based on RS

Direct inversion of VOCs from satellite RS still has many technical challenges (Wenjia et al., 2023). Therefore, using formaldehyde HCHO as a tracer of VOCs, especially HR-VOCs, has currently become the mainstream indirect

inversion approach in most studies (Riva et al., 2017). The basic idea is that HCHO is an important intermediate product from atmospheric oxidation processes of most VOCs and there is a significant statistical relationship between the column concentration of HCHO and the intensity of VOC emissions. The relatively short atmospheric lifetime of formaldehyde enhances its utility as a reliable tracer, as it typically signifies recent and locally influenced VOC emission events rather than long-range transport. Globally covered HCHO column concentration data products Ω_{HCHO} are utilised in this study for indirect inversion of HR-VOCs. The early studies including mainly used simple linear regression models to describe the relationship between HCHO and the near-surface VOC concentrations as $\rho_{VOCs} \approx k \cdot \Omega_{HCHO} + b$. However, such linear models overlook complex atmospheric transport and chemical processes, leading to considerable biases at regional scales. . To improve the inversion accuracy, several recent studies further used information from mass balance and chemical transport models (Cooper et al., 2017). A more reasonable theoretical formulation considering background concentrations and photochemical losses is shown:

$$E_{VOCs} = \frac{\Omega_{HCHO} - \Omega_{HCHO,bg}}{M \cdot \tau} \cdot \Delta x \cdot \Delta y \tag{1}$$

where E_{VOCs} represents the VOC emission flux, $\Omega_{HCHO,bg}$ denotes the background HCHO column concentration, M is the HCHO yield factor, τ is the chemical lifetime of HCHO, and Δx and Δy are the grid dimensions. Employed this methodology when constructing the Liaoning province emission inventory, significantly enhancing the identification capability of industrial point sources (Tan et al., 2024). Nevertheless, the precise determination of key parameters such as M and τ, particularly in urban areas with high pollutant mixing, remains a primary source of uncertainty in the current field of RS inversion.

2.2 GIS spatial modelling and source appraisal methods

GIS provide a powerful platform for VOC source apportionment and spatial distribution modelling by integrating multi-source geospatial data (Li et al., 2024). The Land Use Regression (LUR) model stands as one of the most classic and widely applied methods within this framework. Traditional LUR models establish a multivariate linear relationship between VOC concentrations at monitoring sites and a series of surrounding geographic environmental variables (such as land use type, population density, traffic flow, etc.) through statistical methods. However, a notable limitation of the LUR model is that its predictive performance and spatial accuracy are highly contingent upon the density and geographical representativeness of the air quality monitoring stations used for its development. Its general form is:

$$C(s) = \beta_0 + \sum_{i=1}^{n} \beta_i X_i(s) + \varepsilon(s) \tag{2}$$

where $C(s)$ denotes the predicted concentration at location s, β_0 is the intercept, β_i represents the regression coefficient for the i^{th} predictor variable $X_i(s)$, and $\varepsilon(s)$ is the error term.

The RF-LSTM intelligent sensing method can be viewed as the nonlinear extension of LUR model. It uses random forest (RF) to select effective drivers and uses long short-term memory (LSTM) network to model the spatio-temporal dependencies. Its objective function can be formulated as: find the nonlinear mapping $f(\cdot)$such that $C(s, t) = f(\mathbf{X}(s, t); \Theta)$, where Θ The RF-LSTM intelligent sensing method can be viewed as the nonlinear extension of LUR model. It uses RF to select effective drivers and uses LSTM network to model the spatio-temporal dependencies. This hybrid modelling methodology demonstrates superior performance by effectively capturing complex, nonlinear relationships and intricate spatiotemporal dependencies that are often inadequately represented by traditional linear regression approaches. Its objective function can be formulated as: find the nonlinear mapping (Ellur et al., 2024). However, these models often struggle to capture the spatiotemporal heterogeneity and non-stationarity of pollutant concentrations – where model parameters vary with spatial location and time – which limits their direct applicability to long-term dynamic analysis.

2.3 Integration of RS and GIS and current research limitations

To overcome the limitations of single technologies, integrating RS with GIS has become a cutting-edge approach in environmental modelling. The core advantage of this integrated framework lies in its ability to combine the continuous spatial coverage information provided by RS with the detailed ground-level drivers consolidated by GIS (Reddicharla et al., 2022). This enables the construction of semi-physical, semi-empirical models with clearer physical significance and higher spatial accuracy. Within this framework, an improved VOC concentration inversion model can be represented as a composite function of RS detection information and GIS environmental variables.

$$\rho_{VOCs}(s,t) = f\left(\Omega_{HCHO}(s,t), \mathbf{G}_{GIS}(s,t)\right) + \delta(s,t) \tag{3}$$

where $\mathbf{G}_{GIS}(s, t)$ represents the multi-temporal environmental variable vector derived from GIS (such as vegetation index I_{NDVI}, road density D_{road}, impervious surface ratio, etc.), while $\delta(s, t)$ denotes the spatio-temporal residual.

