International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 12 No. 2, 2026
https://www.theaspd.com/ijes.php

AI-Assisted Monitoring of Atmospheric Greenhouse Gases Over Bangalore Using Satellite And Reanalysis Data

Pranad Navada

Student, Greenwood High International School Chikkawadayapura, Bangalore, Karnataka 560087, India.
Email: pranad.navada@gmail.com

Abstract

Cities concentrate greenhouse gas emissions but rarely have the dense ground monitoring needed to locate them. This study tested whether satellite observations and machine learning can close part of that gap over Bangalore, India. Monthly Sentinel-5P columns of methane, carbon monoxide and nitrogen dioxide for 2020 to 2024 were combined with ERA5 reanalysis on a 0.05° grid, giving 5,619 cell-month records. Random Forest, XGBoost and LightGBM were trained to predict methane from 22 predictors and evaluated under shuffled, year-blocked and cell-blocked cross-validation. Hotspots were detected with HDBSCAN and the models interpreted with SHAP. All three algorithms agreed to within 0.0007 in R^2, reaching 0.956 on a shuffled hold-out split. That figure did not survive stricter validation. Blocking by grid cell left it at 0.936, while blocking by year drove every model to an R^2 near −1.57, worse than predicting the training mean. SHAP explains the asymmetry: season and year carry 54.6 percent of the attributed importance, and methane varies by only 0.86 percent across the grid against a seasonal swing of about 62 ppb. Two nitrogen dioxide hotspots covering 16 of 100 cells were identified, both attributed to traffic and both over the densest part of the road network in independent land-use data. Elevated cell-months were flagged at 88.2 percent with a false positive rate of 2.8 percent. The framework supports spatial gap-filling, hotspot localisation and the ranking of elevated periods. It does not support forecasting. A single shuffled-split score would have hidden that.

Keywords: *Sentinel-Sentinel-5P; TROPOMI; methane; nitrogen dioxide; machine learning; spatial cross-validation; emission hotspots; SHAP; Bangalore*

1. Introduction

Atmospheric greenhouse gas concentrations are now higher than at any point in at least the past 800,000 years. Carbon dioxide has risen from around 280 parts per million before industrialization to over 420 ppm today, while methane has more than doubled, from approximately 700 parts per billion to around 1,900 ppb (Intergovernmental Panel on Climate Change [IPCC], 2021). Methane matters disproportionately: over a twenty-year timeframe it traps roughly eighty times more heat per unit mass than carbon dioxide (IPCC, 2021). It is also short-lived in the atmosphere, so reducing its emissions is one of the few methods capable of slowing warming within a human generation.

Cities contribute greatly to this problem. Urban areas generate the majority of energy-related greenhouse gas emissions while occupying a small fraction of the land, and the fastest growth in emissions is occurring in rapidly urbanising regions of the Global South (IPCC, 2022). Bangalore, one of India's fastest-growing metropolitan areas, illustrates this pattern: sustained expansion of its vehicle fleet, industrial activity and waste volumes has occurred alongside limited local capacity for greenhouse gas measurement. India is among the world's largest emitters, yet dense, city-scale observational coverage of greenhouse gases remains scarce.

Four approaches are mainly used in greenhouse gas monitoring, each with its own strengths and limitations.

Ground-based monitoring stations provide highly accurate, continuous measurements at fixed points. However, they are expensive to install and maintain, and their spatial coverage is limited. A single station represents its immediate surroundings and cannot capture the variation in emissions across a metropolitan area.

Atmospheric sampling campaigns, including flask sampling and aircraft surveys, achieve high precision and can characterize specific plumes (NOAA Global Monitoring Laboratory, 2024). However, campaigns are costly and produce snapshots rather than continuous records.

Emission inventories such as bottom-up national and gridded inventories estimate emissions from activity data and emission factors (European Commission, 2025). They offer comprehensive detail but describe estimated emissions rather than observed atmospheric conditions and are produced at spatial resolutions too low for urban decision-making.

Satellite remote sensing addresses the problem of coverage. Instruments such as TROPOMI (Tropospheric Monitoring Instrument) aboard the European Space Agency's Sentinel-5P satellite, launched in 2017, measure gas columns with full daily coverage outside a narrow band around the equator (Copernicus, n.d.). The methane and carbon monoxide products central to this study have been provided at a resolution of 5.5 × 7 km^2 since August 2019, improved from 7 × 7 km^2 at launch (Sha et al., 2021). However, satellite archives are large, measurements represent total atmospheric columns rather than surface concentrations, and extracting meaning from them requires substantial processing expertise.

Across these approaches, four limitations are seen: limited spatial coverage, high operational cost, delayed reporting, and reliance on manual data processing. Decision-makers in many cities lack timely and spatial information about where greenhouse gases are elevated and why.

Machine learning can overcome some of these limitations. Since it can identify complex patterns in large amounts of data without needing a detailed physical model, it is useful for combining different types of environmental data. Machine learning has been applied to atmospheric environment problems, particularly the prediction of air pollutants such as particulate matter, ozone and nitrogen dioxide, where ensemble and neural approaches have been systematically compared (Zheng et al., 2021).

Two developments are particularly important. First, machine learning methods such as Random Forest and gradient boosting have been shown to work effectively with environmental data. They can often produce more accurate results than simpler linear methods while still requiring relatively little computing power. Second, explainable AI techniques, particularly SHAP (SHapley Additive exPlanations), make it possible to measure how much each input variable contributes to a model's prediction (Lundberg & Lee, 2017). This helps address a major concern about using machine learning in environmental governance: the lack of transparency in models whose decision-making processes can otherwise be difficult to understand.

Improved greenhouse gas monitoring has a direct environmental impact. Faster identification of emission hotspots allows resources to be directed to the locations where they will achieve the greatest reduction.

Other than identification, timely monitoring allows early intervention rather than retrospective accounting, supports targeted emission reduction, and provides the infrastructure on which carbon neutrality strategies depend on.

This research relates directly to several United Nations Sustainable Development Goals. It supports SDG 13 (Climate Action) by strengthening the evidence base for mitigation; SDG 11 (Sustainable Cities and Communities) by generating city-scale information for urban environmental planning; SDG 9 (Industry, Innovation and Infrastructure) through the application of computational methods in environmental infrastructure; and SDG 15 (Life on Land) as reduced emissions and improved air quality benefit ecosystems on land. All together, artificial intelligence-enabled monitoring has the potential to improve environmental mitigation and strengthen climate resilience by making emissions visible and manageable.

This study aims to develop and evaluate an explainable artificial intelligence framework for monitoring atmospheric greenhouse gases in an urban area. It also examines how that framework could support environmental management and climate change mitigation.

