Environmental Data Science (2025), 4: e17, 1–34
doi:10.1017/eds.2025.9

APPLICATION PAPER

A framework for scalable ambient air pollution concentration estimation

Liam J. Berrisford[1,2,3], Lucy S. Neal[4], Helen J. Buttery[4], Benjamin R. Evans[4] and Ronaldo Menezes[1,5]

[1]BioComplex Laboratory, Department of Computer Science, University of Exeter, Exeter, UK
[2]Department of Mathematics, University of Exeter, Exeter, UK
[3]UKRI Centre for Doctoral Training in Environmental Intelligence, University of Exeter, Exeter, UK
[4]Met Office, Exeter, UK
[5]Department of Computer Science, Federal University of Ceará, Fortaleza, Brazil
Corresponding author: Liam J. Berrisford; Email: liam.j.berrisford@bath.edu

Received: 22 January 2024; **Revised:** 24 September 2024; **Accepted:** 03 February 2025

Keywords: air quality; data science; machine learning; sustainable development; urban resilience and justice

Abstract

Ambient air pollution remains a global challenge, with adverse impacts on health and the environment. Addressing air pollution requires reliable data on pollutant concentrations, which form the foundation for interventions aimed at improving air quality. However, in many regions, including the United Kingdom, air pollution monitoring networks are characterized by spatial sparsity, heterogeneous placement, and frequent temporal data gaps, often due to issues such as power outages. We introduce a scalable data-driven supervised machine learning model framework designed to address temporal and spatial data gaps by filling missing measurements within the United Kingdom. The machine learning framework used is LightGBM, a gradient boosting algorithm based on decision trees, for efficient and scalable modeling. This approach provides a comprehensive dataset for England throughout 2018 at a 1 km^2 hourly resolution. Leveraging machine learning techniques and real-world data from the sparsely distributed monitoring stations, we generate 355,827 synthetic monitoring stations across the study area. Validation was conducted to assess the model's performance in forecasting, estimating missing locations, and capturing peak concentrations. The resulting dataset is of particular interest to a diverse range of stakeholders engaged in downstream assessments supported by outdoor air pollution concentration data for nitrogen dioxide (NO_2), Ozone (O_3), particulate matter with a diameter of 10 μm or less (PM_{10}), particulate matter with a diameter of 2.5 μm or less $PM_{2.5}$, and sulphur dioxide (SO_2), at a higher resolution than was previously possible.

Impact Statement

The current high-quality air pollution monitoring station network in the United Kingdom is spatially sparse with heterogeneous placement and commonly suffers from missing data temporally from issues such as power outages. We present a scalable data-driven supervised machine learning model framework to fill missing measurements temporally and spatially, providing a complete dataset for England during 2018 at a 1 km^2 hourly resolution. The approach leverages machine learning and data from the sparse real-world monitoring stations to create 355,827 synthetic monitoring stations across the study. Validation was conducted regarding the model's performance in forecasting, estimating missing locations, and capturing peak concentrations. The dataset provided empowers stakeholders conducting downstream assessments underpinned by outdoor air pollution concentration data for various pollutants, enabling studies to be performed at a higher resolution than previously possible. Furthermore, this work demonstrates that similar approaches can be applied in other countries, as air pollution is a global issue, and many regions face similar challenges of limited data availability.

1. Introduction

Air pollution presents a significant health risk, with between 28,000 and 36,000 deaths per year in the UK associated with exposure (Office for Health Improvement and Disparities, 2022). Estimating ambient air pollution concentrations is crucial in addressing this health burden, although the high cost of individual monitoring stations remains a major challenge. The potential cost for a single multi-pollutant monitoring station could be as high as £198,000 (AEA Technology, 2006). Even for a country such as the United Kingdom that has highlighted tackling ambient air pollution as a key priority (Eustice and Lord Goldsmith of Richmond Park, 2021), there are only 171 monitoring stations across the United Kingdom for all pollutants currently being monitored[1]. Therefore, areas without a dedicated monitoring station must have their ambient air pollution concentrations estimated through models. The outputs of these models inform policy for interventions into air pollution, making the models of pivotal importance.

Existing national-level datasets produce estimations at the annual temporal scale and 1 km^2 spatial resolution (UK-AIR, 2019). However, it is imperative to note that health advisories, as articulated by organizations like the World Health Organization, delineate constraints not solely on the annual mean of air pollution concentrations within a specified region but also on the daily mean. To illustrate, for nitrogen dioxide (NO_2), the stipulated limits include a 10 $\mu g/m^3$ annual mean and a 25 $\mu g/m^3$ 24-hour limit (World Health Organization, 2021), with the absence of an explicitly defined hourly limit (World Health Organization, 2021). The regulatory landscape in the United Kingdom, as governed by the Air Quality Standard Regulations 2010 (King's Printer of Acts of Parliament, 2010), delineates both limit values—legally binding parameters not to be surpassed—and target values, akin to limit values but lacking legal bindings. Notably, this legislation addresses hourly level means for pollutants like NO_2, with a meticulous limit of 200 $\mu g/m^3$, not to be exceeded more than 18 times in a year.

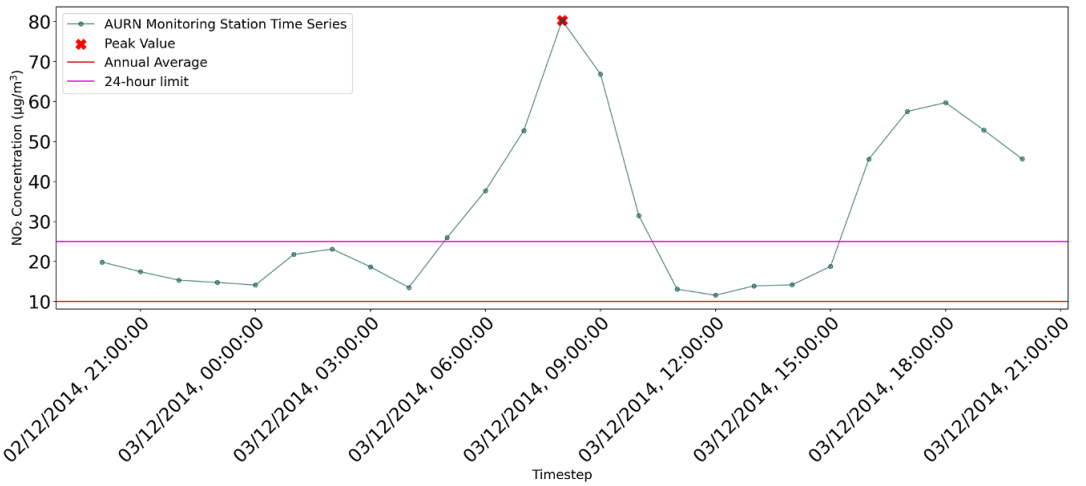
The prevailing methodology, limited to generating mean annual estimates at the national scale, introduces a challenge. This stems from the fact that only specific locales, equipped with monitoring stations, possess hourly data on air pollution. Consequently, areas devoid of such monitoring infrastructure are excluded from any analysis of air pollution levels at a more granular temporal resolution. This discrepancy in data availability raises concerns regarding health inequalities, underscoring the imperative need for a more equitable and comprehensive approach.

The utilization of annual pollution levels provides a broad overview of the pollution within a designated study area. However, a notable challenge arises concerning information loss when transitioning from an hourly to a daily or annual temporal scale. This issue has manifested in the United Kingdom, where instances of divergent narratives emerge between the annual and daily means of specific locations. Take, for instance, Leominster[2] on 03/12/2014 at 08:00, which recorded a peak pollution value for NO_2 of 80.2 $\mu g/m^3$. The 24-hour mean in the vicinity of this peak, spanning 12 hours on either side (from 02/12/2014 20:00 to 03/12/2014 20:00), stands at 31.5 $\mu g/m^3$, as illustrated in Figure 1a. This exceeds the WHO's daily mean guideline of 25 $\mu g/m^3$. The complexity deepens when examining Leominster's annual mean for 2014, registering a value of 9.5 $\mu g/m^3$, deemed safe by WHO guidelines and depicted in Figure 1b. Similar disparities are observable in other monitoring stations. For instance, London North Kensington exhibits unsafe levels at both the annual and daily scales, with a peak value of 209 $\mu g/m^3$, a daily mean of 122 $\mu g/m^3$, and an annual mean of 33 $\mu g/m^3$ for the pollutant NO_2. Supplementary Table S1 details the peak values for the five most polluted stations for NO_2 within the study, encompassing the peak value, daily mean surrounding the peak, and the annual mean for the year of the peak occurrence.

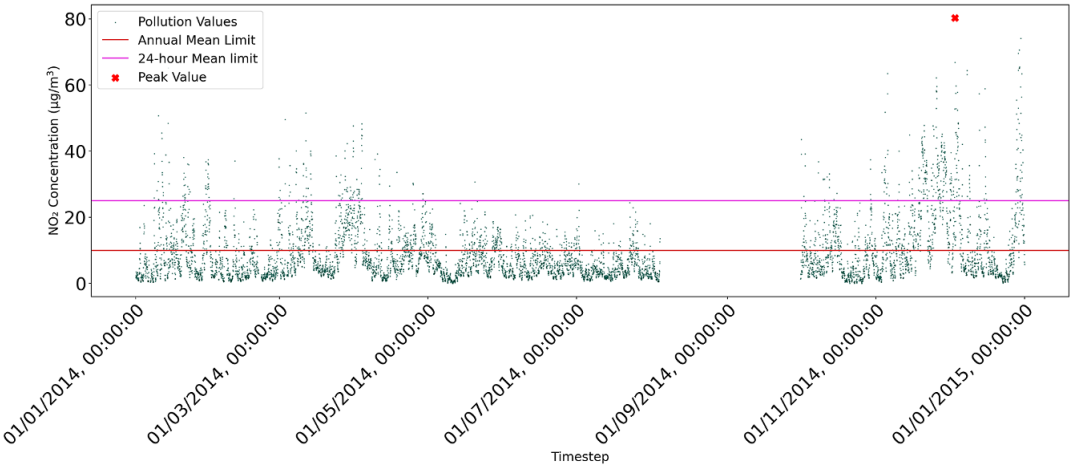
While there is evident importance in hourly air pollution concentration data for compliance and legislation purposes, the data serves a spectrum of other critical purposes. Researchers, policymakers, and public health officials routinely conduct human health assessments (E. Assessment, 1992), enabling informed decisions concerning interventions to protect vulnerable populations (Zou et al., 2009). Further, epidemiological studies assessments are routinely conducted (Atkinson et al., 2016) and are of crucial

[1] https://uk-air.defra.gov.uk/networks/network-info?view=aurn
[2] https://uk-air.defra.gov.uk/networks/site-info?site_id=LEOM

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press



(a) **Leominster day air pollution readings surrounding the 2014-2018 peak.** Shown is the peak value for the Leominster AURN monitoring station spanning 2014-2018, recorded at 08:00 on 03/12/2014. Presented alongside the peak value is the 24-hour window surrounding the peak, along with the annual and 24-hour limit averages.



(b) **Leominster 2014 air pollution readings.** Presented are all ambient air pollution measurements for NO_2 at the Leominster for the year 2014. Emphasised is the peak value, which is further examined in Figure 1a, along with the annual and 24-hour average limits.

Figure 1. *Leominster AURN monitoring station NO_2 measurements. (a) Shows how the peak air pollution reading for NO_2 at the Leominster station dramatically exceeds the 24-hour limit, even more so for the annual limit, showing how there can be periods of quite extreme pollution in the context of the annual limits. (b) shows how there can be extended periods where the air pollution levels are below and exceed the designated limits and the relation of the monitoring station peak to all available data for the station in 2014.*

importance when significant changes in air pollution are being observed, such as during the COVID-19 pandemic (Konstantinoudis et al., 2021). Beyond human health, ecosystem health can be significantly impacted by air pollution, leading to damage to plants, manifested as leaf injury and stunted growth (Molnár et al., 2020). This has critical implications for concerns related to crop yields and food security (Tai et al., 2014).

As such, this work introduces a data-driven machine learning model designed to estimate the hourly concentration of air pollutants at the same spatial resolution (1 km^2) as the existing annual dataset

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

available for the UK[3]. We argue that this work leads to a dataset of substantial value to various stakeholders.

2. Related work

2.1. Measuring air pollution

Various methods exist to offer insight into air pollution concentrations in a given location. The most robust and straightforward method available involves specialized equipment designed to provide direct measurements of air pollution concentrations. In-situ measuring equipment can be broadly categorized into two groups: high-quality stationary monitoring stations discussed in Section 1 and more mobile low-cost air quality sensors (Kang et al., 2022).