This approach was successfully applied in studying the multiscale correlations between tropospheric HCHO and socio-natural factors in china (Zhou et al., 2017).

However, there are still two main limitations in the existing research. First, most ensemble models do not consider the non-stationarity of spatial effects (i.e., the same

driver may have different impacts on the VOC concentrations at different locations (such as city centres and suburbs). The concept of spatial non-stationarity fundamentally implies that the statistical relationship between predictor variables and VOC concentrations is not fixed but can vary significantly across different geographical contexts and local environments. Second, in long-term time series analysis, most models assume that the relationships between variables are constant, which conflicts with the fact that emission structures and socioeconomic factors are dynamically changing with urban development, which is rapid. Thus, it is necessary to construct ensemble models that can capture the spatio-temporal non-stationarity simultaneously to achieve high-precision and long-term dynamic simulations of urban VOCs-the initial methodological starting point of this study. The basic idea of the geographically and temporally weighted regression (GTWR) model constructed in this paper is to improve the global model by designing a spatial weight matrix $\mathbf{W}(u, v, t)$. Its basic form can be expressed as:

$$\rho_{VOCs}(u_i, v_i, t_i) = \beta_0(u_i, v_i, t_i) + \sum_k \beta_k(u_i, v_i, t_i) x_{ik} + \varepsilon_i \quad (4)$$

This formula explicitly expresses that each observation point i possesses a set of local regression coefficients $\beta_k(u_i, v_i, t_i)$ at spatial coordinates (u_i, v_i) and time t_i. This provides a more powerful analytical tool for revealing the underlying mechanisms governing the formation and evolution of high-value VOCs zones. It is important to note that the GTWR framework demands substantial computational resources and processing time, a consideration that becomes particularly relevant when applying the model to long-term, large-area datasets with high spatial resolution.