The specific objectives are:

1. To collect and combine environmental data from multiple sources, including Sentinel-5P satellite measurements of methane, carbon monoxide and nitrogen dioxide, together with ERA5 meteorological data for Bangalore from 2020 to 2024.
2. To develop a reproducible data-processing pipeline that aligns the datasets by location and time, creating a single dataset for analysis.
3. To compare the performance of three machine learning algorithms, Random Forest, XGBoost and LightGBM, in predicting greenhouse gas concentrations. The validation process will account for differences across both locations and time periods.
4. To identify areas with unusually high greenhouse gas concentrations using density-based spatial clustering and to analyse the characteristics of these hotspots statistically.
5. To examine which environmental factors influence the model's predictions using SHAP-based explainable artificial intelligence.
6. To use the identified hotspots to develop recommendations for reducing emissions from specific sources, and to compare these findings with independent land-use data.

Based on these objectives, the study addresses the following research question: How can AI-assisted monitoring improve the assessment of atmospheric greenhouse gases and support sustainable environmental management and climate change mitigation?

2. Methodology

2.1 Research design

This study used a quantitative design based on secondary data. It is non-experimental: no variable was manipulated and no intervention was applied. Observations published by other organisations were combined and processed, then analysed to test how accurately machine learning models could describe greenhouse gas concentrations over a city.

The design is comparative in two ways. Three regression algorithms were trained on identical observations, so any difference in their scores comes from the algorithms rather than the data that was used. Each model was then put through three cross-validation schemes, so the effect of the validation method itself could be measured.

2.2 Study area and spatial sampling

Bangalore is on the Deccan plateau in southern India at about 920 m above sea level, high enough that temperatures stay milder than in the surrounding lowlands (Indian Institute of Science, n.d.). The climate is tropical savanna, with a dry season from December to May and most of the year's rain falling between June and November.

It is also a suitable test case for the question being asked. An emissions inventory for the Greater Bengaluru airshed placed transport ahead of every other source, with vehicle exhaust and resuspended road dust together accounting for 56 percent of $PM_{2.5}$ and 70 percent of PM_{10} emissions, followed by industry at 17.8 percent and open waste burning at 11.0 percent (Guttikunda et al., 2019).

The study area is a rectangle covering Bangalore and the land around it, bounded by latitudes 12.75°N and 13.25°N and longitudes 77.35°E and 77.85°E. It has been divided into a regular grid at a spatial resolution of 0.05°, giving 100 grid cells of about 5.5 km per side. That resolution was chosen to match the satellite products, so the grid can use detail the instrument provides. The study period runs for sixty consecutive months, from January 2020 to December 2024.

All 100 cells inside the boundary were included, and cells were not classified beforehand. Every cell was processed the same way, whether it covers the city centre or farmland at the edge. Land-use information was brought in only after the analysis was finished, as an independent check on the results, and is described in Section 2.9. Temporal sampling was also complete, since every month in the period was used.

2.3 Selection of gases and definition of the variables

Three gases were analysed: methane (CH_4), carbon monoxide (CO) and nitrogen dioxide (NO_2). Methane is the greenhouse gas of main interest here. Carbon monoxide and nitrogen dioxide were included because both come from combustion, which makes them useful for working out the likely source of an emission.

The dependent variable is the monthly methane concentration in a grid cell, measured as the column-averaged dry-air mole fraction in parts per billion. It is a single continuous variable. Each model is therefore a single-output regressor: one model returns one number, the methane value for one cell in one month. No multi-output regressor was used, and the three gases were not predicted jointly.

Carbon monoxide and nitrogen dioxide are independent variables in the methane models rather than separate targets. They enter the feature matrix alongside the meteorological and temporal predictors, on the reasoning that a cell with raised combustion tracers may also have raised methane from the same local activity. The complete predictor set has 22 variables: the carbon monoxide and nitrogen dioxide columns, the seven ERA5 meteorological variables of Section 2.5, the three meteorological variables derived from them in Section 2.7, the six temporal variables described in the same section, latitude and longitude, and the year and calendar month. Every model was fitted on complete records only. A row therefore contributes to the methane models just when methane, both other gases and all meteorological fields are present in that cell and month.

Nitrogen dioxide was also modelled as a dependent variable in its own right, using the same three algorithms and the same validation schemes, with methane and carbon monoxide moved into the predictor set. The purpose was to test whether the validation behaviour seen for methane is a property of that gas or of the method. It is reported alongside the methane results. Carbon monoxide was not modelled as a target. In every case the model

is fitted to one gas at a time, so the validation comparison covers three algorithms applied to two target gases, or six fitted models in each validation scheme.

2.4 Satellite data collection

Satellite observations came from the Sentinel-5P Level 2 product collection, accessed through the Copernicus Data Space Ecosystem. Rather than downloading and processing individual orbit files, monthly composites were requested directly from the processing service (Copernicus Data Space Ecosystem, 2025).

Where a cell had no valid retrieval in a month, the value was recorded as missing rather than as zero. Methane retrieval requires clear skies, so cloudy months sometimes return nothing. Recording those as zero would have biased the data, creating a false pattern of very low methane during the monsoon.

Methane is reported as a column-averaged dry-air mole fraction in parts per billion. Carbon monoxide and nitrogen dioxide are reported as total and tropospheric column densities in mol m^{-2}, with nitrogen dioxide values typically around 10^{-5}. Since August 2019 the methane and carbon monoxide products have been produced at a resolution of 5.5 × 7 km (Sha et al., 2021). Nitrogen dioxide has been produced at 5.5 × 3.5 km since the same date (TROPOMI, n.d.).

2.5 Meteorological data

Meteorological variables came from ERA5, the fifth-generation atmospheric reanalysis produced by the European Centre for Medium-Range Weather Forecasts and distributed through the Copernicus Climate Data Store (Hersbach et al., 2020). ERA5 is a reanalysis rather than a set of direct measurements. It is produced by a numerical weather model into which large volumes of historical observations have been assimilated, which gives a record with no gaps in space or time.

Seven variables were downloaded as monthly means: 2 m air temperature, surface pressure, 2 m dewpoint temperature, the eastward and northward components of the 10 m wind field, boundary layer height, and total precipitation. Each affects what a satellite sees. Wind carries emissions away from their source. Boundary layer height sets the volume of air that surface emissions mix into, so a shallow layer concentrates them. Precipitation removes soluble gases through wet deposition.

2.6 Data integration and preprocessing

ERA5 has a spatial resolution of 0.25°, which gives only six cells across the study area against 100 satellite cells. Each satellite cell was therefore assigned to its nearest ERA5 cell by nearest-neighbour matching, using a k-d tree on the cell centre coordinates (Bentley, 1975). The largest distance between a satellite cell and its assigned meteorological cell was 0.18°, about 20 km. Because the ERA5 grid does not change over time, this assignment was computed once and reused for every month. Temporal matching was exact: records were joined on year and month, pairing each monthly satellite composite with the weather for the same place and period.

Records where all three gases were missing carry no information on the dependent variable and were removed by listwise deletion, taking out 381 rows. The final dataset has 5,619 records across 100 cells and 60 months. Coverage varies a great deal between gases: nitrogen dioxide appears in 5,069 records and carbon monoxide in 5,472, but methane in only 2,197. About 60 percent of methane observations are missing, almost all because of cloud cover.