While high-quality fixed monitoring stations provide a reliable method of obtaining air pollution concentration data, their deployment on a large scale is prohibitively expensive. In 2018, the United Kingdom had 165 high-quality monitoring stations online across the country within its premier monitoring network. Notably, a majority of these monitoring stations are situated in urban areas, comprising 144 urban and 21 rural stations. The strategic decision of where to position these monitoring stations carries the potential to exacerbate inequality between urban and rural areas, potentially fostering a divide between rural and urban communities in terms of insight into air pollution where they live and work (Rosofsky et al., 2018), particularly when considering that Ozone (O_3) air pollution can often be worse in rural locations (Stasiuk and Coffey, 1974; Belgian Interregional Environment Agency, 2024).

The emergence of low-cost sensors has made it possible to monitor air pollution concentrations over a larger geographic area. However, we see two critical problems with low-cost sensors. One such issue is the quality of the sensors themselves, which can be influenced by changes in atmospheric composition and meteorological conditions, or provide false signals if other air pollutants are present in high concentrations (UK-AIR, DEFRA, 2021). Another issue is the quality control that is conducted on the sensors, such as the calibration checks that go into ensuring that the measurement is made under the same conditions, such as the height at which the measurement is taken, affecting the reading that is produced, potentially making comparisons between different low-cost sensors and even the same sensor between locations more challenging (Concas et al., 2021). There is research being conducted to help combat the issues facing low-cost sensors; it is still an open challenge but rapidly improving (Rai et al., 2017). For now, low-cost air pollution sensors are only suitable for raising awareness rather than applications requiring higher accuracy, such as epidemiological studies or compliance with air quality legislation (Castell et al., 2017).

An ex-situ indirect measurement of air pollution concentrations can be achieved with remote sensing. Sentinel 5P (Veefkind et al., 2012) is an ESA satellite platform that can provide insight into air pollution concentrations at a vast spatial extent. However, a major challenge associated with the use of Sentinel 5P is the issue of data completeness. Two primary drivers contribute to missing data from the Sentinel 5P platform. The first challenge arises from the platform's orbit, which follows a near-polar, sun-synchronous path (European Space Agency - Copernicus, 2023). This orbit causes the platform to consistently pass over a region at a similar time each day. While this characteristic is advantageous for comparing locations, it complicates the provision of insight into air pollution concentrations across an entire day. As a result, questions such as the difference between rush hour and midnight air pollution concentrations become difficult to answer. Another factor contributing to data gaps is environmental conditions that may lead to a specific reading not passing quality control, resulting in missing measurements on certain days (European Space Agency - Copernicus, 2017). Another remote sensing platform is the recently operational TEMPO (Zoogman et al., 2017), which provides hourly air pollution concentration measurements. However, TEMPO shares similar limitations with Sentinel 5P and only offers coverage over North America. While remote sensing is a valuable tool in certain circumstances, it cannot

[3] UK-AIR Annual Modelled Background Air Pollution Data.

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

provide a complete picture of air pollution concentrations. This comprehensive understanding is crucial for designing effective interventions to tackle air pollution.

2.2. Modelling air pollution

As it is evident that all methods of measuring air pollution concentrations have drawbacks, models have been extensively utilized to complement real-world observations. Various types of model frameworks exist, each offering distinct benefits and drawbacks.

There are two widely used air pollution model frameworks: Lagrangian and Eulerian. Lagrangian models track individual air parcels (or particles) moving through the atmosphere. Each parcel is associated with a set of equations of motion, making the parcel the focal point of the model as it moves through space and time (Eliassen, 1984). On the other hand, Eulerian models do not concentrate on a single air parcel; instead, they divide the atmosphere into regions, using fixed points or cells to represent specific locations. The goal is to understand the concentrations of air pollutants at these specific locations at different times (Byun et al., 2003). Lagrangian models are particularly well-suited for studying problems where a specific pollution source is of interest, such as the ash emitted from a volcano (Vitturi et al., 2010).

On the other hand, an Eulerian model is well-suited for studying the spatial distribution and long-term trends of air pollutants, albeit at the expense of not being able to provide specific information about a given source and pollutant. Consequently, this work will primarily focus on Eulerian models, as they provide the necessary data for conducting the analyses and assessments discussed in Section 1.

Statistical Eulerian models, such as Land use regression (LUR), exist as a method for creating stochastic air pollution models. LUR incorporates a variety of predictors, including meteorological, terrain, land use, and road network data (Hoek et al., 2008). Another paradigm of Eulerian air pollution models is represented by mechanistic models, such as GEOS-Chem (Henze et al., 2007). These models are open source and available for use, providing comprehensive spatial coverage of air pollution concentrations. However, they demand a high level of expertise in the domain field for interrogation due to their complexity. Additionally, these models come with extensive requirements for supporting infrastructure, with a GEOS-Chem 4.00°x5.00° degree standard simulation requiring 15GB of RAM[4].

A rapidly emerging area is the use of deterministic models to address the current gap within the existing suite of models, providing high-resolution air pollution concentration data both temporally and spatially; this empowers stakeholders to make informed decisions concerning air pollution. Several models in this category are based on data-driven supervised machine learning, where a target vector, typically representing air pollution concentrations, is estimated from a feature vector, such as meteorological variables (e.g., wind speed). The model's objective is to learn the relationship between the target and feature vectors in situations where both are available, enabling subsequent predictions of target vectors when only the feature vector is available. In the scientific literature, numerous studies utilize machine learning techniques to forecast air pollution concentrations (Freeman et al., 2018; Tao et al., 2019; Harishkumar et al., 2020). However, a limitation exists, as this approach requires air pollution concentration data from the location being predicted before the time that is to be predicted. Therefore, there is a need for historical air pollution data to be available. For example, a forecasting model will use air pollution concentration data from T-X to estimate air pollution at time T, where X is some defined time, such as 1/3/9 hours. If historical air pollution concentrations are used, it restricts the method's applicability to locations where an air pollution monitoring station exists.

Existing studies have tackled the problem of estimating air pollution concentrations in locations without monitoring stations. However, the studies focus either on small geographical areas, such as the Bay of Algeciras (Spain) with hourly temporal resolution (Van Roode et al., 2019) or a large geographical area with low temporal resolution, such as monthly (Chen et al., 2021). Some work has been able to achieve higher spatial coverage with daily temporal resolution (He et al., 2023; Li et al., 2020). As such,

[4] https://geos-chem.readthedocs.io/en/stable/getting-started/memory.html

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

this work presents a model combining these aspects, predicting hourly temporal resolution concentrations across England's large geographical area, a considerable challenge due to the variance of air pollution concentrations in locations covered.

This manuscript aims to use machine learning to produce data similar to an Eulerian model framework. While traditionally, the concentration is resolved over an area in an Eulerian model, the model presented here can be considered an approach of using machine learning as a synthetic monitoring station. The process that is followed for the model is answering the question of the air pollution concentration reading of a monitoring station that experiences the environmental conditions described by the input data. The model takes the training data to learn the relationship between environmental conditions and air pollution, allowing us to use the environmental conditions that are known in all locations across England and make predictions of air pollution concentrations that are not so readily available, providing a complete picture of air pollution concentrations in the England. Compared with other deterministic methods, such as mechanistic models, a key benefit of the approach is the improvement in computational speed. In contrast, more traditional Eulerian models involve spatial dependencies between grids, where, for example, two adjacent grids impact each other. The framework presented in this work, however, treats each synthetic monitoring station as independent from one another. This approach offers a significant speedup in computation through the parallelization of predictions, while also enabling more accessible exploration of data by predicting air pollution locations independently. This novel approach is a key contribution of this work, utilizing machine learning to underpin a scalable estimation of ambient air pollution concentrations. Importantly, this approach is linearly scalable concerning computational complexity, allowing stakeholders to employ a model capable of predicting air pollution concentrations at any spatial and temporal resolution.

3. Data

The data-driven supervised machine learning model this paper proposes for air pollution concentration prediction is based on two primary sets of data: feature vectors and target vectors. In the case of air pollution concentration estimation, the target vector is the air pollution concentration itself, the data to be estimated, and the feature vector represents the data used to make predictions, for example, the wind speed. The model aims to understand the relationship between the feature and target vectors, e.g., what is the given NO_x concentration at given wind speeds? The supervised machine learning model proposed in this work aims to learn a function f that maps from the feature vector $\mathbf{X}$ (which consists of input variables such as wind speed, temperature, etc.) to the target vector $\mathbf{X}$ (air pollution concentrations). This can be mathematically expressed as:

$$y = f(\mathrm{X}) \tag{1}$$

The model attempts to estimate f by learning from historical data, aiming to minimize the difference between the predicted values and the actual values of air pollution concentrations.

3.1. Target vector: air pollution concentrations

We obtained air pollution data for our study from the UK Automatic Urban and Rural Network (AURN) using the OpenAir package (Carslaw and Ropkins, 2012). Our study focuses on seven pollutants: nitrogen oxide (NO), nitrogen dioxide (NO_2), nitrogen oxides (NO_x), particulate matter < 10 μm (PM_{10}), particulate matter < 2.5 μm ($PM_{2.5}$), ozone (O_3), and sulphur dioxide (SO_2). All air pollutants are measured in micrograms per cubic meter ($\mu g/m^3$). All types of monitoring stations were included in the study. The number of station types per pollutant varied, resulting in different data point distributions, as shown in Table 1, with apparent gaps in some locations for certain air pollutants, such as suburban Industrial for PM_{10}, $PM_{2.5}$, and SO_2. Clear spatial differences exist in the locations of monitoring stations. Supplementary Figure S2 shows the spatial distribution of all AURN monitoring stations used in this study across three high-level environmental area classifications.

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

Table 1. *AURN monitoring station counts by environmental classification per air pollutant*

Pollutant name	Urban background	Urban traffic	Rural background	Suburban background	Urban industrial	Suburban industrial
NO_x	40	41	11	3	6	2
NO_2	40	41	11	3	6	2
NO	40	41	11	3	6	2
O_3	32	3	13	2	3	1
PM_{10}	17	22	2	0	5	0
PM_{25}	30	15	2	2	4	0
SO_2	9	1	5	0	3	0

Note. The number of stations for each pollutant within the UK AURN network within England is shown. It can be seen that there is an unequal distribution across the different environment types, alongside some pollutants such as SO_2, missing some environmental types completely.

Each environmental area has a representative area over which the station's measurements relate (UK-AIR, 2023). Urban stations are defined as representative of a few square kilometres (km^2), suburban stations cover tens of square kilometres, and rural stations encompass at least 1000 square kilometres. Each station within the network also has a location type that specifies the primary source of air pollution at the station. Background stations are strategically located to ensure that no single source or street significantly influences the readings at the station. Instead, the measurements reflect an integrated contribution from all sources upwind. Traffic stations are located so that the measurements represent a street segment of at least 100 meters, and industrial stations have a representative area of 250 square meters (m^2).

The monitoring station location was abstracted to the closest grid centroid for ease of creating the needed datasets. Consequently, there is some distance between the true location and the location where we created the feature vector for the monitoring station. This abstraction of location provided a common framework, reducing the computation required to build the associated feature and target vectors, and facilitating a more straightforward interpretation of the framework. While the AURN monitoring station guidelines specify a minimal representative sample area, and the maximum abstraction distance across the monitoring network locations was 399 meters for the London.

Hackney monitoring station, we deemed this to be a worthwhile tradeoff. Full details of the abstraction distance can be found in Supplementary Table S3. It is noteworthy that as the spatial resolution increases, the associated errors will decrease, leading to an improvement in the approach. Eventually, the error from abstracting the location of the monitoring station will be eliminated when the abstracted distance is below the monitoring station representative sample area. However, this comes at the cost of considerable additional computational expenses. Therefore, the experiments in this study represent a lower bound for the framework's performance, as any operational deployment could utilise increased spatial resolution for potential performance improvement.

For the study, we used the years 2014–2016 as the training set, 2017 as the validation set, and 2018 as the test set. To be included in the study, a station needed to have at least one measurement in each of these sets.

We conducted preprocessing on the collected air pollution concentration data. While UK-AIR performs some data validation (DEFRA, Department for Environment Food and Rural Affairs, 2017), we undertook additional preprocessing steps. The initial step involved removing negative values, which are possible in the UK-AIR dataset (DEFRA, Department for Environment Food and Rural Affairs, 2023). The number of observations removed per air pollutant due to the presence of negative concentrations is detailed in Supplementary Table S2. The distribution of the positive air pollution concentration values can vary widely across air pollutants and exhibit apparent differences between different environmental locations of monitoring stations. To visualize the distribution of the different air pollution concentrations, Kernel Density Estimation (KDE) (We˛glarczyk, 2018) was used. KDE is a non-parametric way to

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

estimate the probability density function of a variable. It provides a smooth curve that represents the distribution of data points without making assumptions about the underlying distribution. In the context of air pollution data, the KDE helps visualize the distribution of air pollution concentrations, highlighting patterns such as skewness or multimodality that may not be apparent from raw data alone. The KDE for each air pollutant is shown in Supplementary Figure S3. This variability raises the question of how to identify outliers, alongside the issue of developing a model that can handle target vectors with stark differences in distribution. For example, O_3 is the only air pollutant with a non-zero-inflated distribution, with the skewing between distributions for different environmental classifications for each air pollutant showing varying degrees of skew.