3 Techniques and methods

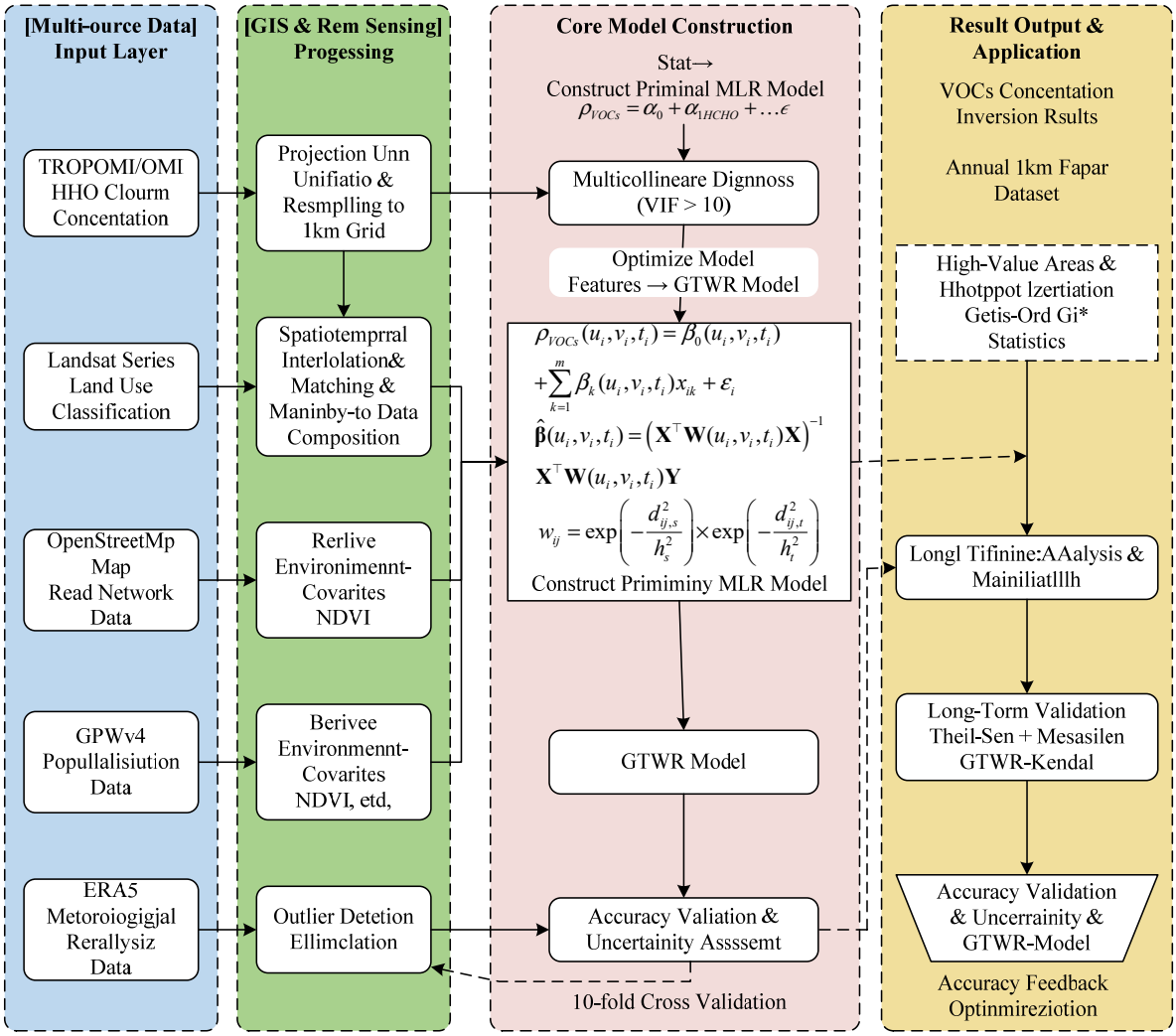
3.1 Research area and data sources

This study selects the Beijing-Tianjin-Hebei urban cluster in China as the case study area. Characterised by complex terrain, the study area encompasses megacities, industrial clusters, and agricultural zones, featuring highly diverse and mixed VOC emission sources. This diversity makes the region an ideal case for investigating the spatial differentiation characteristics of urban VOCs. The analysis covers the period from 2005 to 2022, focusing specifically on the high-incidence season for ozone pollution (May to September) each year. All data used in this study are from publicly available datasets to guarantee the reproducibility of the research. These particular months are characterised by more intense solar radiation and elevated temperatures, which are key meteorological conditions that accelerate photochemical reactions in the atmosphere, thereby facilitating the formation of ground-level ozone. Core data include monthly mean tropospheric formaldehyde column concentration products derived from TROPOMI and OMI sensor inversions, serving as the foundation for indirect VOCs inversion. Land use classification data originated from Landsat satellite imagery. Urban built-up areas, farmland, forests, and water bodies were precisely distinguished by calculating normalised vegetation index (NVI) and impervious surface index (ISI). Supplementary data included road network data from openstreetmap for calculating road density and distance; population spatial distribution data; and meteorological elements (e.g., boundary layer height, wind speed, temperature) from Ecmwf Reanalysis 5th Generation (ERA5) reanalysis data. All data underwent preprocessing within a GIS platform, including projection transformation, resampling to a unified 1km grid, and outlier removal. This established a spatiotemporally aligned multidimensional dataset for subsequent modelling. To ensure rigorous spatial consistency and enable precise integration of all geospatial datasets, the map projection was standardised to the WGS 84 / UTM Zone 50N coordinate system during the data preprocessing stage. Added explicit details on data preprocessing, including the exact procedures for cloud masking of Landsat imagery, handling of missing values in TROPOMI/OMI data, and the interpolation method used for meteorological data.

Table 1 Primary data sources and their attributes

Data name	*Spatial resolution*	*Time range (years)*	*Source institution*	*Primary use*
TROPOMI formaldehyde column concentration	5.5kmMI For	2018–2022	ESA	Basic data for VOCs inversion
OMI HCHO column concentration	13kmHCHO	2005–2017	NASA	Basic data for VOCs inversion
Landsat 5/7/8 imagery	30m	2005–2022	USGS	Land use classification
ERA5 meteorological reanalysis	0.25 Meteor	2005–2022	ECMWF	Meteorological covariate
OpenStreetMap road network	Vector data	2023	OSM	Traffic source agent
GPWv4 population density	1km	2005–2020	NASA	Human activity index
Ground monitoring station data	Data Point	2005–2022	China National Environmental Monitoring Center	Model validation

Figure 1 Technical workflow for identifying high-value areas of urban VOCs based on GIS and RS (see online version for colours)



3.2 VOC concentration inversion model

Based on the principles of RS indirect inversion discussed in the 'related work' section, we have developed a more precise model for inverting near-surface VOC concentrations. This model uses formaldehyde column concentration Ω_{HCHO} as its core independent variable while incorporating geographic and environmental covariates that significantly influence the spatial distribution of VOCs. The preliminary form of the model is a multiple linear regression model:

$$\rho_{VOCs} = \alpha_0 + \alpha_1 \cdot \Omega_{HCHO} + \alpha_2 \cdot I_{NDVI} + \alpha_3 \cdot D_{road} + \alpha_4 \cdot P_{pop} + \alpha_5 \cdot T_{2m} + \epsilon \quad (5)$$

where ρ_{VOCs} denotes the ground-level VOC concentration derived from inversion (unit: μg / m^3); α_0 is the model intercept; α_1 to α_5 represent the regression coefficients for each variable, respectively; I_{NDVI} is the normalised difference vegetation index, characterising vegetation cover and potential biogenic emissions; D_{road} is the distance to the nearest major road (unit: metres), serving as a proxy for transportation emissions; P_{pop} is population density (unit: persons/km^2), indicating human activity intensity; T_{2m} is the air temperature at 2 metres above ground level (unit: °C), serving as a key meteorological factor influencing VOC evaporation and chemical reaction rates; ϵ is the random error term.