2.7 Feature engineering

Several predictor variables were derived from the retrieved data. Wind speed was calculated from the two wind components as the square root of u^2 plus v^2, and wind direction from the same components using the standard meteorological convention, in degrees clockwise from north. Relative humidity was derived from 2 m temperature and 2 m dewpoint temperature through a Magnus-form approximation to saturation vapour pressure (Alduchov & Eskridge, 1996). Season was assigned from the calendar month using the Indian seasonal classification of winter, summer, monsoon and post-monsoon. Month and day-of-year were also transformed into sine and cosine pairs. This cyclical encoding stops the model treating December and January as far apart when they are in fact next to each other.

2.8 Model development, training and validation

Three supervised regression algorithms were trained: a random forest (Breiman, 2001) with 300 estimators, and two gradient boosting models, XGBoost (Chen & Guestrin, 2016) and LightGBM (Ke et al., 2017), each with 400 estimators and a learning rate of 0.05, with XGBoost also limited to a maximum tree depth of six. Hyperparameters were left at their default values rather than tuned. The aim was to see whether different

algorithms show similar performance on the same set of data, and tuning each one separately would have made that comparison harder to figure out.

Records with a missing target value or any missing predictor were dropped, so all three models trained on identical observations. The data was split into training and hold-out test sets in an 80:20 ratio using a fixed random seed, so the split can be reproduced exactly. Performance is reported as mean absolute error, root mean squared error and the coefficient of determination (R^2).

Three cross-validation schemes were applied. Shuffled folds are known to overstate performance on data with spatial or temporal structure, because records lying close together in space or time end up on both sides of the split and the model is scored on observations it has effectively already seen; blocking the folds by a grouping variable is the standard correction (Roberts et al., 2017). The first is five-fold cross-validation with the records shuffled. The second blocks the folds by year, withholding a complete year and asking the model to generalise to a period it has never seen. A third scheme blocks by grid cell, withholding complete locations. In the year-blocked scheme the year variable was removed from the predictors, and in the cell-blocked scheme latitude and longitude were removed, since keeping the grouping variable as a predictor would cause the model to fail.

2.9 Hotspot detection, model interpretation and source attribution

Emission hotspots were found in three stages. Concentrations were first aggregated over time, taking the mean across the study period to give one value per grid cell. Cells at or above the 75th percentile were marked as elevated. The coordinates of those cells were then clustered using HDBSCAN. A minimum cluster size of three was used, which suits a grid of 100 cells. Coordinates were standardised first so latitude and longitude contributed equally to the distance measure.

Nitrogen dioxide was used for hotspot detection rather than methane. Averaged over the study period, methane varies by less than one percent across the whole grid, which is too little spatial contrast for density-based clustering to work with. Nitrogen dioxide varies much more across the same cells.

Model behaviour was interpreted using SHAP (SHapley Additive exPlanations), computed with the exact tree estimator (Lundberg & Lee, 2017). The method breaks each individual prediction into additive contributions from each predictor variable, in the units of the target variable. For the methane models, for example, each feature's contribution is measured in parts per billion.

Likely sources were assigned to each cluster using a rule-based classification. The mean concentration of each gas inside a cluster was compared against the study-area baseline mean. A gas counted as elevated if it exceeded that baseline by at least three percent. Elevated nitrogen dioxide alone maps to traffic and vehicular emissions, elevated nitrogen dioxide together with carbon monoxide to industrial combustion, and elevated methane to waste management or agriculture.

For road mapping, 8,688 major road segments and 1,602 industrial land-use polygons for the study area were downloaded from OpenStreetMap through the Overpass API and plotted beneath the detected clusters (OpenStreetMap contributors, 2025). Those layers were brought in only to see whether the hotspots fit with the existing infrastructure.

2.10 Software

The analysis was written in Python 3.11. Satellite data were retrieved with the sentinelhub library and raster data handled with rasterio. Meteorological files were read with xarray and netCDF4, and nearest-neighbour matching used scipy (Virtanen et al., 2020). Models came from scikit-learn (Pedregosa et al., 2011), xgboost (Chen & Guestrin, 2016) and lightgbm (Ke et al., 2017), clustering from hdbscan, and interpretation from shap. Figures were produced with matplotlib.

3. Results

3.1 Temporal and spatial patterns of greenhouse gas concentrations

The merged dataset holds 5,619 cell-month records across 100 grid cells and 60 months. Coverage is uneven between the three gases. Carbon monoxide was retrieved in 5,472 records and nitrogen dioxide in 5,069, but methane in only 2,197, so 60.9 percent of the methane values are missing. The gaps are not spread evenly through the year. Methane was retrieved in just 27 of the 60 months, and in no month between June and September in any of the five years. Figure 1 shows the monthly city-mean series.

Over the records available, methane averaged 1,925.13 ppb, with a standard deviation of 24.06 ppb and individual values between 1,845.46 and 2,001.03 ppb. The spatial variation is small. Averaged over the study

period, the per-cell means run from 1,917.75 ppb to 1,934.27 ppb, a spread of 16.53 ppb, or 0.86 percent of the mean. The difference between the highest and lowest cell is close to the size of the measurement scatter. Temporal variation is roughly four times larger. Averaged by calendar month, methane peaks in November at 1,956.9 ppb and falls to 1,894.5 ppb in May, a swing of about 62 ppb, and the post-monsoon mean of 1,956.6 ppb stands well above the summer mean of 1,913.0 ppb.

Annual means rose from 1,888.9 ppb in 2020 to 1,931.3 ppb in 2024, but those figures are not directly comparable, because the months that produced usable retrievals differ from year to year and no calendar month is present in all five. Comparing a single month across years removes that problem. February was retrieved in all 100 cells in each of 2021, 2022, 2023 and 2024, and the February mean rose from 1,896.96 ppb to 1,930.03 ppb over those four years, an increase of 33.1 ppb, or roughly 11 ppb per year. March shows the same direction, from 1,878.28 ppb in 2020 to 1,932.59 ppb in 2024. The rise is therefore present in the data rather than an artefact of which months were observed, and its rate is consistent with the continued global increase reported by the World Meteorological Organization (WMO, 2024).

Carbon monoxide averaged 0.0323 mol m^{-2}, ranging from 0.0193 to 0.0515 mol m^{-2}. Its spatial spread, at 7.19 percent of the mean, is larger than methane's but still modest. The seasonal cycle is far more pronounced: the April mean of 0.0409 mol m^{-2} is close to twice the July mean of 0.0237 mol m^{-2}, and the summer mean of 0.0375 mol m^{-2} sits well above the monsoon mean of 0.0249 mol m^{-2}.