We considered removing outliers from the dataset but ultimately decided against it for several reasons. Points that are distant from the mean have the potential not to be genuine outliers, but rather data points generated by different phenomena compared with the other data points in the distribution. We aimed to identify and remove both outliers and anomalies within the dataset. The challenge in identifying outliers lies in the context of the dataset, where a single urban traffic data point within the context of rural background monitoring station data points might be flagged as an outlier using established methods like Interquartile Range (IQR) (Crosby, 1994). We were also cognizant of the potential presence of anomalies in the dataset. We recognized that a single localized event could drive a high-value concentration data point. While this reading might be accurate, it does not align with the AURN monitoring stations' purpose, where concentrations are intended to represent a larger geographic area. Consequently, we considered identifying and removing these values beyond the scope of this work and proceeded with the study, acknowledging the presence of outliers and anomalous observations within the dataset that we could not explicitly identify.

3.2. Feature vectors

The data considered in this study can be categorized into different dataset families, each containing a set of distinct but related datasets describing a phenomenon associated with air pollution concentrations. Addressing the temporal and spatial resolution differences between the datasets was a key challenge in creating a consistent feature vector to estimate the air pollution target vector. The common framework employed consisted of 355,827 1 km^2 grids covering the extent of the England land mass. England was chosen as the study area since it was a common geographical region in all the datasets examined during this study, as illustrated in Supplementary Figure S1. For the study, seven different dataset families were used, each providing a set of datasets describing a range of related phenomena that correlate with air pollution concentrations. Across all dataset families, there are 152 feature vector elements, with Figure 2 showing example feature vectors across England for each dataset family.

3.2.1. Transport infrastructure structural properties, 28 features

Transport infrastructure has been shown to provide information concerning air pollution concentrations (Berrisford et al., 2022). We used Open Street Maps (Bennett, 2010) to create annual snapshots of 14 transport infrastructure networks, from motorways[5] to residential[6] roads. Using the road network, we calculated two sets of feature vectors. One detailing the distance to each road type from the grid centroid, and the second detailing the total length of the given road type within the grid. Further details on the process conducted can be seen in Supplementary Section S1.3.

3.2.2. Transport infrastructure use, 5 features

Vehicles themselves are a primary driver of air pollution through multiple processes. Road vehicles exhaust gas air pollutants such as NO_x (Watkins, 1991) alongside causing PM air pollution (Yan et al.,

[5] Open Street Maps Motorway Highway Classification (https://wiki.openstreetmap.org/wiki/Tag:highway%3Dmotorway#:~:text=The%20tag%20highway%20%3D%20motorway%20is,local%20context%20and%20prevailing%20convention).

[6] Open Street Maps Motorway Residential Classification (https://wiki.openstreetmap.org/wiki/Tag:highway%3Dresidential#:~:text=The%20highway%20%3D%20residential%20tag%20is,have%20also%20some%20transit%20traffic).

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

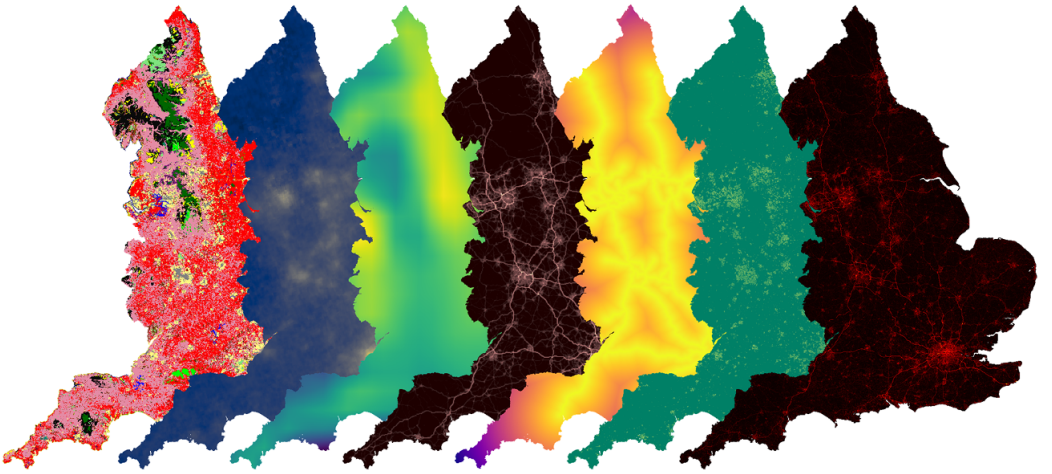


Figure 2. *Example feature vector dataset from each dataset family. From left to right, the example datasets are the majority land use classification for each grid (geographic family, discussed in Supplementary Section S1.8), Sentinel 5P NO_2 measurements (remote sensing family, discussed in Supplementary Section S1.6), 100 m U component of wind (meteorological family, discussed in Supplementary Section S1.5), NAEI SNAP sector 7 (road transport) NO_x emissions (emissions family, discussed in Supplementary Section S1.7), road infrastructure distance from the nearest motorway and total length of residential road per grid (transport infrastructure structural properties family, discussed in Supplementary Section S1.3), and the car and taxis score (transport use family, discussed in Supplementary Section S1.4).*

2011). Further vehicles can cause air pollution through traffic resuspension (Amato et al., 2010) and fuel spillage and evaporation (Haagen-Smit, 1959), alongside $PM_{2.5}$ and PM_{10} from brake, tyre, and road dust emissions (Matthaios et al., 2022). Within England, only traffic counts from point locations detailing daily traffic flows across different road types for key vehicle types such as Cars and Heavy Goods vehicles (HGVs) were available from the Department of Transportation (Department of Transport, UK Government, 2023). Using OpenStreetMaps, we created a spatially complete dataset by providing the average daily traffic flow by road type per meter across the United Kingdom. We then used a spatial microsimulation using data from the UK Census (providing sociodemographic details of different regions of the United Kingdom) (Office for National Statistics, 2017) and the UK Time Use Survey (providing details of how different sociodemographic groups travel, both temporally and by which transportation mode) (Sullivan and Gershuny, 2023) to spatially distribute the daily traffic counts to produce an hourly spatially complete dataset of traffic counts. Complete details of the process are covered in Supplementary Section S1.4.

3.2.3. Meteorology, 11 features

Meteorological phenomena play a pivotal role in air pollution concentrations. Wind speed and direction advects air pollution both to and from locations of interest through horizontal transport (Jurado et al., 2021; Cichowicz et al., 2017). Temperature can have a range of impacts on air pollution, impacting temperature inversions (Wallace et al., 2010), the production of O_3 air pollution (Bloomer et al., 2009). UV radiation directly impacts O_3 production (Finlayson-Pitts and Pitts, 1986). The removal of air pollution from the atmosphere via deposition by precipitation is notable (Jolliet and Hauschild, 2005), alongside wash-off from surfaces (Yuan et al., 2017; Xu et al., 2019). Pressure can also influence air pollution concentrations, either by vertical mixing in low-pressure systems (Ning et al., 2018) or high-pressure systems, causing an accumulation of air pollution concentrations near the ground through a lack of vertical mixing and advection (Vukovich, 1979). Further, O_3 production is increased at higher pressures

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

(Hippler et al., 1990). The boundary layer also influences air pollution concentrations through vertical mixing within the layer. Larger boundary layer heights tend to produce less concentrated air pollution at the surface. For smaller boundary layer heights, the inverse is true. (Xiang et al., 2019; Davies et al., 2007). We retrieved data from the ECMWF Re-Analysis Version 5 (ERA5) dataset (Hersbach, 2016). ERA5 is a global dataset that details the environmental conditions at equal space points, at 0.25°x0.25° hourly resolution. We choose the variables commonly associated with air pollution concentrations in the scientific literature. We interpolated across the study area to provide a value for each $1km^2$ grid centroid; the details of the process of creating the dataset can be seen in Supplementary Section S1.5.

3.2.4. Remote sensing, 5 features

While there are limitations to the data produced by remote sensing, as discussed in Section 2.1, they provide valuable insight into air pollution concentrations between locations. We used monthly averages of Sentinel 5P data (Veefkind et al., 2012) to ensure that all locations had a measurement value. Further details of the feature vector are discussed in Supplementary Section S1.6.

3.2.5. Emissions, 77 features

Emissions of air pollutants are the primary driver of a wide range of air pollutant concentrations. The emissions are classified into 11 SNAP (Selected Nomenclature for Air Pollutant) sectors denoting the emissions source, with particular details discussed in Supplementary Section S1.7. The first sector "Combustion Energy Production and Transformation" (SNAP 1) includes power generation which can produce air pollutants such as SO_2 (Chaaban et al., 2004; Shi and Wu, 2021). Road vehicles exhaust gas air pollutants such as CO, CO_2, NO_x, SO_2 (Watkins, 1991) and are included in the "Road Transport" category (SNAP 7). SNAP 8, "Other Transport and Mobile Machinery" includes shipping which emits NO_x, PM, CO_2 and VOCs (Corbett and Fischbeck, 1997), particularly SO_x from the marine fuels which has a high sulfur content (Tao et al., 2013). Organic waste in landfills ("Waste Treatment and Disposal", SNAP 9) can produce VOCs, a precursor to O_3 (Nair et al., 2019). Agriculture emissions (part of SNAP 10, "Agriculture, Forestry and Land Use Change"), comprise a large source of air pollutants, for example, 39% of global $PM_{2.5}$ is caused by ammonia from livestock manure and urine and synthetic nitrogen fertilisers (Gu et al., 2021). Other emission sectors are "Combustion in Commercial, Institutional, Residential and Agriculture" (SNAP 2), "Combustion in Industry" (SNAP 3), "Production Processes" (SNAP 4), "Extraction and Distribution of Fossil Fuels" (SNAP 5), "Solvent Use" (SNAP 6) and "Nature" (SNAP 11).

3.2.6. Land use, 22 features

The land use composition of a given area is related to air pollution concentrations, such as throughout greenspace (Nowak et al., 2002; Nowak et al., 2006) and urbanization (Arnfield, 1990; Yassin, 2011). Land use composition profiles were created for each grid using the UKCEH 25m Land Cover Maps (Rowland et al., 2017). Details of the process and the different land use types are discussed in Supplementary Section S1.8.

3.2.7. Temporal aspects, 4 features

Air pollution displays various temporal cyclical elements, including diurnal cycles caused by rush hour for NO_2 (Goldberg et al., 2021), UV radiation for O_3 (Garland and Derwent, 1979), and boundary layer height for all pollutants (Su et al., 2018). Weekly trends also emerge due to the working week affecting transportation and industrial emissions for NO_x (Beirle et al., 2003), with similar patterns observed for PM (Gietl and Klemm, 2009). Seasonal cycles for PM are evident due to winter residential heating (Feng et al., 2014), with similar factors contributing to an increase in SO_2 (Meng et al., 2018). Furthermore, winter has a higher probability of adverse meteorological conditions, which reduces vertical mixing (Li et al., 2022). Additionally, colder temperatures and reduced sunlight in winter months affect O_3 production (Cichowicz

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

et al., 2017). Consequently, the hour, day of the week, week number, and month were all included as elements in the feature vector.

The following section justifies the inclusion of such a broad range of different datasets by exploring the relationship between the feature vector and the air pollution concentrations at AURN monitoring stations. This analysis highlights the differences between types of air pollution and variations experienced at different environmental types for a single type of air pollution, ensuring robust estimation of air pollution concentrations across all environment types found in England.

4. Feature selection

4.1. Air pollution and feature vectors

Air pollution can be attributed to various sources, with different processes influencing its concentration, as discussed in Section 3.2. The issue of different drivers of air pollution is further complicated when considering that at different locations, the key phenomena driving the pollution concentrations are different, as detailed within the AURN environment classification discussed in Section 3.1. This tangled web of sources, such as road transport, and sinks, such as wind speed, makes it challenging to identify a common set of datasets that can provide insight into all the air pollutants of interest. Therefore, this section aims to disentangle the relationship between the different air pollutants and monitoring station environment types to provide insight into the different feature vectors and their relationship with the air pollutant measurements. This provides insight into the benefit of each dataset, detailed in Section 3.2.

Figure 3 shows the average Spearman correlation coefficient across all monitoring stations for NO_x and O_3. These figures depict the 10 highest magnitude feature vector elements in both directions. The differences in the contributing sources of air pollution for a given pollutant are evident and appear to support the scientific literature regarding the relationship between different air pollutants and their sources and sinks.

For example, NO_x has the highest positive Spearman correlation coefficient with the emissions dataset, particularly SNAP sectors 1 and 2, indicating a strong relationship between energy production and transformation emissions and the commercial, institutional, residential, and agriculture sectors. It is also notable that the highest magnitude negative Spearman correlation belongs to the sinks, namely wind speed and the boundary layer height, as expected.