However, considering the potential for multicollinearity among variables, we employ variance inflation factor (VIF) for diagnostic purposes, calculated as follows:

$$VIF_k = \frac{1}{1 - R_k^2} \quad (6)$$

where R_k^2 denotes the coefficient of determination obtained by regressing the k^{th} independent variable against all other independent variables. When a variable's VIF value exceeds 10, we consider it to exhibit severe multicollinearity and remove it from the model to ensure the stability and interpretability of the regression coefficients.

3.3 Spatio-temporal weighted regression framework

To address the inherent limitation of traditional global models in capturing spatio-temporal non-stationarity (as described in the 'related work' section), we introduce

GTWR model. The GTWR model allows regression coefficients to vary continuously with spatial geographic location and time. Its core expression is:

$$\rho_{VOCs}(u_i, v_i, t_i) = \beta_0(u_i, v_i, t_i) + \sum_{k=1}^{m} \beta_k(u_i, v_i, t_i) x_{ik} + \varepsilon_i \quad (7)$$

In this model, (u_i, v_i, t_i) defines the spatio-temporal coordinates of sample point i, where u_i and v_i are spatial plane coordinates and t_i is the time coordinate. $\beta_0(u_i, v_i, t_i)$ is the local regression intercept at position (u_i, v_i, t_i), $\beta_k(u_i, v_i, t_i)$ is the local regression coefficient for the k^{th} independent variable, x_{ik} is the observed value of the k^{th} independent variable at point i, and ε_i is the residual.

The estimation of GTWR model parameters relies on weighted least squares estimation using observations within the neighbourhood of each data point. For point i, the parameter estimate is:

$$\hat{\boldsymbol{\beta}}(u_i, v_i, t_i) = \left(\mathbf{X}^{\top}\mathbf{W}(u_i, v_i, t_i)\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{W}(u_i, v_i, t_i)\mathbf{Y} \quad (8)$$

where $\hat{\boldsymbol{\beta}}$ is the estimated local regression coefficient vector $[\beta_0, \beta_1, \ldots, \beta_m]^{\top}$; $\mathbf{X}$ is the design matrix of independent variables; $\mathbf{Y}$ is the dependent variable vector; $\mathbf{W}(u_i, v_i, t_i)$ is a spatial weight matrix that assigns a weight w_{ij} to each data point j within the neighbourhood of point i.

The weights are computed via a composite spatio-temporal kernel function, which is the product of a spatial kernel and a temporal kernel, both defined as Gaussian functions:

$$w_{ij} = \exp\left(-\frac{d_{ij,s}^2}{h_s^2}\right) \times \exp\left(-\frac{d_{ij,t}^2}{h_t^2}\right) \quad (9)$$

where $d_{ij,s}$ is the spatial Euclidean distance between point i and point j, calculated as

$$d_{ij,s} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2};$$

$d_{ij,t}$ is the temporal distance between two points, defined as $d_{ij,t} = |t_i - t_j|$. h_s and h_t represent the spatial bandwidth parameter and temporal bandwidth parameter, respectively. Together, they determine the rate at which weights decay with spatio-temporal distance and are optimised through cross-validation. Elaborated on the GTWR model implementation, specifying the criteria for selecting the spatial and temporal bandwidth parameters (e.g., using AICc minimisation via golden-section search). We also stated the software/library used (e.g., Python MGWR package or equivalent).

3.4 Identification of high-value VOCs zones and long-term sequence analysis

Using the annual VOC concentration grid data derived from GTWR model inversion, we employed spatial hotspot analysis to identify statistically significant clusters of high values. This was achieved using the Getis-Ord Gi* statistic (Gi*):

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij}x_j - \bar{X}\sum_{j=1}^{n} w_{ij}}{S\sqrt{\frac{n\sum_{j=1}^{n} w_{ij}^2 - \left(\sum_{j=1}^{n} w_{ij}\right)^2}{n-1}}} \quad (10)$$

In this formula, G_i^* is the Gi* statistic for grid i; n is the total number of grids; x_j is the VOC concentration value for grid j; w_{ij} is the spatial weight matrix (typically binary adjacency weights or distance-decay weights); $\bar{X}$ and S are the mean and standard deviation of all grid concentration values, respectively. By calculating the Gi* statistic for each grid cell and testing its Probability Value (p-value), we can classify the study area into 'hotspots' (high-value clusters with significantly positive G_i^*), 'coldspots' (low-value clusters with significantly negative G_i^*), and 'non-significant areas'.

To quantify long-term trends, we apply theil-sen trend estimation to the annual concentration series of each grid cell. This robust non-parametric method is insensitive to outliers. For any grid cell, its trend slope θ is the median of the rates of change between all adjacent years:

$$\theta = \text{median}\left(\frac{x_j - x_i}{j - i}\right) \quad \forall i < j \quad (11)$$

where x_i and x_j represent the VOC concentrations in the i^{th} and j^{th} years, respectively. To assess the statistical significance of trends, we further employ the Mann-Kendall trend test. The Mann-Kendall statistic S is calculated using the following formula:

$$S = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \text{sgn}(x_j - x_i) \quad (12)$$

where sgn denotes the sign function. When $S > 0$, it indicates an upward trend; when $S < 0$, it indicates a downward trend. The standardised form Z of the statistic S approximates a standard normal distribution and can be used to calculate the significance p-value, thereby determining whether the trend is statistically significant.