Nitrogen dioxide behaves differently from both. It averaged 3.931×10^{-5} mol m^{-2} with a standard deviation of 1.873×10^{-5}, and its spatial variation is the largest of the three by a wide margin. Per-cell means range from 2.646×10^{-5} to 7.194×10^{-5} mol m^{-2}, a spread of 115.7 percent of the mean. The highest cell lies at 12.975°N, 77.525°E, in the centre of the city. Its seasonal cycle follows carbon monoxide, peaking in February at 5.292×10^{-5} and reaching a minimum in July at 2.400×10^{-5} mol m^{-2}. Annual means, by contrast, are flat, moving only between 3.885×10^{-5} and 3.963×10^{-5} mol m^{-2} across the whole period. Ten of the 5,069 nitrogen dioxide records are slightly negative. That reflects retrieval noise around very low column values rather than a physical measurement, and those records were left in place rather than filtered.

Methane varies in time but barely in space, which is what a well-mixed gas with a long atmospheric lifetime should do. Nitrogen dioxide varies strongly in space, which is what a short-lived gas emitted from identifiable local sources should do. Carbon monoxide falls between them. The monsoon suppresses both combustion tracers through ventilation and wet deposition, and removes the methane observations altogether.

3.2 Predictive performance of the machine learning models

After records with a missing target or any missing predictor were removed, 2,137 complete observations remained, split into 1,709 for training and 428 for the hold-out test set. Each model used the 22 predictors listed in Section 2.3.

With methane as the dependent variable, all three algorithms performed almost identically, as shown in Table 1. XGBoost recorded the highest coefficient of determination at 0.956, with a mean absolute error of 3.57 ppb and a root mean squared error of 4.91 ppb. Random Forest had the lowest mean absolute error at 3.55 ppb and the highest root mean squared error at 4.95 ppb, and LightGBM sat between the two. Across the three, R^2 differs by 0.0007. Mean absolute percentage error is below 0.19 percent for all of them, but that figure flatters the model, because the test-set methane mean is 1,925.4 ppb and any error expressed as a percentage of a number that large will be small. A more informative comparison is the root mean squared error of 4.91 ppb set against the test-set standard deviation of 23.53 ppb, about 21 percent.

That the three algorithms agree so closely is itself a result. Random Forest and gradient boosting build their trees in different ways, so if the choice of algorithm were driving the score they would not land within a thousandth of each other. The accuracy reported here is a property of the data and the way it was split, not of any one model. Ren et al. (2020), comparing thirteen algorithms for ozone across the United States, reached the same conclusion about the weight of the data-splitting design relative to the choice of algorithm.

Nitrogen dioxide was modelled separately as its own dependent variable, and the three algorithms did not agree there. LightGBM reached an R^2 of 0.893 and Random Forest 0.860, but XGBoost returned −0.001 with a mean absolute percentage error of 32.9 percent, meaning it performed no better than predicting the mean of the data. The likely reason is the scale of the target. Nitrogen dioxide columns are of order 10^{-5} mol m^{-2}, and the default settings described in Section 2.8 left XGBoost unable to fit a target this small. Tuning would probably correct it, but the models were deliberately left untuned so that the comparison used identical settings.

Table 1. Performance of the models on the hold-out test set (n = 428).

Target	Model	R^2	MAE	RMSE	MAPE (%)
CH_4 (ppb)	XGBoost	0.9563	3.572	4.913	0.186
CH_4 (ppb)	LightGBM	0.9561	3.622	4.925	0.188
CH_4 (ppb)	Random Forest	0.9556	3.551	4.949	0.184
NO_2 (mol m^{-2})	LightGBM	0.8929	3.82×10^{-6}	5.58×10^{-6}	9.07
NO_2 (mol m^{-2})	Random Forest	0.8597	4.43×10^{-6}	6.39×10^{-6}	10.25
NO_2 (mol m^{-2})	XGBoost	−0.0006	1.36×10^{-5}	1.71×10^{-5}	32.85

3.3 Generalisation under blocked validation

The scores in Table 1 come from a randomly shuffled split, and Section 2.8 gave the reason for not stopping there (Roberts et al., 2017). Table 2 reports the same models under the three cross-validation schemes.

Under shuffled five-fold cross-validation the methane models reproduce the hold-out result, with R^2 between 0.9521 and 0.9529 and a fold-to-fold standard deviation below 0.007. Blocking by grid cell barely changes anything: R^2 falls to between 0.9358 and 0.9376 and the root mean squared error rises from about 5.3 ppb to about 6.1 ppb. Asked to estimate methane over a location it has never been trained on, the model loses almost nothing.

Blocking by year produces the opposite outcome. Every methane model returns a negative R^2, between −1.43 and −1.75, with a fold-to-fold standard deviation above 1.5 and a root mean squared error between 24.1 and 25.9 ppb, roughly five times the shuffled figure. A negative R^2 means the model did worse than simply predicting the training mean. Withhold a year, and the model cannot estimate that year's methane at all.

The two results follow from the patterns in Section 3.1. Methane's variance over this study area is almost entirely temporal, so an unobserved cell can be filled in from its neighbours with very little loss, while removing the year removes most of the information the model was using. Two further factors deepen the year-blocked collapse. The concentration is rising, so a withheld year lies outside the range the model was trained on, and tree-based models cannot extrapolate beyond the values they have seen. Because the months with usable methane retrievals differ from year to year, withholding a year also changes which seasons appear in the test fold.

Nitrogen dioxide behaves more evenly. Random Forest falls from 0.8698 under shuffled folds to 0.6466 blocked by year and 0.6492 blocked by cell, and LightGBM from 0.9049 to 0.6656 and 0.6345. Both lose roughly a quarter of their explanatory power under blocking and both stay clearly positive. XGBoost remains near zero under every scheme, consistent with its failure in Table 1.

The honest reading is that the headline R^2 of 0.956 describes interpolation within observed conditions and should not be presented as forecasting skill. These methane models fill spatial gaps in a partially observed field, at an error of about 6 ppb. Forecasting is beyond them. Five years of data over a single city is not enough to predict a year the model has never seen.

Table 2. Mean R^2 across five folds under three cross-validation schemes.

Target	Model	Shuffled	Blocked by year	Blocked by cell
CH_4	Random Forest	0.9521	−1.5650	0.9376
CH_4	XGBoost	0.9529	−1.7498	0.9362
CH_4	LightGBM	0.9527	−1.4300	0.9358
NO_2	Random Forest	0.8698	0.6466	0.6492
NO_2	XGBoost	−0.0041	−0.0099	−0.0452

NO_2	LightGBM	0.9049	0.6656	0.6345

3.4 Detection of emission hotspots

Hotspot detection used nitrogen dioxide, for the reason given in Section 2.9. Averaging each cell over the five years and taking the 75th percentile gave a threshold of 4.442×10^{-5} mol m^{-2}, which marked 25 of the 100 cells as elevated. HDBSCAN grouped 16 of those 25 into two clusters and labelled the remaining nine as noise, meaning they were elevated but too isolated to form a dense group.