In contrast, O_3 presents a very different situation compared to NO_x; the highest magnitude correlation is inverse. Both wind speed and boundary layer height have a high positive magnitude Spearman correlation, highlighting that the same phenomenon can have a completely opposite relationship on the concentrations of air pollutants. It also follows that the correlation between "Downward UV Radiation At Surface" and O_3 has a positive correlation, given that O_3 is produced under sunlight by the precursors of NO_x and Volatile Organic Compounds (VOCs) (United States Environmental Protection Agency, 2023), highlighting that more sunlight results in more O_3 being produced.

Figure 4 illustrates the relationship between feature vectors for Rural Background and Urban Traffic monitoring station environment classifications (discussed in Section 3.1) for the air pollutant NO_x. Each subclassification of monitoring stations indicates the primary contributor to the measured air pollution. Background stations have no single primary source, while traffic stations are primarily driven by traffic.

For the Urban Traffic station, the strongest positive correlation across air pollutants is with SNAP Sector 7, denoting road transport emissions. Notably, there is a strong relationship with SNAP Sector 6 NMVOCs, indicating solvent use for NMVOCs. This might be explained by the small number of data points for monitoring stations (41 Urban Traffic stations), alongside emissions data being based on extensive scaling depending on the hour, week, and month of interest (UK National Atmospheric Emissions Inventory (NAEI), 2023). The potential for confounding variables, such as NMVOC emissions arising from vapor from petrol (UK National Atmospheric Emissions Inventory (NAEI), 2023), makes the 0.014 magnitude difference in relationship strength negligible. The current data quality is suitable for identifying general trends rather than pinpointing the most substantial relationship by sector.

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

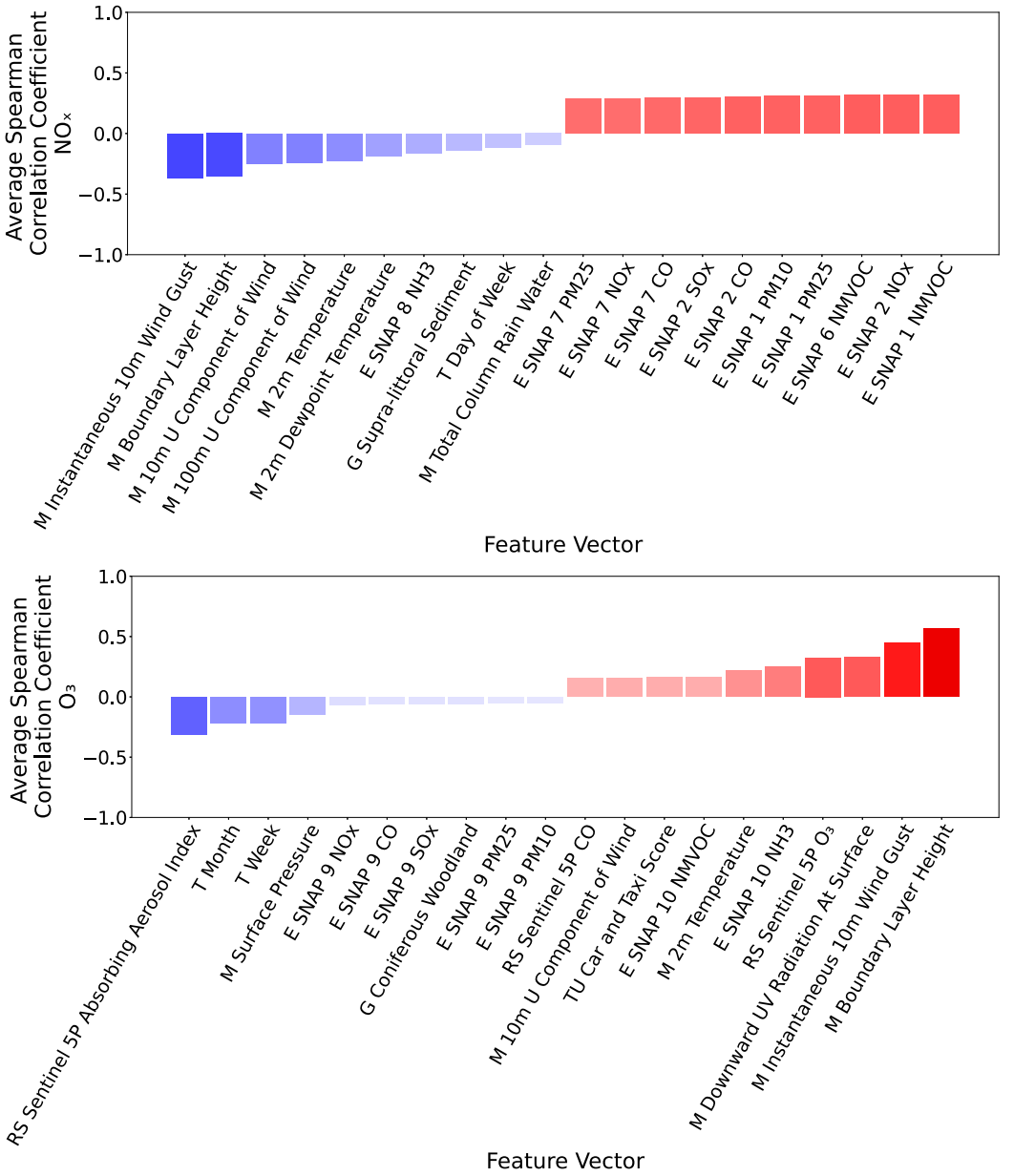


Figure 3. *Spearman correlation coefficients overall mean for all pollutants. The mean Spearman correlation coefficients for NO_x and O_3 across all the environmental classifications of the AURN network for the 10 most extreme, both positive and negative, for the feature vectors are shown. The sources and sinks of the air pollutants are different, aligning with the scientific literature (Section 3.2), with NO_x being highly positively correlated with emission features, whereas O_3 exhibits such a relationship mainly with meteorological features, such as wind gusts. Regarding negative correlations, the two air pollutants exhibit counter relationships, with NO_x having a negative correlation with the meteorological. The analysis highlights how the relationships between a particular phenomenon and a given air pollutant can be widely different in strength.*

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

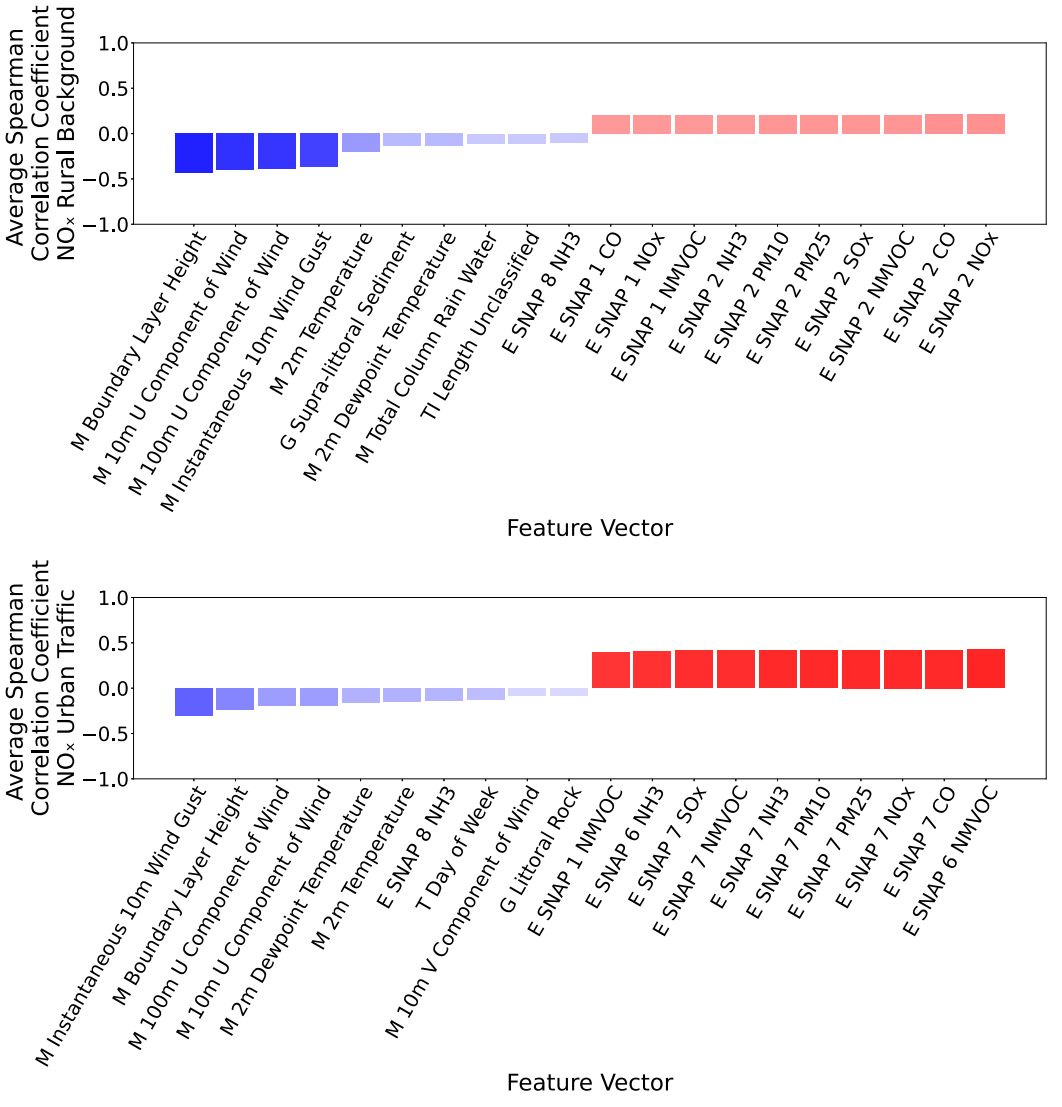


Figure 4. *Spearman correlation coefficients for NO_x monitoring station environmental subclassification locations, Rural Background and Urban Traffic. While Figure 3 highlights the difference between phenomena and air pollutants, there exists a further difference between environmental subclassifications. For the Urban Traffic monitoring stations, it can be seen that the primary positive correlations are related to road transport as would be expected (the strong relationship with solvent use is likely an artefact of the scaling performed and discussed in Section 3.2 and Supplementary Section S1.7, alongside a limited sample size of 41 stations). In contrast, the Rural Background monitoring stations show a strong relationship with emissions from the residential sector, highlighting that the sources and sinks for an air pollutant depend on the air pollutant itself and the location of interest.*

The transport use dataset exhibits a positive but weaker Spearman correlation with the Urban Traffic site, with an average score of 0.36 across the five datasets. In contrast, the Rural Background sites for the transport use datasets have an average of 0.05. This aligns with the literature, showing a clear signal for increased traffic near an Urban Traffic monitoring site and increased NO_x concentrations. It also agrees with the AURN environment classification, with a still positive but significantly reduced magnitude correlation.

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

As the AURN environment classification is based on the primary emitters closest to the station, it then follows that both station types have the same feature vector for the highest negative magnitude Spearman correlation with the expected boundary layer height and wind, namely the meteorological variables that are present across all monitoring stations, and the sinks in the case of NO_x.

4.2. Inter feature vectors

While there is considerable existing literature about the relationship between different air pollutants and the phenomena covered in the datasets used in this study, considerably less literature covers the relationship between the phenomena comprising the feature vector. This section aims to understand the relationship between the different feature vectors to address the issue of multicollinearity, which can have significant implications for the machine learning approach implemented.