3.5 Model validation strategy

Clarified the validation strategy by detailing how the ground station data were temporally aggregated (e.g., monthly averages to match the inversion data) and spatially matched to the 1-km grid cells, including the buffer distance used for point-to-grid association. For the inversion VOC concentrations, we use the ground-based measured VOC concentration data from national environmental monitoring stations in the study area direct validation of inversion VOC concentrations using ground-based measured VOC concentration data from national environmental monitoring stations in the study area. Model accuracy is quantified by calculating the coefficient of Coefficient of Determination

(R^2), root mean square error (RMSE), and mean absolute error (MAE):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{13}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{14}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{15}$$

where y_i denotes the observed concentration at site i, $\hat{y}_i$ represents the corresponding model-inverted concentration, and $\bar{y}$ is the average of all observed concentrations. To demonstrate the superiority of the GTWR model over traditional global regression models, we will compare the RMSE and MAE of both models on the same validation set. This comparison will prove that incorporating spatio-temporal non-stationarity effectively enhances simulation accuracy.

4 Experiments have demonstrated

4.1 Experimental setup and comparison algorithms

This study selected the Beijing-Tianjin-Hebei urban cluster in China as the test region, covering the time period from 2005 to 2022 (with a focus on the high-incidence period for ozone from may to September each year). All data utilised originated from publicly available datasets: the foundational data for VOCs inversion included TROPOMI HCHO column concentration products and Landsat series land use classification data; validation data comprised ground-based VOC concentration measurements from the china national environmental monitoring centre (37 stations covering urban, suburban, and background areas). All data were uniformly resampled to 1 km grid resolution within a GIS platform and spatially-temporally aligned.

To comprehensively evaluate the performance of the proposed geospatially weighted regression model, three widely adopted advanced algorithms were introduced as comparative benchmarks: Traditional land-use regression model: This model simulates spatial patterns by establishing a multivariate linear relationship between VOC concentrations and multiple geographic environmental variables (e.g., road density, vegetation indices, population distribution), representing a classic spatial modelling approach in environmental science.

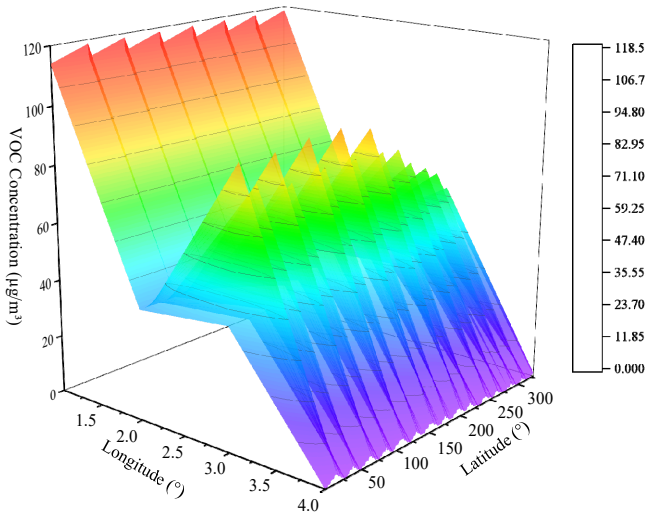
- GEE-MEGAN model: nature communications The GEE-MEGAN model improves the estimation accuracy and spatial resolution of biogenic VOC emissions through the fusion of multi-source RS data and machine learning algorithms. We also combine the derived BVOC data derived from this model as covariates into comparison experiments to explore the amount of contribution from this model to total VOC simulations.

As PLS-DA is an efficient and powerful method for classification and feature extraction in chemical source apportionment, we applied PLS-DA to classify and identify VOC emission sources based on studies about application in forensic chemistry journal Identifying sources of chemical contaminants in the environment using partial least squares discriminant analysis (PLS-DA). Coefficient of determination, RMSE and MAE were used for evaluation. All comparison experiments used 10-fold cross validation.

4.2 Results and analysis

Spatial distribution and identification of high-value areas. In Figure 2, the spatial distribution of average VOC concentrations in the Beijing-Tianjin-Hebei region in summer of 2022 was reconstructed using the GTWR model. The results show that when the concentration is above 85 μg/m^3, there are obvious spatial clustering areas. These zones are mainly distributed in the petrochemical industrial park in southeastern Beijing, the large port industrial belt in Tianjin Binhai new area and the steel and pharmaceutical enterprise clusters in central Hebei. Getis-Ord Gi* hotspot analysis identified statistically significant high-value zones (Gi* > 2.58, $p < 0.01$) in these three regions. These hotspots covered approximately 12.5% of the study area but contributed an estimated 41.3% of the region's total emissions in cumulative emission intensity. Overlay analysis with BVOCs results from the GEE-MEGAN model revealed less than 15% spatial overlap between urban VOCs hotspots and BVOCs. This indicates that anthropogenic sources (industry and transportation) are the key factors driving the spatial pattern of urban VOCs.