Cluster 0 is the larger of the two. It covers 13 cells centred on 12.948°N, 77.544°E, with a convex hull of 0.0213 square degrees, and its nitrogen dioxide values run from 4.551×10^{-5} to 7.194×10^{-5} mol m^{-2} around a mean of 5.877×10^{-5}. Cluster 1 is compact, three cells centred on 13.025°N, 77.575°E with a mean of 5.952×10^{-5} mol m^{-2}. Both clusters sit about 50 percent above the study-area baseline of 3.931×10^{-5} mol m^{-2}, and both fall over the central and western parts of the city rather than its edges. Figure 2 shows the concentration surface with both clusters outlined.

The rule-based attribution of Section 2.9 assigned both clusters to traffic and vehicular emissions, which is the largest emission source in the Greater Bengaluru airshed inventory (Guttikunda et al., 2019). Nitrogen dioxide exceeded the baseline by well over the three percent margin in each, while carbon monoxide did not, and it is the combination of the two that the rules treat as a signature of industrial combustion. Plotting the clusters over the 8,688 road segments and 1,602 industrial land-use polygons retrieved from OpenStreetMap supports that assignment, and Figure 3 shows the result. Both clusters sit over the centre of the city, where the road network is at its densest. Industrial land use is a weaker match. A few small parcels do fall inside cluster 0, but the largest industrial concentrations lie outside both clusters, mainly in the north-east and towards the southern edge of the study area. The land-use data took no part in the clustering, so this is an independent check rather than a circular one.

No methane hotspots were found, and none could have been. With a spatial spread of 0.86 percent across the grid there is no density structure for a clustering algorithm to detect. That is a property of the gas rather than a failure of the method.

3.5 Environmental drivers of the predictions

SHAP values were computed for the best methane model on the hold-out test set, and Table 3 gives the ten largest contributors expressed both as a mean absolute contribution in parts per billion and as a share of the total.

Season is the single largest driver, moving each prediction by 11.61 ppb on average and accounting for 33.9 percent of the total attributed importance. Year is second at 7.11 ppb, or 20.7 percent. Between them the two temporal variables account for 54.6 percent of everything the model uses. Meteorological variables follow at some distance: the eastward wind component contributes 3.33 ppb, the northward component 1.58 ppb, and relative humidity 1.44 ppb, together a little under 19 percent. Least of the named predictors are the two co-pollutant columns, nitrogen dioxide at 0.99 ppb and carbon monoxide at 0.76 ppb, about 5 percent combined.

The direction of each relationship was checked by correlating each feature's value against its own SHAP contribution. Year is strongly positive, which simply records the rising trend. Relative humidity and surface pressure are positive, boundary layer height is negative, and both wind components are negative, all of which is physically sensible: a deeper mixing layer and stronger winds dilute and disperse what is below them, while a shallow humid stagnant atmosphere holds it in place. Latitude, longitude and carbon monoxide show no consistent direction, which fits a well-mixed gas with no spatial gradient to track. Boundary layer height and the wind components have been reported as leading meteorological controls in comparable urban nitrogen dioxide models built from the same two data sources (Kukavica et al., 2026).

A model that takes more than half its information from the date is one that will collapse when the date is withheld, which is the result reported in Section 3.3. The ranking also shows the limit of what the co-pollutants add. Nitrogen dioxide and carbon monoxide were included on the reasoning that shared combustion sources might raise methane alongside them, and at roughly 5 percent of attributed importance that reasoning is only weakly supported over this area.

Table 3. Ten most important predictors of methane concentration (SHAP, XGBoost).

Predictor	**Mean |SHAP| (ppb)**	**Relative importance** (%)	**Relationship**
Season	11.61	33.85	Positive
Year	7.11	20.73	Positive
Eastward wind (u)	3.33	9.71	Negative
Northward wind (v)	1.58	4.61	Negative
Relative humidity	1.44	4.20	Positive
NO_2 column	0.99	2.89	Positive
Longitude	0.89	2.60	Nonlinear
Precipitation	0.86	2.51	Nonlinear
Surface pressure	0.85	2.47	Positive
CO column	0.76	2.21	Nonlinear

3.6 Detection of elevated-concentration months

Predicting a concentration and flagging an elevated one are different tasks, and the second is closer to what a monitoring authority needs. To test it, a threshold was set at the 75th percentile of the training target alone, so that no information from the test set entered the threshold. For methane this gave 1,938.57 ppb and for nitrogen dioxide 5.466×10^{-5} mol m^{-2}. Each model's predictions were then classified against that same threshold, and the resulting labels compared with the labels implied by the observed concentrations.

Of the 428 methane records in the test set, 102 were genuinely above the threshold. XGBoost flagged 99 of them, of which 90 were correct, giving a detection rate of 88.2 percent, a precision of 90.9 percent, a false positive rate of 2.8 percent and an overall accuracy of 95.1 percent. Random Forest and LightGBM both detected 87.3 percent at slightly lower precision. For nitrogen dioxide, 118 of the 428 records were elevated, and LightGBM detected 87.3 percent of them with 93.6 percent precision and a false positive rate of 2.3 percent. Random Forest detected 78.0 percent. XGBoost detected none at all, which follows directly from its near-zero R^2: a model predicting close to a constant never crosses the threshold.

The rates in Table 4 carry the same caveat as Table 1, having been computed on the shuffled hold-out split. Read with that restriction, they suggest that the useful part of the output may be the ranking rather than the exact value: a model that cannot predict the 2024 methane level may still identify which cells and months in 2024 are the high ones.

Table 4. Classification of elevated cell-months in the hold-out test set (n = 428).

Target	**Model**	**Detection rate** (%)	**Precision** (%)	**False positives** (%)	**Accuracy** (%)
CH_4	XGBoost	88.2	90.9	2.8	95.1
CH_4	Random Forest	87.3	89.0	3.4	94.4
CH_4	LightGBM	87.3	88.1	3.7	94.2
NO_2	LightGBM	87.3	93.6	2.3	94.9
NO_2	Random Forest	78.0	90.2	3.2	91.6
NO_2	XGBoost	0.0	not applicable	0.0	72.4

3.7 Relevance to emission reduction

Nothing in this study reduces an emission. It produces information of a kind that a reduction effort would need, along a path that runs from monitoring to pattern detection, then to hotspot identification and targeted intervention.

The narrowing that happens along that path is worth stating in numbers. A quarter of the study area, 25 of 100 cells, sits above the 75th percentile for nitrogen dioxide. Clustering reduced that to 16 cells in two groups, and the attribution step assigned both groups to traffic rather than to industry or waste, which selects a different set of measures. The recommendations attached to those clusters concern public transport capacity, electric vehicle uptake and low-emission zones. An authority working from the raw percentile map would be looking at a quarter of the city; working from the clusters it would be looking at two defined areas with a stated likely cause.

No emission-reduction scenario was modelled, and no figure in tonnes of carbon dioxide equivalent can be offered. Estimating that would require activity data, emission factors and an assumption about compliance, none of which this study holds. The claim is narrower: the method finds where concentrations are elevated and ranks which of those areas should be dealt with first.