The Spearman correlation coefficient was again used to calculate the relationship between each pair of feature vectors. Figure 5 is a heatmap representing the Spearman correlation coefficient value for every pairing. There are no air pollution monitoring stations for some feature vectors—nine feature vectors for the emissions dataset family and four in the geographic dataset family. The lack of target vector data at

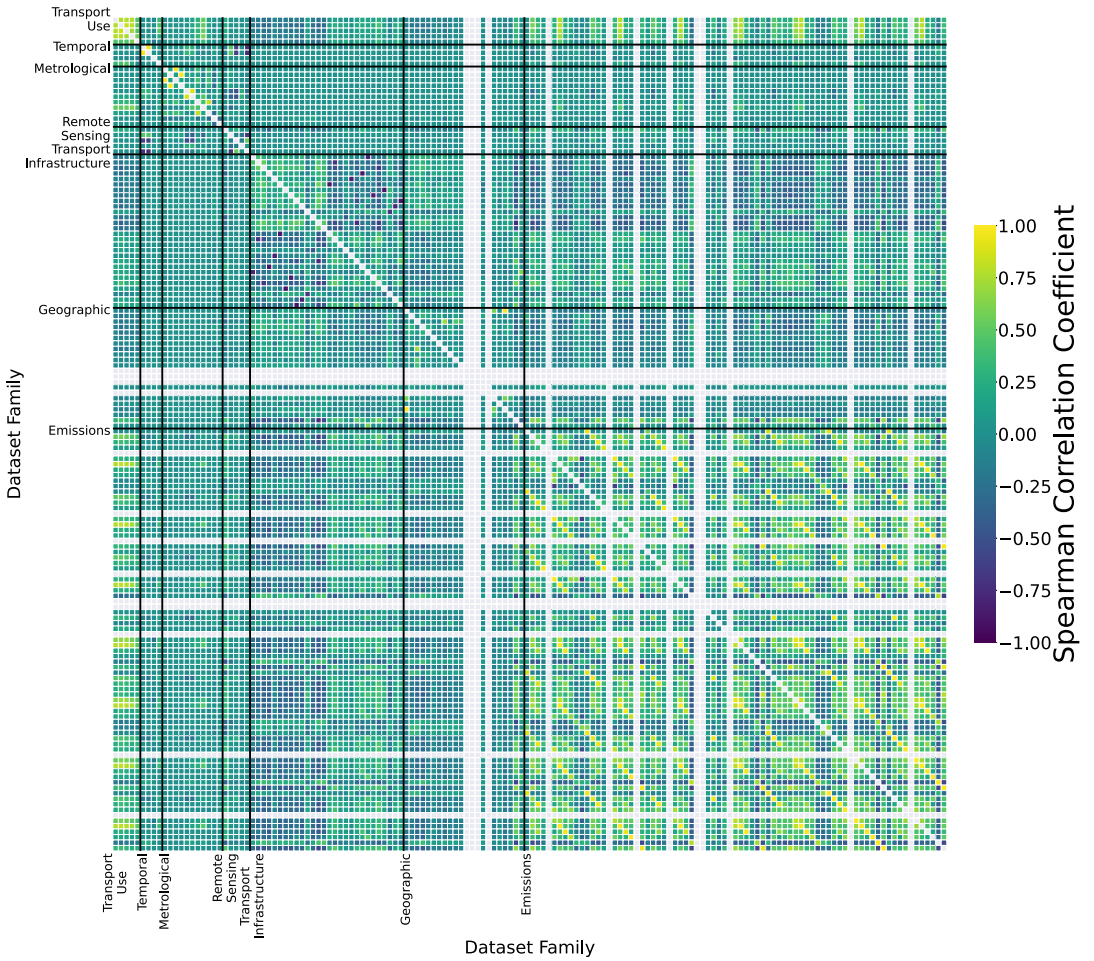


Figure 5. *Spearman correlation heatmap between all feature vectors. The grey lines throughout the heatmap show the data points missing from the dataset, phenomena with no monitoring stations across all pollutants, including four geographic features and nine emissions features.*

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

some key feature vector locations presents the first significant problem with any model developed: not all environmental condition types experienced within England have observations. For example, there are no air pollution concentration data for the Land Use classification of Saltwater, representing a scenario in which the model has no exposure.

From the heatmap in Figure 5, it is clear that multicollinearity is present between some features, for example between different species within the same emissions sector. Two main concerns were identified with the complete set of feature vectors: the implications on model interpretation depending on the model chosen and the redundant information between features. The model interpretation would impact future stakeholder engagement, and redundant features would make the hyperparameter search more complex. Therefore, we considered removing features entirely or creating a new set of features through dimensionality reduction. Stakeholder engagement with the model is crucial, as many stakeholders rely on the interpretability of the model to make informed decisions. Creating a new set of features through dimensionality reduction techniques would have reduced the ability to explain or understand the significance of certain sets of features. Preserving the original semantic meaning of features allows stakeholders to directly relate model inputs to real-world phenomena, ensuring that the model remains transparent and interpretable. Therefore, we chose not to create a new set of features and instead prioritized maintaining the original feature set for better communication and practical use by stakeholders.

Hierarchical clustering was performed between the Spearman correlation of the feature vectors, allowing us to create a more complex method of grouping together feature vectors. Figure 6 shows the clustering results. We used Ward's linkage method (Ward, 1963), which minimizes the variance within clusters to ensure homogeneity. The linkage distance provides a consistent metric across all feature vectors to explore the similarity of features and provide clusters of features depending on the value of the linkage distance provided. Figure 6 shows how related some of the feature vectors are; for example, the 100m and 10m components of wind in both directions are highly correlated and therefore have a very low linkage distance. There are also more complex relations between the feature vectors, such as within the transport use datasets. Still, there are differences within the data set, such as car and taxi and bus and coach being highly related but not to the same degree as HGVs. The motivation for performing hierarchical clustering is to allow for a subset of features to be selected that provide the same information as one another, aiding model interpretation. In the case above, the idea is that including the 100m U component of wind provides the same information as the 10m U component of wind, so there isn't a need to include both.

Supplementary Table S7 shows the number of clusters at varying linkage distances, where increasing the linkage distance results in fewer clusters as the information provided between datasets isn't required to be as strong.

As including redundant feature vectors increases the computation costs of creating and using the model rather than impacting the performance of the predictions, we decided to keep all feature vectors when training the model.

The models discussed in the following sections use all 152 feature vector elements. The intention is to provide a baseline performance of a machine learning model that utilizes all the datasets covered while allowing for an understanding by individual stakeholders of the redundant feature through the hierarchical clustering performed, allowing them to subset the datasets as desired for their particular use case. The idea of using different subsets of the 152 feature vector elements is explored in Section 5.6; however, up until that section, all 152 feature vector elements are used. Using all the feature vectors does mean, however, that a machine learning approach that is robust to multicollinearity needs to be chosen. The second issue of model interpretability implications is discussed in Section 5.1.

5. Modeling

Section 5 starts by describing the reasoning behind different modelling choices. The model's performance in two critical scenarios is then explored: forecasting and estimating missing stations. Forecasting aims to answer the question of, given a location the model has already seen, how well the model performs when estimating a future year it has not. The estimating of a missing station then experiments with

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

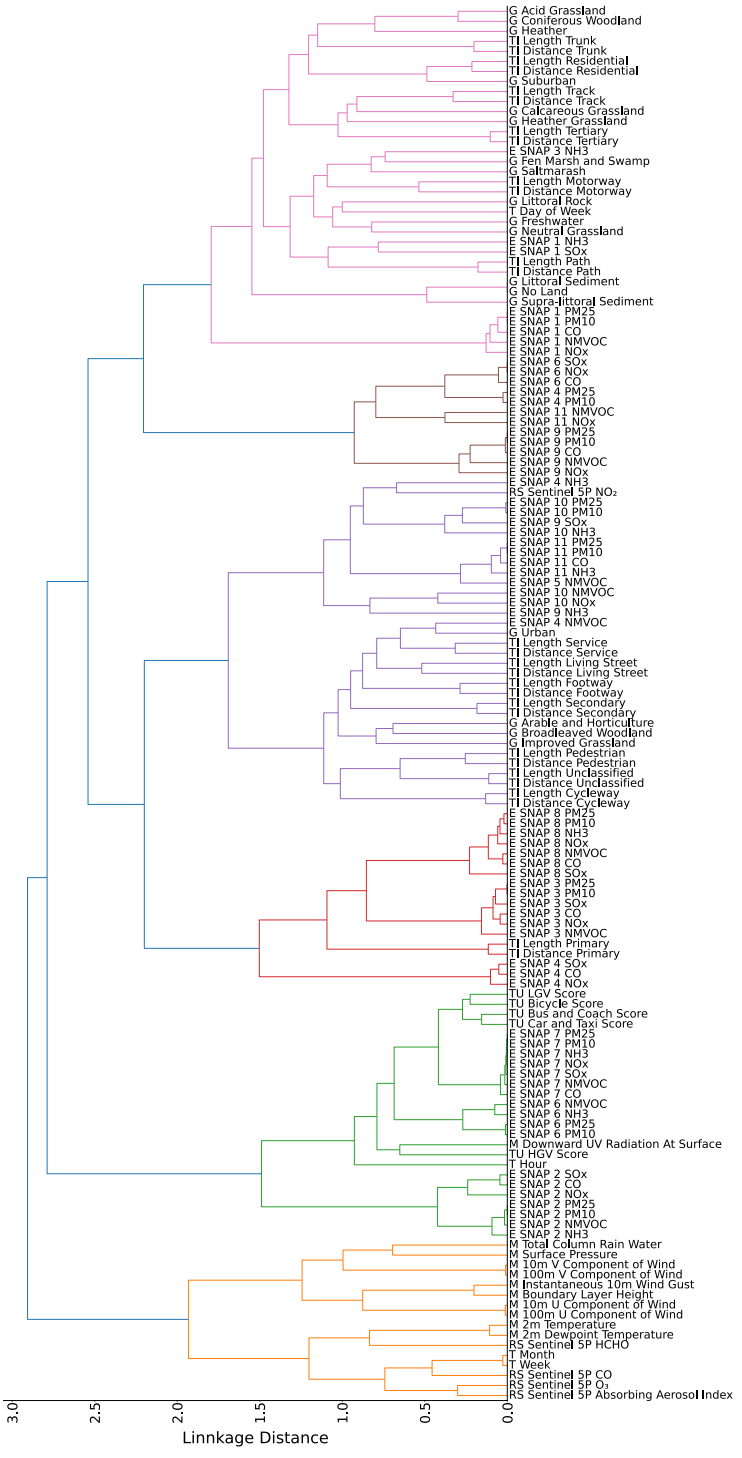


Figure 6. *Dendrogram depicting hierarchical clustering of feature vectors. The lower the linkage distance between feature vectors, the more correlated the features are, indicating that they provide similar information. Supplementary Table S7 details the number of clusters for different linkage distances.*

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

understanding model performance when predicting air pollution at a location it has never seen before. The model's performance on peak concentrations is also analyzed, a critical situation for the model to perform well. Finally, a justification for including such a wide range of datasets is motivated by experimenting with a model that predicts based on one dataset family.

5.1. Model design and training

The first consideration when choosing the model framework was the need for the model to be robust to multiple uninformative and redundant features. As discussed in Sections 4.1 and 4.2, uninformative features exist for various reasons. Some features are uninformative for specific air pollutants, such as the transport use features with O_3. Features can also be uninformative for specific environmental locations of monitoring stations, as seen in Section 4.1 for rural background NO_x stations. Multicollinearity further compounds the issue, allowing multiple features to extract the same information about air pollution measurements, as discussed in Section 4.2. The second consideration when choosing the model was for it to be robust to outliers and anomalies, as discussed in Section 3.1. The model we chose to use was LightGBM, a gradient-boosting algorithm based on decision trees.

LightGBM (Ke et al., 2017) was identified as a machine learning approach that could address the concerns raised in the study through various techniques while also providing state-of-the-art performance on tabular prediction problems. Tabular prediction problems are those that are based on structured data, which is typically organized in rows and columns. Each row represents an observation, and each column a feature of that instance. In our case, the air pollution data is structured in this tabular format, with various environmental factors (such as wind speed and temperature) serving as the features used to predict air pollution concentrations. LightGBM is a gradient-boosting decision tree (GBDT) algorithm where an ensemble of decision trees is trained in sequence, with the n + 1 decision tree fitting the residuals of the first n decision trees, learning the difference between the actual target vector and the weighted sum of predictions of the first n decision trees. For illustration purposes consider a decision tree that predicts air pollution concentration y based on wind speed X_1 and temperature X_2. A possible structure of the tree could look like:

$$y = \begin{cases} y_1, \text{ if } X_1 \leq 5\,\text{m/s and } X_2 \leq 20^{\circ}C \\ y_2, \text{ if } X_1 \leq 5\,\text{m/s and } X_2 > 20^{\circ}C \\ y_3, \text{ if } X_1 > 5\,\text{m/s and } X_2 \leq 25^{\circ}C \\ y_4, \text{ if } X_1 > 5\,\text{m/s and } X_2 > 25^{\circ}C \end{cases} \tag{2}$$

Here, X_1 (wind speed) and X_2 (temperature) are feature vectors that the decision tree uses to make predictions about y (air pollution concentration). Each branch of the tree represents a different split based on feature values, leading to different predictions y_1, y_2, y_3, y_4. A parameter in this context would be the predicted values y_1, y_2, y_3, y_4, which are learned during model training. A hyperparameter would be the number of splits or branches in the decision tree, which is set before training and influences the model's complexity.

LightGBM allows us to mitigate the impact of uninformative and redundant features on training time through the tree-building algorithm. The approach that LightGBM takes when building the decision trees is to split observations based on the feature vector values, looking for the best possible split regarding information gain and reducing the uncertainty regarding the target vector. This involves grouping homogenous instances of data points, such as instances where there is high transport use at a monitoring station that is measuring high concentration readings for NO_x.

One of the core issues with our air pollution training data is that many data points within the datasets repeat the same information due to the cyclical nature of air pollution measurements, causing a considerable amount of bloat in the datasets. The standard approach to identifying split points within a GBDT is the pre-sorted algorithm where all possible split points are explored, an approach which, in this use case, would be highly costly regarding computation and memory. LightGBM helps tackle this issue by

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

using histograms when performing the splits, where continuous variables are put into discrete bins, changing the computational cost from being dependent on the number of data points to the number of discrete bins created.