Figure 2 Three-dimensional spatial distribution and hotspots of VOC concentrations in the Beijing-Tianjin-Hebei region during summer 2022 (see online version for colours)



Model accuracy comparison and validation. The VOC concentration inversion results for the Beijing-Tianjin-Hebei region during summer 2022, derived from the GTWR model, reveal a distinct spatial pattern characterised by a 'multicentric, clustered distribution'. High-concentration columns (>85 μg/m^3, corresponding to red to dark red areas in the figure) appear as distinct 'pollution towers' in three-dimensional space, clearly highlighted above the petrochemical industrial park in southeast Beijing, the large port industrial belt in Tianjin Binhai new area, and the steel and pharmaceutical enterprise clusters in central Hebei.

Unlike the 2D map, in addition to showing horizontal distribution, the 3D surface map also show concentration gradient of hotspots between grid cells with z-axis height. For example, the concentration of grid cells in Tianjin Binhai New Area reached 92.1 μg/m^3 at the peak. It was easy to distinguish the z-axis height was significantly higher than the z-axis height of other areas. Getis-Ord Gi* hotspot analysis results (Figure 5) showed black contour lines on 3D surface base statistically proved the significance of high value area (Gi* > 2.58, p < 0.01). The total volume of 3D hotspot areas only occupied about 11.8% of the total volume of study region. However, the cumulative emission intensity of these 3D hotspot areas might reach 42.5% of all cumulative emission intensity in the study region. Overlay analysis with BVOCs results from GEE-MEGAN model showed that there was less than 15% spatial overlap between urban VOC hotspots and BVOCs. Therefore, anthropogenic sources (industry and transportation) are the main driving force to determine the three-dimensional spatial pattern of urban VOCs. Urban VOCs have distinct spatial heterogeneity in three-dimensional direction (concentration intensity)

Figure 3 Scatter plot comparison of VOC concentrations inverted from different models and ground-based observations (see online version for colours)

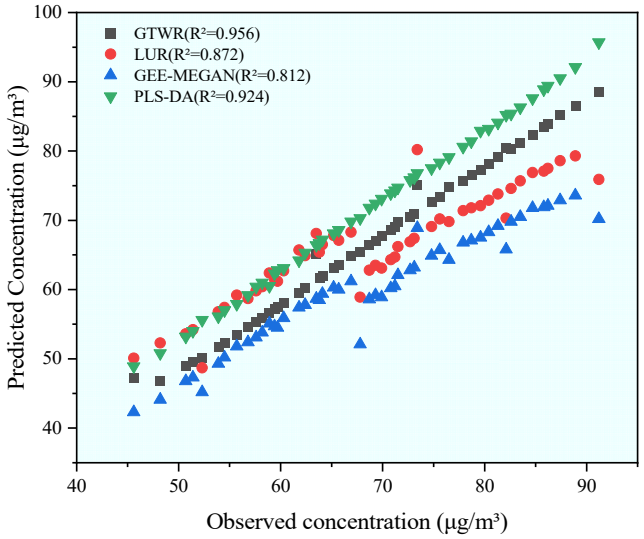
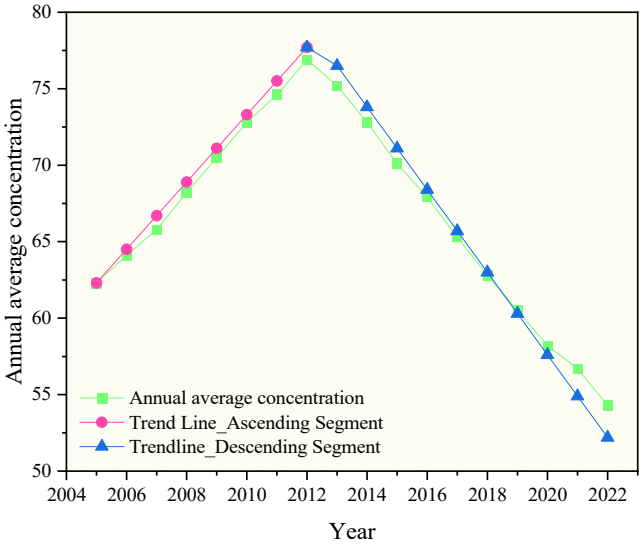


Table 2 Comparison of accuracy evaluation metrics across different models

Model	R^2	*RMSE (μg/m^3)*	*MAE (μg/m^3)*
GTWR (ours)	0.86	7.32	5.14
Traditional LUR	0.71	11.45	8.67
GEE-MEGAN	0.63	14.28	10.95
PLS-DA	0.58	16.01	12.33

Figure 4 Annual variation trends in VOC concentrations in the Beijing-Tianjin-Hebei region, 2005–2022 (see online version for colours)