3.8 Implications for environmental management

The practical case for the approach rests on coverage rather than accuracy. A value was produced for every one of 100 cells in every month that returned a retrieval, across five years and for three gases, using instruments already in orbit and a reanalysis product already published. No instrument was installed and no field campaign was run.

Three limits apply to everything above. The satellite measures a total or tropospheric column rather than the concentration at street level, so a column hotspot shows where the overlying air is loaded and not where a person is breathing. Methane is missing for every monsoon month, which is the season when boundary layer and rainfall conditions differ most from the rest of the year. The methane results therefore describe a dry-season city. And no ground measurement from Bangalore entered the analysis, so the validation in Section 3.3 tests whether the models generalise, not whether the underlying observations are accurate over this city.

3.9 Overall results in relation to the research question

Table 5 collects the main quantities and the condition attached to each. They give a split answer to the question the study set out to address. AI-assisted monitoring improved the assessment of greenhouse gases over Bangalore on three counts. Spatial gaps in a partially observed methane field were filled at an error of about 6 ppb. Two nitrogen dioxide hotspots were located and characterised, with independent land-use data supporting the attribution. And elevated cell-months were identified at a detection rate near 88 percent. It did not deliver a forecast, and the negative year-blocked R^2 is the evidence against any claim that these models can predict a year they have not seen.

Table 5. Summary of results and the conditions attached to each.

Indicator	Result	Condition
Prediction R^2, CH_4	0.956	Shuffled split; interpolation within observed conditions
R^2, unobserved location	0.936	Blocked by grid cell; spatial gap-filling holds
R^2, unobserved year	−1.57	Blocked by year; worse than predicting the mean
MAE / RMSE, CH_4	3.57 / 4.91 ppb	Against a mean of 1,925 ppb
MAPE, CH_4	0.186%	Small by construction on a large-valued target
Agreement across algorithms	R^2 within 0.0007	Accuracy is a property of the data, not the model
Hotspot clusters detected	2, covering 16 cells	NO_2 only; CH_4 has no usable spatial contrast

Elevated months detected	88.2% at 2.8% false positives	Same shuffled split, same optimism
Spatial coverage	100 cells, 60 months	No instrument installed
CH_4 data completeness	39.1%	No retrievals at all from June to September
Dominant driver	Season and year, 54.6%	Explains the year-blocked collapse
Emission reduction estimated	None	Requires activity data not held here

4. Discussion

4.1 Comparison with other studies

Three recent studies make useful comparators, because each combines satellite retrievals with reanalysis meteorology and tree-based regression in roughly the way done here.

Kukavica et al. (2026) estimated ground-level nitrogen dioxide over Zagreb from TROPOMI columns, Sentinel-2 indices and ERA5 variables, using Random Forest and XGBoost. Their final models reached an R^2 of 0.74 and 0.78 respectively, with test-set root mean squared errors of 9.86 and 8.76 μg m^{-3}. Two things in that study line up with what is reported here. Their training error was roughly half their test error, 4.35 against 9.86 μg m^{-3} for Random Forest, which is the same direction of movement seen in Section 3.3 when validation is made stricter. And their SHAP ranking placed boundary layer height fourth for both models and the northward wind component third for XGBoost, close to the ordering of meteorological drivers in Table 3, where the two wind components and relative humidity lead the weather variables and boundary layer height follows.

The nitrogen dioxide scores obtained here under blocked validation, between 0.63 and 0.67, sit a little below the Zagreb figures. That is the expected direction. Kukavica et al. trained against ground-station measurements with station identity available to the model, whereas the nitrogen dioxide models here were asked to generalise to withheld locations and withheld years.

Cedeno Jimenez and Brovelli (2024) estimated ground-level nitrogen dioxide over Mexico City from Sentinel-5P and ERA5 alone, without ground monitors in the feature set, and reported temperature and wind speed as the leading meteorological controls. The present study finds the same family of variables carrying the weather signal for methane, though wind enters through its components rather than its magnitude: the eastward and northward components contribute 3.33 and 1.58 ppb against 0.42 ppb for wind speed itself.

Ren et al. (2020) compared thirteen algorithms for ozone across the United States and concluded that careful design of data splitting and validation is necessary for reliable and stable results, with Random Forest and gradient boosting among the best performers. Both findings are reproduced here. The three algorithms land within 0.0007 in R^2 of one another, and the choice of validation scheme changes the headline figure far more than the choice of algorithm does.

4.2 What the validation result means

The gap between an R^2 of 0.956 under shuffled folds and −1.57 under year-blocked folds is the most consequential number in this study, and it is consistent with the warning in Roberts et al. (2017) that ignoring dependence structure in cross-validation produces serious underestimation of predictive error. What this study adds is the size of the effect for a well-mixed gas over a single city, which is large enough to invert the sign of the score.

The asymmetry between the two blocking schemes is the informative part. Blocking by location costs almost nothing, 0.956 to 0.936, while blocking by year costs everything. Both numbers come from the same data and the same models, so the difference is a property of methane over this area rather than of the method. Section 3.1 gives the reason: the per-cell spread is 0.86 percent while the seasonal swing is about 62 ppb. A model fitted to a field that is nearly flat in space and strongly structured in time will interpolate across space and fail across time, and that is what happened.

This has a practical consequence for how such models should be deployed. Gap-filling a partially observed field is a legitimate application of the methane models reported here, at an error of about 6 ppb. Forecasting is not. Published studies that report a single shuffled-split score for satellite-derived greenhouse gas prediction are not necessarily reporting forecasting skill, and the distinction matters when the output is intended to support policy.

4.3 Why XGBoost failed on nitrogen dioxide

XGBoost returned an R^2 of −0.001 for nitrogen dioxide while Random Forest and LightGBM reached 0.86 and 0.89 on identical data. The target is of order 10^{-5} mol m^{-2}, and the model was run with the default settings used for every algorithm in the comparison. The failure is therefore a property of the default configuration meeting a very small-valued target, not evidence about gradient boosting as a method, and the contrast with LightGBM, which is also a gradient boosting implementation, makes that clear. It is reported rather than removed because leaving it in shows what an untuned comparison actually produces.

4.4 Interpretation of the hotspot result

Both detected clusters were attributed to traffic, and both sit over the densest part of the road network. That is consistent with the emissions inventory for the Greater Bengaluru airshed, which places transport ahead of every other source and attributes 56 percent of $PM_{2.5}$ emissions to vehicle exhaust and resuspended road dust (Guttikunda et al., 2019). The agreement is worth stating carefully. The inventory is a bottom-up estimate built from activity data, while the clusters here come from measured columns, so the two are independent routes to the same conclusion rather than one confirming the other.

The limits of the attribution should also be stated. The rule set distinguishes traffic from industry by whether carbon monoxide is elevated alongside nitrogen dioxide, which is a coarse test. A few small industrial parcels fall inside cluster 0, and the method as implemented would not separate an industrial source sitting inside a dense road network from the traffic around it.