The second concern identified when exploring the datasets was the presence of outliers and anomalies within the dataset. LightGBM inherently tackles this problem via decision trees being the underlying learner within the model. The decision tree's goal is to group homogenous data instances, and there is an ability to set a minimum number of data instances that comprise a valid leaf on the decision tree via the "minimum data in leaf" model parameter.

The "minimum data in leaf" parameter allows for a minimum threshold of homogeneous data points for the LightGBM algorithm to view as a set of data points that should be learned from and used in predictions. In this context, homogeneous refers to instances where the data points share similar feature values, such as numerous data points with a short distance to a motorway road feature and a high air pollution concentration. Concerning the air pollution prediction problem as a notional example, say there is a high wind speed and low traffic count but a high air pollution concentration reading, which only occurs once in the dataset; LightGBM will not create a leaf for this data instance. The scenario described could plausibly happen if a single air pollution emitter passes by the station, causing an artificially high measurement that would not represent the geographic area intended for the station as outlined by the AURN documentation, as discussed in Section 3.1.

There are, however, some trade-offs to the LightGBM solutions presented above. The feature importances given for the feature vectors via the model will likely be misleading due to the multicollinearity present. For example, the most extreme case seen in the hierarchical clustering of multicollinearity is for the wind speeds at 10 m and 100 m, where both features exhibit a strong correlation, meaning that we can extract information about air pollution from either feature. Therefore, during model building, in the split performed, the model would use only the 10 m or 100 m component of the wind direction, as they would present the same information gain about the target variable as each other. As the feature importance given by LightGBM is based on the number of times a feature vector is used, the total number of times the two feature vectors are used may be split across the two features, reducing the feature importance given to each one. There are also implications for any sensitivity analysis conducted, as it is possible that we could increase/decrease the 100 m component of wind and there be a misleading change in the air pollution concentration prediction if the model used the 10 m wind component as the split point. The feature importances given must be analyzed considering the clusters presented in Supplementary Table S7, treating each of the clusters' feature importances together. Another method to understanding the model would be using SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017). SHAP is a game-theory-based approach used in ML to explain the contribution of each feature to a model predictions. It assigns each feature a "SHAP value" that represents its impact on the prediction, providing a more nuanced understanding of feature importance by accounting for feature interactions and multicollinearity. Unlike traditional feature importance methods, SHAP ensures that the contributions of all features are fairly distributed, making it a robust tool for model interpretability.

A key consideration during the model design was the choice of the loss function. The loss function represents the error of a given prediction, in this case, quantifying the difference between the prediction and the actual air pollution concentration measurement of a model, thereby allowing for comparisons between models and subsequent choice of the optimal model. The choice of the loss function in this situation was between the mean absolute error (MAE) and the mean squared error (MSE) (Hodson, 2022). The MAE would help reduce the influence of higher air pollution measurements on the model present due to the known presence of outliers and anomalies within the dataset. However, these high air pollution measurements are of vital interest within the context of air pollution, even if they are potentially erroneous. So, a tradeoff of potentially overfitting on these higher values was seen as a worthwhile tradeoff, and as such, the MSE was chosen as the loss function. The underlying premise is that 10 $\mu g/m^3$ is more than twice as bad for human health than 5 $\mu g/m^3$, so using the MSE is more appropriate given the domain in which the model would be used, supporting the existing literature that there is a non-linear

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

relationship between the detrimental effects of air pollution concentrations (Yang et al., 2022; Zhao et al., 2019).

As an air pollution concentration can never measure less than zero, we trained the model on the log transformation of the target vector. We added a small constant of 1×10^{-7} to the target vector and then performed the log transformation due to the presence of 0 concentration measurements within the dataset. The log transformation and addition ensured that the model would never predict a negative value as the model output, as the reverse transformation of calculating the exponential and subtracting 1×10^{-7} was performed on the output. A further hyperparameter explored during model training was L2 regularization (Hoerl and Kennard, 1970). Including L2 regularization helps distribute the weights within the decision tree, encouraging the weights to be closer to 0 but keeping all feature vectors, ensuring that no single feature vector drives predictions, which is key with the considerable number of feature vectors used.

The framework for choosing the model's hyperparameters was a randomized grid search of 40 hyperparameter sets. The hyperparameters we optimized during the randomized search were the L2 regularization and the min data in each leaf already discussed, alongside the number of leaves, the number of trees, and the max depth (Microsoft, 2023). The number of leaves search space was given the range of 1,000 to 4,095 (Mishra, 2023), with the optimal values being chosen near the centre of this range, validating its choice. The number of trees was controlled via early stopping, where no additional tree would be added after 30 trees had been added without any improvement in the loss function performance. Similarly, the max depth was not limited and left to grow as needed until performance did not improve during training.

Some model parameters were kept constant throughout the search, such as the max bin, kept constant at 255. The max bin refers to the number of discrete bins created for a continuous feature vector. 255 was chosen to ensure that a range of different splits during model training could be created while also helping to reduce training time by allowing data to be stored optimally as an int8 data type. The boosting type used during training was Gradient-based One-Side Sampling (GOSS) (Ke et al., 2017). GOSS is a method of boosting that allows the $n + 1$ decision tree discussed at the start of this section to be trained on a subsample of the data. The subsample of data chosen is the data that has a large gradient, that is, the data has yet to learn well from in the model and a random sample of the small gradient data, helping to reduce the amount of data used drastically, and therefore training time. The tradeoff with GOSS is the potential for overfitting when the datasets are small; however, this was not a concern in the context of air pollution.

The final consideration was the grouping and number of models to develop. One possible choice was creating a single unified model with all seven air pollutants comprising the target vector. However, due to the considerable imbalance in the number of data points and the issue of every monitoring station measuring a different subset of pollutants, the number of locations with every air pollutant measured at the same timestamp was minuscule. Another possibility was to create an individual model for each environment type covered in the AURN environment classification, such as Urban Traffic, Rural Background. However, this approach presented the problem of requiring a determination of the environment type of every grid in the study where there wasn't an existing monitoring station. Therefore, we created a single model for each of the different air pollutants mixed with all the different environment types, the benefit of which is simplifying the process of estimating a never before seen location while making use of all of the air pollution observations possible.

During the hyperparameter grid search, data from 2014 to 2016 was used as the training set, with the validation set being 2017 and 2018 as the test set. We split the dataset temporally to ensure no data leakage and to give an intuitive sense of the performance metrics gathered. We chose the best parameter set based on the model's MSE on the validation set across the parameter sets. Subsequently, using the best parameter set to train a model with both the training and validation set, with performance evaluated using the test set, data the model has never seen. The R^2 score for each model on the different sets was calculated at each stage.

To allow flexibility in extending the model with new data, we deliberately excluded feature vector elements that would identify monitoring station details, such as names or locations. Additionally, we opted not to include lags of air pollution concentrations, like using the concentration at T-1 to estimate the

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

concentration at time T. This decision enables us to make predictions even in the absence of a specific air pollution measurement, promoting the use of observations as independent entities. This structure facilitates the tabular format, leveraging the state-of- the-art performance of LightGBM. The temporal and spatial independence of observations supports the creation of a lightweight model, conducive to parallel computation for different locations and time points. Our experiments addressed two crucial scenarios, providing insights into the model's temporal and spatial performance. We assessed its ability to forecast air pollution concentrations into the future (discussed in Section 5.2), and extended the analysis to evaluate the model's performance in estimating air pollution concentrations in spatial locations not previously encountered (discussed in Section 5.3).

5.2. Filling temporal missing data

The initial set of experiments focused on evaluating the models' capability to predict future air pollution concentrations at locations already included in the model. These experiments aimed to assess the model's performance in forecasting scenarios. The achieved performance serves as a conservative estimate for filling in missing data temporally in a given monitoring station's time series, considering that air pollution concentration readings at the estimated time would be available in an operational hindcast situation.

Table 2 presents the R^2 score for the model developed at each stage, as discussed in Section 5.1. The results illustrate a degradation in the model's performance as it moves temporally away from the data it was initially optimized for during the randomized parameter grid search. These experiments demonstrate that the model parameters identified during the randomized search remain consistent across the three datasets, with minimal performance loss observed between the validation and test sets. This observation supports the idea that the model is effectively learning the true relationship between the feature and target vectors.

In the best-case scenario, NO_2 shows no drop in performance (rounded to 2 decimal places) between the validation and test sets. The most significant performance decrease is observed in SO_2; however, it is important to note that this may be influenced by a data issue, as SO_2 has significantly fewer stations (18) compared to NO_2 (103), as detailed in Table 1.

While a benefit of the model presented is the ability to forecast air pollution concentrations into the future, answering the question of what air pollution concentrations at a station will look like in the next year, the adaptable temporal and spatial independence discussed in Section 5.1 allows for the model to be used to estimate missing data. Figure 7 shows the model used to estimate the missing data in the NO_2 observations for the Chesterfield Loundsley Green monitoring station from 2014 to 2018. There are two possible cases for the missing data being filled in. The first is to backdate or postdate the observations

Table 2. R^2 *scores depicting forecasting performance (2014–2016 train score, 2017 validation score, 2018 test score)*

Pollutant name	Dataset train score	Dataset validation score	Dataset test score
NO_2	0.85	0.77	0.77
NO_x	0.82	0.75	0.74
NO	0.73	0.67	0.65
O_3	.80	0.70	0.67
SO_2	0.45	0.43	0.30
PM_{10}	0.51	0.38	0.32
PM_{25}	0.55	0.35	0.29

Note. The dataset train score shows the model's performance in capturing the relationship of the training data shown with the validation showing the performance in 2017 and the test score on 2018 data. The similar performance between the validation and test scores shows that the model optimized during the parameter search is learning the true relationship between the features and air pollution that is robust to data never seen before.

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

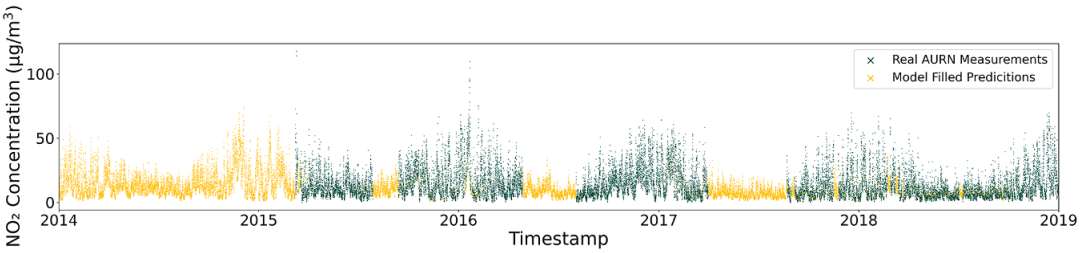


Figure 7. *Chesterfield Loundsley Green NO_2 concentrations augmented dataset, with missing AURN measurements filled with model predictions. This figure shows that the station's measurements (green) started in early 2015 with three clear periods of long-term missing data. The model predictions (yellow) can create a complete augmented time series using the model.*

depending on when the station came online or was decommissioned. Chesterfield Loundsley Green came online on 01/03/2015 (UK-AIR, DEFRA, 2023a), but the model can backdate the observations to 01/01/2014, extending the readings and helping to create a complete dataset. It is also possible to extend the life of a station if it was taken offline by filling in observations since the station was decommissioned (UK-AIR, DEFRA, 2023b).

The second situation where the model can fill in missing data is when an issue at the monitoring station or associated infrastructure causes the station to go offline and measurements not to be reported (UK-AIR, DEFRA, 2023c). However, one potential issue with this approach is if there is a particular reason that the site cannot report data, for example, when wind speeds are over a defined speed. This situation would indicate that no data is within the training set concerning this specific situation, indicating the model is extrapolating. However, this is not a concern in this situation as AURN reports the reasoning behind data not being reported, such as a communication issue or an instrument error. It is something, however, to be understood in the context of any future work that uses this framework where this situation could occur. In Figure 7, there are three prolonged periods in which Chesterfield was not reporting NO_x measurements. The model presented can fill in these periods alongside the periods from before the station came online to create a time series for the station that has all available data as seen in Figure 7 where the real measurements from the station have been augmented with the model output where real measurements are not available.

5.3. Filling spatial missing data

The second set of experiments that we conducted explored the ability of a model to be trained and predict the complete time series for another station, never seen before. We used 5-fold leave-one-out validation (LOOV) to experiment with this scenario. The results from this experiment provide an understanding of how the model performs when filling in missing air pollution concentration data spatially, a situation akin to using the model as synthetic stations across England at locations where no station has ever existed.