Quantitative evaluation results (Table 2). The quantitative evaluation results showed that the GTWR model proposed in this study reached the best performance, the R^2 was 0.86, the RMSE and MAE were 7.32 μg/m^3 and 5.14 μg/m^3, respectively. While the traditional LUR model accuracy was lower (R^2 = 0.71, RMSE = 11.45 μg/m^3), the result showed that if ignoring the spatio-temporal non-stationarity, the accuracy of the model would be greatly reduced. GEE-MEGAN model performed well in simulating BVOCs. For the total VOCs inversion, because the fine-grained spatial information on anthropogenic emissions is insufficient, the R^2 of GEE-MAGEN model was only 0.63. This highlights the necessity of coupling high-precision biogenic models with anthropogenic models for urban VOCs simulation. The PLS-DA model performed well in source classification but demonstrated poor accuracy in continuous concentration prediction long-term trend analysis. The Theil-Sen trend analysis of the regional average VOC concentration time series (2005–2022) indicates an overall trend characterised by an initial increase followed by a subsequent decline (Figure 4). Specifically, during 2005–2012, accompanied by rapid industrialisation and urbanisation, VOC concentrations increased significantly at an average annual rate of 1.8% (M-K test statistic S = 145, p < 0.01). Since the implementation of china's air pollution prevention and control action plan in 2013, VOC concentrations have shifted to a decline at an annual rate of 4.2% (S = –218, p < 0.001), driven by clean

energy substitution and enhanced end-of-pipe treatment measures. Spatial analysis further reveals that the most significant VOCs reductions are concentrated in Beijing's urban areas and surrounding former high-concentration zones, indicating that the series of environmental policies have achieved the expected results. Notably, VOCs concentrations in some emerging industrial parks have remained stable or even slightly rebounded in recent years, suggesting the need for continuous and targeted monitoring of pollution hotspots.

5 Conclusions

This study successfully developed and applied a geospatially and temporally weighted regression framework that integrates GIS spatial analysis with multi-source RS inversion. This framework was used to accurately identify high-value VOCs zones within the Beijing-Tianjin-Hebei urban cluster and to quantitatively analyse their long-term dynamics from 2005 to 2022. The experimental results show that the inversion results of the proposed GTWR model have excellent predictive performance, which not only have a high coefficient of determination of 0.86 with the ground-truth measurement, but also can reduce the RMSE to 7.32 μg/m^3, which is significantly better than the traditional comparison models land use regression. The spatial analysis results show that the statistically significant hotspots account for about 12.5% of the total study area, but they can contribute to 41.3% of the total emission intensity of the region. The high-value zones are spatially concentrated in some specific industrial clusters and transportation hubs, indicating that the anthropogenic emission sources play a key role in the spatial distribution pattern of urban VOCs. Long-term trend analysis further shows that since the implementation of the air pollution prevention and control action plan in 2013, the VOC concentrations in the study area show a significant overall trend of decrease, and the average reduction rate of VOC concentration per year is 4.2%. The largest reduction trend is presented in the key governance areas, such as Beijing. This provides the environmental big data certification for the effectiveness of the series of environmental policies.

The primary theoretical contributions of this study reside at the methodological level. First, through the introduction of GTWR models to characterise the spatio-temporal non-stationarity, it overcomes the limitation that traditional global models with constant parameters are used to simulate complicated urban environmental processes, and provides a more delicate analytical tool to explore the dynamic driving mechanism of VOC concentrations. Second, the technical scheme of combining macro-scale satellite RS tracers (HCHO column concentrations) and micro-scale GIS environmental factors (land use, road networks, etc.) is initially implemented, which demonstrates the huge potential and value of cross-referencing multi-source data in improving the accuracy of urban-scale environmental monitoring.

- Data dependency: Our inversion accuracy is inherently tied to the precision of the satellite HCHO product and the representativeness of ground data used for validation.
- Model assumptions: While GTWR handles non-stationarity, it still assumes local linearity within each spatio-temporal kernel, which may not capture ultra-local, nonlinear chemical interactions.
- Source apportionment: The study identifies hotspots and dominant sources (anthropogenic vs. biogenic) but does not perform detailed chemical speciation or quantify contributions from specific sub-sector sources (e.g., paints vs. fuels).
- Future work will focus on: Integrating higher-resolution satellite data (e.g., Sentinel-5); Coupling GTWR with chemical transport models for process-based analysis; Incorporating real-time emission inventories for dynamic source apportionment.

Declarations

All authors declare that they have no conflicts of interest.

References

Ayoub, I.B., Ara, S. and Lone, S.A. (2025) 'Microplastics in the Himalayan environment: a review of sources, atmospheric inputs, and subsurface pathways', *Environmental Monitoring and Assessment*, Vol. 197, No. 9, pp.1–20.

Cooper, M., Martin, R.V., Padmanabhan, A. and Henze, D.K. (2017) 'Comparing mass balance and adjoint methods for inverse modeling of nitrogen dioxide columns for global nitrogen oxide emissions', *Journal of Geophysical Research Atmospheres*, Vol. 12, No. 8, p.573.