5. Conclusion

This study set out to develop and evaluate an explainable artificial intelligence framework for monitoring atmospheric greenhouse gases over an urban area, and to assess what such a framework contributes to environmental management. Five years of Sentinel-5P observations of methane, carbon monoxide and nitrogen dioxide over Bangalore were combined with ERA5 reanalysis into a single gridded dataset. Three regression algorithms were trained on it, tested under three validation schemes, interpreted with SHAP, and used to locate and attribute emission hotspots.

The three gases behaved differently in ways that shaped everything after. Methane varied in time but almost not at all in space, with a per-cell spread of 0.86 percent across the grid. Nitrogen dioxide varied by 115.7 percent across those same cells. That difference decided which gas could support hotspot detection and which could not.

Model accuracy was high and needs qualifying. Random Forest, XGBoost and LightGBM agreed on methane to within 0.0007 in R^2, reaching 0.956 on a shuffled hold-out split. Because three differently built algorithms land that close together, the figure describes the data rather than any one model. Blocked by grid cell it held at 0.936, so the models fill spatial gaps in a partially observed field with little loss. Blocked by year, every methane model returned a negative R^2 and so performed worse than predicting the training mean. SHAP shows why. Season and year together carry 54.6 percent of the attributed importance, and withholding the year removes most of what the model was using.

Detection worked better than prediction. Two nitrogen dioxide hotspots covering 16 of 100 cells were identified, both about 50 percent above the study-area baseline and both attributed to traffic by the rule-based classifier. Plotted against OpenStreetMap data that took no part in the clustering, both sit over the densest part of the road network. Elevated cell-months were flagged at 88.2 percent with a false positive rate of 2.8 percent.

On the research question, AI-assisted monitoring improved the assessment of greenhouse gases over this city in three respects: it filled spatial gaps, it localised hotspots with an independent check behind the attribution, and it ranked which cells and months were elevated. It did not produce a forecast. On five years of data over one city it is not in a position to, and the negative year-blocked score is the evidence against any claim that it could.

Three constraints bound all of this. The satellite measures a column of air rather than what is at street level. Methane is missing for 60.9 percent of cell-months and for every monsoon month, so the methane findings describe a dry-season city. And no ground measurement from Bangalore entered the analysis at any stage, so nothing here establishes whether the satellite values are accurate over this particular place.

Each constraint points at the work that would follow. A longer record would allow the year-blocked test to run across more folds and would show whether temporal generalisation improves once more years are available. A comparison against a local ground station, were one measuring these gases to be installed, would settle whether

the column values track surface conditions. Extending the grid to other Indian cities would test whether the split seen here between a well-mixed gas and a short-lived one is general or particular to Bangalore.

References

1. Alduchov, O. A., & Eskridge, R. E. (1996). Improved Magnus form approximation of saturation vapor pressure. Journal of Applied Meteorology, 35(4), 601–609. https://doi.org/10.1175/1520-0450(1996)035<0601:IMFAOS>2.0.CO;2
2. Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. Communications of the ACM, 18(9), 509–517. https://doi.org/10.1145/361002.361007
3. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
4. Cedeno Jimenez, J. R., & Brovelli, M. A. (2024). Estimating ground-level NO2 concentrations using machine learning exclusively with remote sensing and ERA5 data: The Mexico City case study. Remote Sensing, 16(17), 3320. https://doi.org/10.3390/rs16173320
5. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785
6. Copernicus. (n.d.). Sentinel-5P mission. SentiWiki. Retrieved October 8, 2026, from https://sentiwiki.copernicus.eu/web/s5p-mission
7. Copernicus Data Space Ecosystem. (2025). Sentinel Hub Process API. European Space Agency. Retrieved October 5, 2026, from https://documentation.dataspace.copernicus.eu/APIs/SentinelHub/Process.html
8. European Commission, Joint Research Centre. (2025). EDGAR: Emissions Database for Global Atmospheric Research. Retrieved August 29, 2026, from https://edgar.jrc.ec.europa.eu/
9. Guttikunda, S. K., Nishadh, K. A., Gota, S., Singh, P., Chanda, A., Jawahar, P., & Asundi, J. (2019). Air quality, emissions, and source contributions analysis for the Greater Bengaluru region of India. Atmospheric Pollution Research, 10(3), 941–953. https://doi.org/10.1016/j.apr.2019.01.002
10. Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., … Thépaut, J.-N. (2020). The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730), 1999–2049. https://doi.org/10.1002/qj.3803
11. Indian Institute of Science, Centre for Ecological Sciences. (n.d.). Geography and environment. Energy and Wetlands Research Group. Retrieved October 5, 2026, from https://wgbis.ces.iisc.ac.in/energy/water/paper/bangalore/geography.htm
12. Intergovernmental Panel on Climate Change. (2021). Climate change 2021: The physical science basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press. https://doi.org/10.1017/9781009157896
13. Intergovernmental Panel on Climate Change. (2022). Climate change 2022: Mitigation of climate change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press. https://doi.org/10.1017/9781009157926
14. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.
15. Kukavica, M., Gašparović, M., & Henebry, G. M. (2026). Developing a machine learning approach for estimating urban NO2 concentrations using multi-source remote sensing data. Smart Cities, 9(10), 166. https://doi.org/10.3390/smartcities9100166
16. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://arxiv.org/abs/1705.07874
17. NOAA Global Monitoring Laboratory. (2024). Global greenhouse gas reference network. National Oceanic and Atmospheric Administration. Retrieved August 31, 2026, from https://gml.noaa.gov/ccgg/about.html
18. OpenStreetMap contributors. (2025). OpenStreetMap. OpenStreetMap Foundation. Retrieved August 31, 2026, from https://www.openstreetmap.org/
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
20. Ren, X., Mi, Z., & Georgopoulos, P. G. (2020). Comparison of machine learning and land use regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States. Environment International, 142, 105827. https://doi.org/10.1016/j.envint.2020.105827
21. Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913–929. https://doi.org/10.1111/ecog.02881
22. Sha, M. K., Langerock, B., Blavier, J.-F. L., Blumenstock, T., Borsdorff, T., Buschmann, M., Dehn, A., De Mazière, M., Deutscher, N. M., Feist, D. G., García, O. E., Griffith, D. W. T., Grutter, M., Hannigan, J. W., Hase, F., Heikkinen, P., Hermans, C., Iraci, L. T., Jeseck, P., … Zhou, M. (2021). Validation of methane and carbon monoxide from Sentinel-5 Precursor using TCCON and NDACC-IRWG stations. Atmospheric Measurement Techniques, 14(9), 6249–6304. https://doi.org/10.5194/amt-14-6249-2021
23. TROPOMI. (n.d.). Nitrogen dioxide. Netherlands Institute for Space Research. Retrieved October 8, 2026, from https://www.tropomi.eu/data-products/nitrogen-dioxide
24. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Vázquez-Baeza, Y. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods, 17(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