The same experimental design as Section 5.2 was repeated alongside a final step that calculates the LOOV score for every station not included in the training, validation or test set. Table 3 shows that the models trained during the 5-fold LOOV can retain their future predictive performance, with minor differences for the performance of ai pollutants across the different subsets of stations used, showing the results from Section 5.2 to be robust to changing input datasets. Table 4 shows the LOOV summary statistics for the experiments conducted, based on the R^2 retrieved from the model estimating the complete time series of a monitoring station's data. Four different summary statistics were considered from the set of LOOV results, namely the mean, median, min, and max results. The max LOOV results are positive for all of the pollutants, indicating some merit to this approach across all pollutants. This is further supported by the majority of positive results in the mean and median LOOV for all pollutants apart from SO_2. Of central interest is the LOOV min, where for all pollutants other than PM_{10} there is a negative R^2, indicating

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

Table 3. R^2 *scores depicting forecasting performance for 5-fold leave-one-out-validation*

Pollutant name	Dataset train score	Dataset validation score	Dataset test score
NO_2	0.85	0.77	0.77
NO_x	0.82	0.75	0.74
NO	0.73	0.67	0.65
O_3	0.81	0.70	0.67
SO_2	0.46	0.42	0.29
PM_{10}	0.54	0.38	0.32
PM_{25}	0.58	0.34	0.29

Note. The experiment conducted aimed to ensure that with different subsets of monitoring stations, the forecasting performance of the model remains robust. Shown with Table 3 having similar performance as the experiment result shown in Table 2, particularly the test score, data the models have never seen before.

Table 4. R^2 *scores for missing monitoring stations performance summary statistics for 5-fold leave-one-out- validation*

Pollutant name	Estimation LOOV max	Estimation LOOV min	Estimation LOOV mean	Estimation LOOV median
NO	0.62	−2.28	0.10	0.21
NO_2	0.70	−1.75	0.25	0.37
NO_x	0.67	−1.46	0.20	0.32
O_3	0.78	−1.95	0.45	0.62
PM_{10}	0.59	−0.17	0.35	0.38
PM_{25}	0.59	0.23	0.45	0.46
SO_2	0.10	−1.65	–0.20	–0.02

Note. The summary statistics show the R^2 for each monitoring station in the study for a model that has never seen the stations' data before. The approach has clear merit, with all air pollutants having a positive maximum score. However, some monitoring stations have a negative minimum score, driven by their unique nature concerning the feature vectors and phenomena driving the air pollution concentrations at a given location. The mean and median R^2 scores show that the approach works for most stations for most air pollutants in estimating air pollution concentrations at a missing location.

a prediction across the time series that is worse than simply predicting the average concentration. In this case, a potential hypothesis for the model performance is a lack of data available. The performance of SO_2 supports this hypothesis; the worst out of all pollutants and has the least available data.

Through the definition provided for the AURN station environment types, we know that air pollutant concentrations exhibit different signatures in different locations. There is the possibility of further subclassifications within these environment types. For example, taking the Urban Traffic environment type, there could be a distinction between the Urban Traffic stations within London and outside of London. Particularly for the approach used within this work, a data-driven model, this can have wide-ranging implications. Suppose it does exist that a subset or single station within the LOOV dataset is unique compared to others, resulting in no similar data present within the training data. In that case, the model presented will fail to replicate the time series measured. This hypothesis is supported as the temporal experiment framework performance is consistent across all the experiments conducted, shown in Figures 2 and 3, with the issue only appearing in the spatial experiments. As a more concrete example of this scenario, the Aston Hill AURN monitoring station is the only monitoring station in the NO_x dataset with no roads nearby, a clearly unique monitoring station in the dataset. This issue is further complicated when considering the results from Section 4.2, where not all feature vector elements have an air pollution monitoring station present, showing clear environment types that have no data available, denoting situations where the model is extrapolating and potentially widely wrong.

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

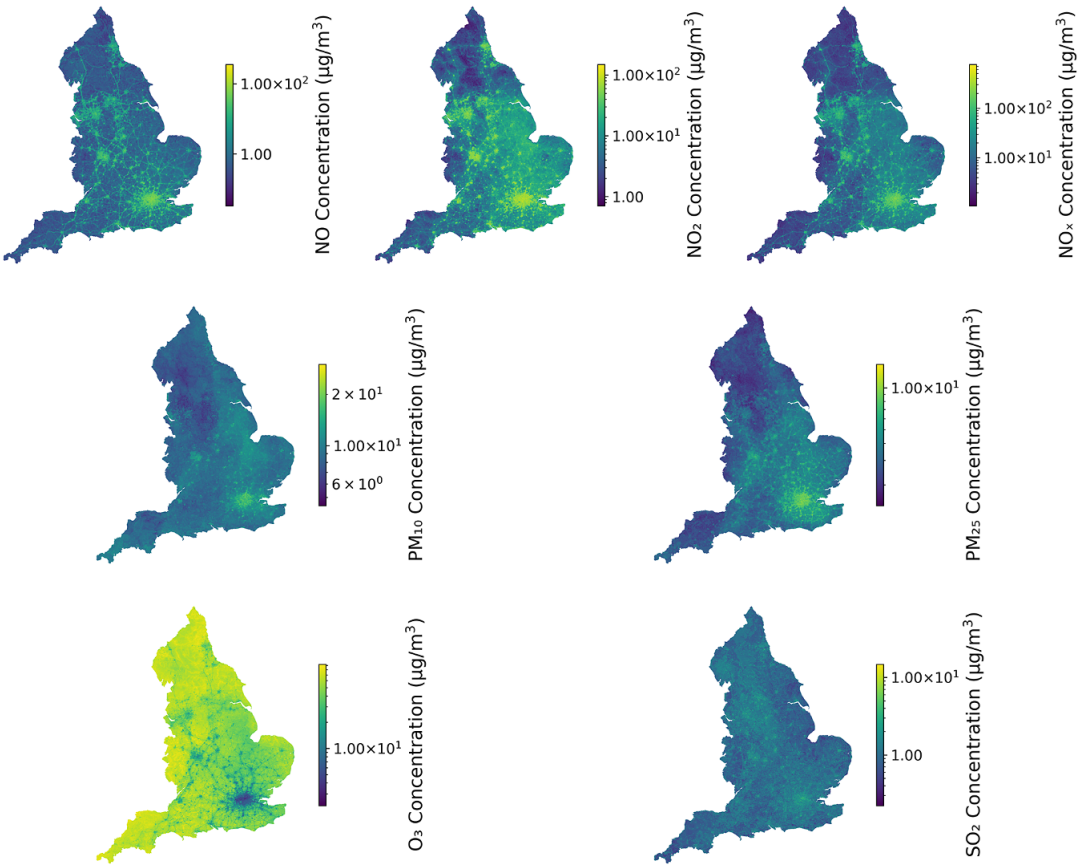


Figure 8. *Full spatial map of England for all pollutants for 8AM on 19/01/2018, chosen arbitrarily as a typical working day away from national holidays in England. Plotted on a log scale to help highlight the differences within regions in the map.*

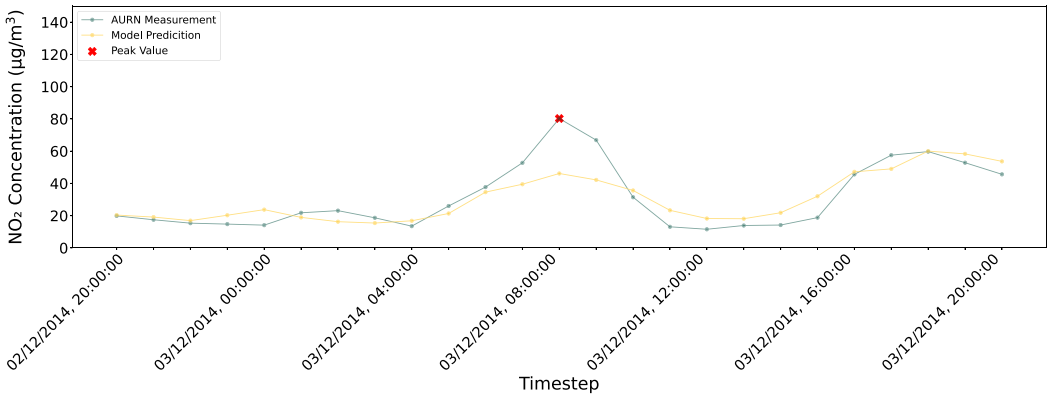
The experiment here provides the basis for using the model to create a complete spatial map of pollution across England and Wales with the framework of synthetic stations. In the case of the grid system used in this study, the framework acts as if 355,827 synthetic stations are present at the centroid of each grid, which gives a point sample measurement of the ambient air pollution concentration at a given time. Figure 8 shows the resulting air pollution concentration map we can create from the model, predicting air pollution at every location across England.

5.4. Prediction of peak values

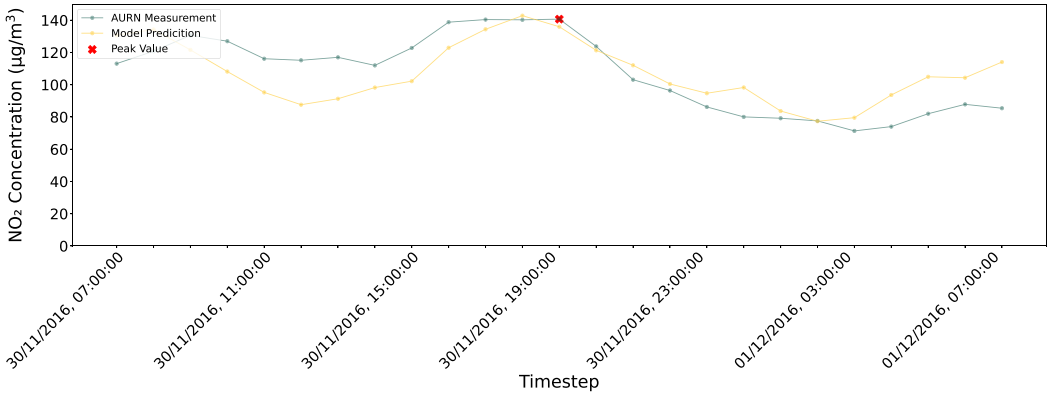
While the R^2 score provides a metric for evaluating the overall performance of the model on the entire dataset, a critical consideration in the context of air pollution concentration estimation is the model's performance during peak concentrations. Given that peak concentrations have the most significant impact on human health and well-being and are the focus of policymakers when designing interventions, it is crucial to assess how well the model performs in these high-concentration scenarios.

We conducted an analysis of the model predictions during peak concentration events observed at each station. Figure 9 illustrates the model's predictions compared to real-world measurements from AURN monitoring stations. Specifically, Figure 9a focuses on the Leominster monitoring station discussed in Section 1. The visual comparison reveals that while the model did not capture the exact magnitude of the peak concentration at the station, it did exhibit an uptick at the correct time. This raises concerns about the model's ability to make high-magnitude predictions. However, further investigation indicates otherwise,

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press



(a) Leominster NO_2.



(b) Stanford-le-Hope Roadside NO_2.

Figure 9. *Prediction of peak values for NO_2 monitoring stations. In (a), it is evident that the model failed to capture the peak concentration for the Leominster monitoring station. However, there is a noticeable uptick in the concentration prediction at the correct time, raising concerns about a consistent underestimation by the model. Conversely, (b) illustrates the peak prediction for the Stanford-le-Hope monitoring station. The model not only captures the peak but also yields a magnitude considerably higher than that for Leominster, offering an initial indication that the model may not be systematically underpredicting concentrations.*

as evidenced by the Stanford-le-Hope Roadside AURN station, which had a high-magnitude prediction of approximately 140 µg/m^3 for its overall peak concentration reading between 2014 and 2018.

This prompts the question of whether the model's prediction for the Leominster peak value was an underprediction or if the peak value itself was an anomalous reading. The percentage difference for the Leominster monitoring station at 03/12/2014 8 AM was 42.5%. In contrast, the mean peak percentage difference across all NO_2 monitoring stations was 22.1% during the Leominster peak, considerably lower than the overall NO_2 average peak distance of 50.72%. The considerable difference in expected peak distance suggests that the Leominster station differed from nearby stations at this time. The performance variation across different pollutants sheds light on a potential issue with the model stemming from the training data, as depicted in Table 5. Notably, O_3 exhibits the best performance in predicting peak concentrations. As outlined in Section 2, meteorological conditions predominantly drive O_3 concentrations. The ERA5 data used in this study stands out as the most robust dataset, featuring temporally and spatially unique data points.