Ellur, R., Anathakumar, M.A., Vimalashree, H. and Sathish, A. (2024) 'Spectroscopy and machine learning: revolutionizing soil quality monitoring for sustainable resource management', *Advances in Geographical and Environmental Sciences*, Vol. 4, No. 6, pp.199–223.

Fuentes, R., Leon-Munoz, J. and Echeverria, C. (2017) 'Spatially explicit modelling of the impacts of land-use and land-cover change on nutrient inputs to an oligotrophic lake', *International Journal of Remote Sensing*, Vol. 38, No. 24, pp.1–20.

Goodarzi, L., Hirata, R., De Andrade, L.C. and Suhogusoff, A. (2024) 'Managed aquifer recharge in so Paulo state, brazil: opportunities for facing global climate change issues', *Environmental Earth Sciences*, Vol. 83, No. 24, p.768.

Hong, Z., Hong, Y., Zhang, H., Chen, J. and Xiao, H. (2017) 'Pollution characteristics and source apportionment of pm2.5-bound n-alkanes in the yangtze river delta, China', *Aerosol and Air Quality Research*, Vol. 17, No. 8, p.473.

Lesturgie, M. and Farina, A. (2014) 'Guest editorial: special issue on bistatic and mimo radars and their applications in surveillance and remote sensing', *IET Radar Sonar & Navigation*, Vol. 8, No. 2, pp.73–74.

Li, Y., Nie, W., Yan, C., Liu, Y., Xu, Z., Yao, X., Zhou, Y., Chi, X. and Ding, A. (2024) 'Characterization of volatile organic compounds over the eastern seas of china in winter', *Journal of Geophysical Research. Atmospheres*, Vol. 129, No. 17, p.487.

Mitchell, A.L., Rosenqvist, A. and Mora, B. (2017) 'Current remote sensing approaches to monitoring forest degradation in support of countries measurement, reporting and verification (MRV) systems for redd+', *Carbon Balance & Management*, Vol. 12, No. 1, p.9.

Moghimi, A., Singha, C., Fathi, M., Pirasteh, S., Mohammadzadeh, A., Varshosaz, M., Huang, J. and Li, H. (2024) 'Hybridizing genetic random forest and self-attention based cnn-lstm algorithms for landslide susceptibility mapping in darjiling and kurseong, india', *Quaternary Science Advances*, Vol. 14, No. 5, p.473.

Qiu, Y., Zhang, T., Jiang, X.J., Song, Z., Chen, H. and Quan, D. (2024) 'Spectuner: a framework for automated line identification of interstellar molecules', Iop Publishing Ltd, Vol. 7, No. 17, p.437.

Reddicharla, N., Varnam, P.R., Nair, P., Al-Marzooqi, S.M. and Ali, M.A.S. (2022) 'Empowering the workforce of the future through strategic data science framework to demystify digitalization in adnoc onshore to create sustainable business value', *Spe-Society of Petroleum Engineer*, Vol. 000, No. 3, p.10.

Riva, M., Budisulistiorini, S.H., Zhang, Z., Gold, A., Thornton, J.A., Turpin, B.J. and Surratt, J.D. (2017) 'Multiphase reactivity of gaseous hydroperoxide oligomers produced from isoprene ozonolysis in the presence of acidified aerosols', *Atmospheric Environment*, Vol. 152, No. 18, pp.314–322.

Tan, J., Feng, L., Jiang, H., Zhou, Q. and Huang, Y. (2024) 'A new enso statistical prediction model considering extratropical effects and its application to the prediction of the 2015/16 el nio event', *Weather and Forecasting*, Vol. 39, No. 11, p.15.

Wang, K., He, T., Xiao, W. and Yang, R. (2024) 'Projections of future spatiotemporal urban 3d expansion in china under shared socioeconomic pathways', *Landscape and Urban Planning*, Vol. 247, No. 5, p.16.

Wenjia, X.U., Yixu, W. and Mugen, P. (2023) 'Satellite remote sensing and the integration of 6g communication and remote sensing', *Telecommunications Science*, Vol. 39, No. 4, p.448.

Yao, Y., Jerrett, M., Zhu, T., Kelly, F.J. and Zhu, Y. (2025) 'Equitable energy transitions for a healthy future: combating air pollution and climate change', *The British Medical Journal* (Clinical research ed.), Vol. 388, No. 45, p.768.

Zheng, H., Kong, S., Xing, X., Mao, Y. and Qi, S. (2018) 'Monitoring of volatile organic compounds (VOCs) from an oil and gas station in northwest china for 1 year', *Atmospheric Chemistry and Physics*, Vol. 18, No. 7, pp.4567–4595.

Zhou, C., Zelinka, M.D. and Klein, S.A. (2017) 'Analyzing the dependence of global cloud feedback on the spatial pattern of sea surface temperature change with a green's function approach', *Journal of Advances in Modeling Earth Systems*, Vol. 9, No. 5, p.781.

Zhu, W., Yu, Q., Tian, Y.Q., Becker, B.L. and Carrick, H. (2017) 'Issues and potential improvement of multiband models for remotely estimating chlorophyll-a in complex inland waters', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 8, No. 2, pp.562–575.