25. World Meteorological Organization. (2024). WMO greenhouse gas bulletin No. 20. Retrieved October 8, 2026, from https://wmo.int/resources/publication-series/greenhouse-gas-bulletin/wmo-greenhouse-gas-bulletin-no-20
26. Zheng, L., Lin, R., Wang, X., & Chen, W. (2021). The development and application of machine learning in atmospheric environment studies. Remote Sensing, 13(23), 4839. https://doi.org/10.3390/rs13234839

Figures

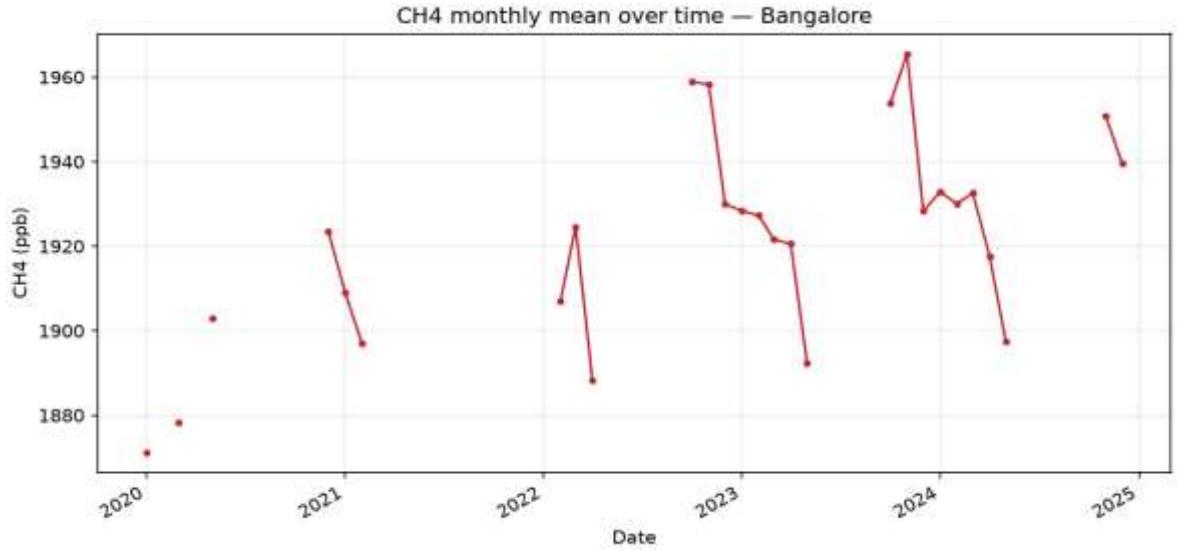
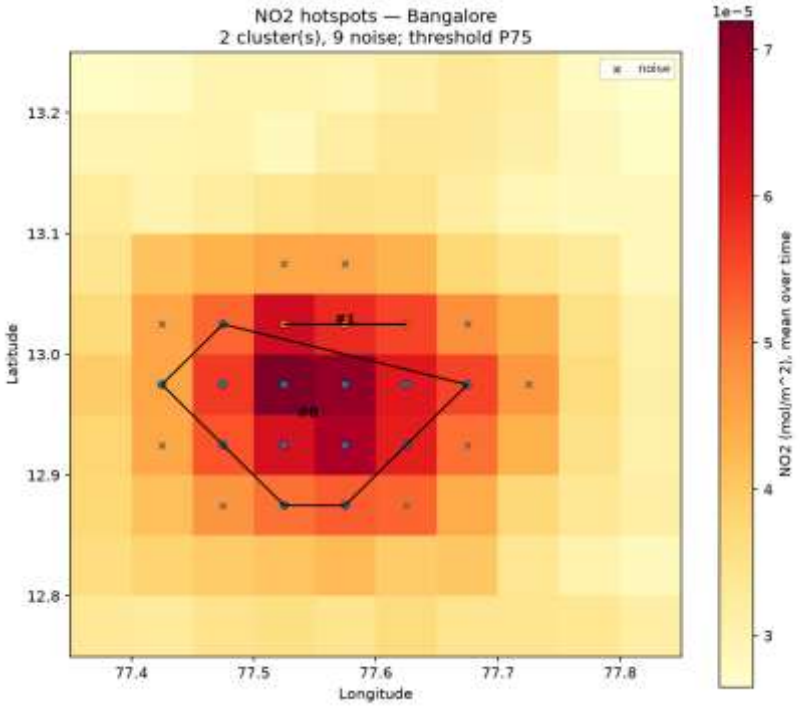


Figure 1. Monthly city-mean methane, 2020–2024. Breaks in the line are months with no usable retrieval.



27.

Figure 2. Mean nitrogen dioxide column over 2020–2024, one value per grid cell, with the two detected clusters outlined. Filled circles mark cells at or above the 75th percentile. Crosses mark elevated cells that HDBSCAN labelled as noise.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 12 No. 2, 2026
https://www.theaspd.com/ijes.php

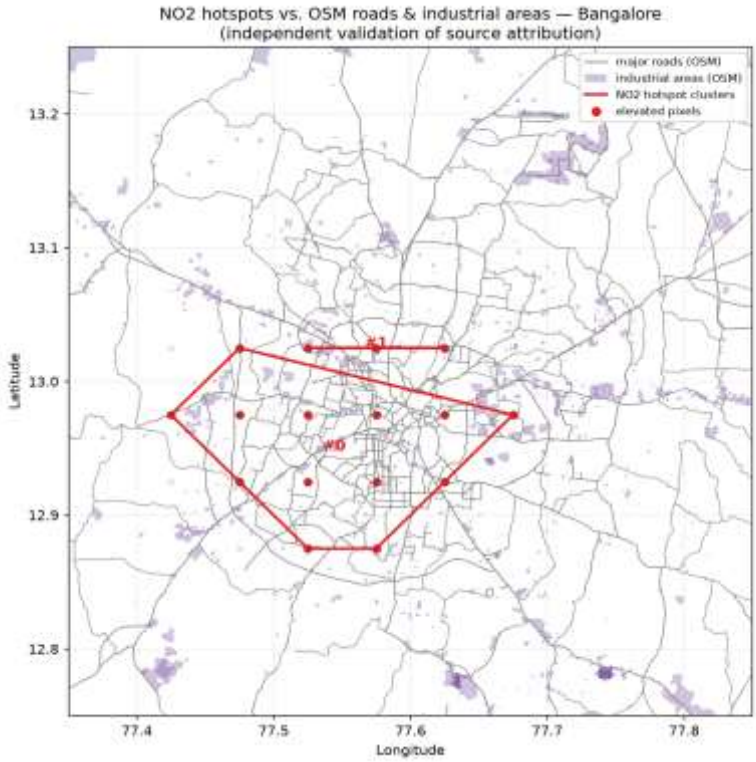


Figure 3. The same two clusters plotted over OpenStreetMap major roads and industrial land-use polygons. Neither layer was used in the clustering.