In contrast, other datasets used in the model lack this level of uniqueness, relying on idealized values that may not accurately represent the true variability. For instance, the transport use dataset family follows

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

Table 5. *Average peak concentrations prediction difference*

Pollutant name	Average peak distance percentage (% of μg/m^3)
O_3	32.07
NO_2	50.72
NO_x	63.82
NO	72.32
PM_{25}	84.21
PM_{10}	87.37
SO_2	93.00

Note. The peak percentage difference is calculated according to equation 3. O_3 has the best performance for predicting the peak concentrations across all the monitoring stations, with SO_2 having the worst performance. This ordering presents further evidence that the likely explanation for the model not capturing the peak concentrations is not the model framework itself but rather the input data. SO_2 has by a considerable margin the least amount of data across the air pollutants (Table 1), alongside O_3 being most correlated (Section 4.1) and driven by meteorological phenomena according to the scientific literature (Section 3.2), which given that ERA5 is the highest quality dataset, with unique points spatially and temporally indicates that the difference in data is likely driving the difference in peak concentration estimation performance.

a common temporal distribution, emissions are scaled, and datasets like transport infrastructure lack sufficient variability. These idealized values present an avenue for improving the model by seeking enhanced data representations of phenomena influencing air pollution. Additionally, addressing the potential impact of outlier and anomalous data points could further enhance model performance. The peak distance percentage is defined in Equation 3.

$$\left(\frac{(\text{Measured Peak-Model Prediction})}{\text{Peak Value}}\right) \times 100 \tag{3}$$

5.5. Final model performance summary analysis

Additional performance metrics have been calculated to improve understanding of the final models on the input air pollution concentration data. Whilst the R^2 provides a good indication of the model's performance for a given monitoring station's prediction, the bias and correlation of the prediction can provide further insight into the model's performance. Bias represents the average difference between the monitoring station measurements and the model predictions, providing a metric for the overall tendency of the predictions to be higher or lower than the observations. Correlation quantifies the linear relationship between the monitoring station measurements and the model predictions, reflecting how well the model captures the temporal variations. We use the Pearson Correlation Coefficient to capture this characteristic. MSE quantifies the average squared difference between the monitoring station measurements and the model predictions. It provides a measure of how close the model's predictions are to the actual observations, with larger errors contributing more heavily to the score. The MSE is provided as a further indicator of bias, providing an indication of the effect of more extreme differences between observed and predicted. Table 6 show the mean, max and min values for each of these metrics for all air pollution monitoring stations across each of the air pollutants. The mean, min, and max values for the correlation highlight that the final models can accurately capture the overall trends of the air pollution concentrations across all air pollutants. Within the context of the air pollution concentrations used in this study, as shown in Figure 9, the bias across the air pollutants indicates strong performance across the models, systematically predicting within single-digit values even though the magnitude of concentrations can exceed 100 (μg/m^3).

5.6. Data subsetting

The scalability of the framework has been considered primarily within temporal and spatial resolution dimensions. However, the framework's adaptability extends to different amounts of data, contingent on

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

Table 6. *Mean, max, and minimum values for bias, correlation and MSE for each air pollutant across all air pollution monitoring stations*

Air pollutant	Mean correlation	Max correlation	Min correlation	Mean bias ($\mu g/m^3$)	Max bias ($\mu g/m^3$)	Min bias ($\mu g/m^3$)	Mean MSE $(\mu g/m^3)^2$	Max MSE $(\mu g/m^3)^2$	Min MSE $(\mu g/m^3)^2$
NO_2	0.87	0.92	0.75	−1.12	−0.50	−2.48	74.80	336.94	7.65
O_3	0.89	0.92	0.82	−1.74	−0.70	−2.62	112.37	164.38	47.30
NO_x	0.84	0.90	0.72	−3.72	−0.53	−9.85	1008.38	7587.10	11.02
NO	0.74	0.88	0.45	−2.61	−0.07	−8.03	365.12	2925.59	0.21
PM_{10}	0.74	0.82	0.36	−1.52	−1.16	−2.16	76.41	335.28	36.14
PM_{25}	0.75	0.82	0.66	−1.44	−1.10	−2.39	44.11	73.46	26.51
SO_2	0.58	0.88	0.35	−0.45	−0.14	−1.45	4.85	30.77	0.25

Note. Supplementary Section S3.1 provides the bias, correlation and MSE for each individual monitoring station across each air pollutant.

the availability of specific datasets. Variations in input datasets allow for the development of models tailored to different use cases. For instance, datasets that are only available after historical dates, such as remote sensing, may be excluded when creating models for forecasting purposes. Supplementary Table S18 provides experiment results for this type of model. Additionally, in situations where a location lacks certain datasets, such as traffic estimates for an entire country, models can be built using only the available datasets. Both meteorological and remote sensing datasets have global availability, making them suitable for creating baseline hindcast models for locations not just in England. Supplementary Table S17 showcases the performance of such a model.

Further experimentation delved into assessing the performance of each dataset family, as depicted in Table 7. These findings support the concepts presented in Section 4.1, emphasizing that no single dataset alone can achieve a positive mean LOOV score. In the tables presented in this section and corresponding supplementary section, the training, validation and test scores format follows the same framework used throughout this section, with the additional LOOV summary data included.

6. Research data output

As part of our ongoing efforts, we plan to release an open-source dataset consisting of two components. The first component is the augmented AURN dataset, as illustrated in Figure 7. This dataset includes

Table 7. *Repeat experiments results of Tables 2 and 4 for models trained on individual dataset families (Section 3.2) for NO_2*

Dataset family	Dataset train score	Dataset validation score	Dataset test score	Mean LOOV
Emissions	0.46	0.42	0.42	−0.23
Geographic	0.31	0.25	0.29	−0.46
Meteorological	0.15	0.17	0.14	−0.53
Remote sensing	0.35	0.31	0.36	−0.38
Temporal	0.00	0.03	0.02	−0.64
Transport infrastructure	0.32	0.29	0.29	−0.41
Transport use	0.35	0.23	0.16	−0.47

Note. The framework presented can be used on varying amounts of input data, depending on available data, providing a basic understanding of limitations when moving the model between areas, such as being used to predict countries other than England. Table 7 shows that in the case of England, while individual dataset families can forecast into the future, the performance of estimating missing monitoring stations is limited and requires datasets that cover a wide range of phenomena to achieve the same performance as 4.

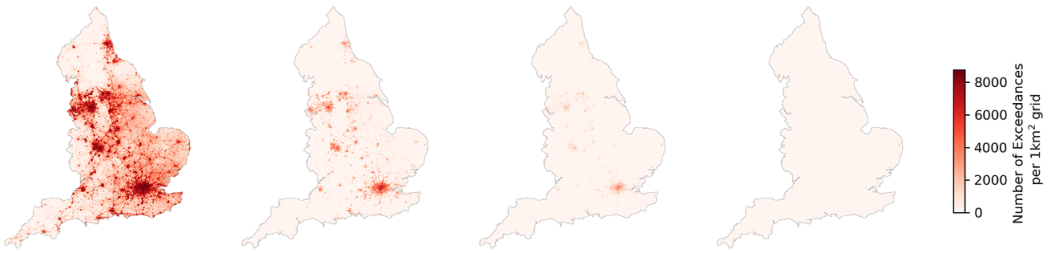
https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

model predictions for air pollution concentrations at all AURN monitoring stations for the period 2014-2018, which were utilized in this study. The second component is a comprehensive air pollution concentration map for England, encompassing each air pollutant for the year 2018. This dataset provides a spatial resolution of 1 km^2 and hourly temporal resolution. We anticipate that this dataset will be of significant interest to various stakeholders, as outlined in Section 1. Moreover, it opens avenues for the research community to explore a diverse range of research topics that were previously constrained, given that current air pollution estimations at this spatial resolution typically operate at the annual temporal level.

The presented dataset opens up diverse research possibilities, with one illustrative example being the examination of air pollution concentration variations across different locations concerning legislation compliance, as discussed in Section 1. Figure 10 showcases a series of heatmaps representing the grids employed in this study. Each grid is color-coded based on the number of times it exceeded specific concentration thresholds in 2018 at an hourly granularity. The maximum possible number of exceedances per grid is 8,760, representing the total hours in 2018.

The thresholds considered include 10 $\mu g/m^3$, aligning with the WHO NO_2 air quality guideline level for the annual temporal period. The second threshold is 25 $\mu g/m^3$, corresponding to the WHO 24-hour aggregate air quality guideline (World Health Organisation, 2021). The third threshold is 40 $\mu g/m^3$, reflecting the UK National Air Quality Guideline annual limit. Lastly, 200 $\mu g/m^3$ represents the UK National Air Quality Guideline hourly target for NO_2 concentrations (King's Printer of Acts of Parliament, 2010). While not all these thresholds directly pertain to hourly concentration legislation, they provide a comprehensive set of benchmarks derived from actual legislation, offering insights into the distribution of air pollutants across England.

The analysis reveals compelling insights into air pollution concentration exceedances across various thresholds. It was found that 99.96% of locations surpassed the 10 $\mu g/m^3$ threshold at least once, 63% exceeded the 25 $\mu g/m^3$ threshold at least once, 26.2% exceeded the 40 $\mu g/m^3$ threshold at least once, and only a single grid exceeded the 200 $\mu g/m^3$ threshold at least once. This analysis serves as a valuable tool to pinpoint locations demanding further investigation into air pollution concentration causes and potential interventions at the local level. For instance, the coordinates at latitude 51.5, longitude −0.15, representing a location exceeding the 200 $\mu g/m^3$ threshold with concentration predictions surpassing 10 for every hour in 2018, underscore the need for targeted attention. Moreover, leveraging the temporal precision of the predictions allows for flexible aggregation to various temporal levels stipulated in United



(a) 10$\mu g/m^3$ threshold. (b) 25$\mu g/m^3$ threshold. (c) 40$\mu g/m^3$ threshold. (d) 200$\mu g/m^3$ threshold.

Figure 10. *Count of times that a grid exceeded the outlined thresholds for NO_2 in 2018. (a) shows the 10 $\mu g/m^3$ threshold where one grid exceeds the threshold for every hour of the year, with 99.6% of grids exceeding the threshold at least once. (b) depicts the counts for the 25 $\mu g/m^3$ where the max count was 8,656 exceedances across the year, with 63% of grids exceeding the threshold at least once. (c) uses a threshold of 40 $\mu g/m^3$ where the max count for exceedances was 8,086 across the year, with 26.2% of grids exceeding the threshold at least once. (d) denotes a 200 $\mu g/m^3$ threshold, where only a single grid exceeded the threshold twice across the year. Latitude 51.5, longitude −0.15 was the location that exceeded the threshold, a central London location with the postcode W1G 6JA.*

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

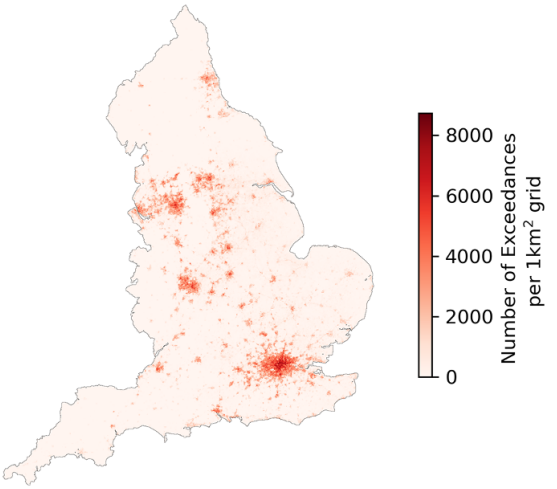


Figure 11. *24-hour mean (UK-AIR, DEFRA, 2023d) exceedance counts example. The threshold used is a mean of 25 μg/m3. As the hourly level is the most common high-resolution temporal level mentioned in air quality legislation, pursuing data at this level allows for a more coarse temporal level to be calculated from the input data, resulting in the dataset providing complete legislation coverage no matter the resolution of interest.*

Kingdom or EU legislation. For instance, aggregating to a 24-hour mean (UK-AIR, DEFRA, 2023d) for each grid facilitates a comprehensive assessment of legislation compliance, as depicted in Figure 11.

7. Discussion

We have released two datasets that hold significant value for the scientific community, policymakers, and the public. The comprehensive spatial dataset offers valuable insights into locations where air pollution concentration data might be non-existent at the hourly temporal level. If one were to acquire the data produced by our model through real-world measurements, the cost would amount to £70B[7] through AURN monitoring stations. We do not imply that our approach is equivalent in value to this figure; rather, we present it to highlight the impracticality of installing monitoring stations on such a large scale, underscoring the need for cost-effective alternatives like our proposed model.

The augmented AURN dataset, generated from this study, provides a complete temporal perspective of all monitoring station measurements across England. This is particularly crucial for compliance assessments related to absolute threshold exceedances, such as NO_2, where a detailed limit of 200μg/m^3 should not be exceeded more than 18 times a year (King's Printer of Acts of Parliament, 2010). Ensuring a complete time series with measurements at each time step is essential when comparing two locations, especially in situations where missing data during peak pollution periods, like NO_x during rush hour, could potentially mask crucial information. This consideration becomes paramount when creating higher temporal resolution statistics through UK AIR[8].

While the presented model has demonstrated its effectiveness in filling missing data from high-quality air quality monitoring stations, its most significant advantages lie in its potential application to low-cost monitoring sensors and citizen science initiatives. The symbiotic relationship between the model and low-cost sensors addresses a core issue present in both approaches. The model's performance is notably impacted

[7] Calculated based on 355,827 monitoring stations at a cost of £198,000 per station, as outlined in Section 1.

[8] UK AIR Statistics Using Incomplete Data, denoted by data capture rate (https://uk-air.defra.gov.uk/data/exceedance?f_exceedence_id=S3&f_year_start=2006&f_year_end=2007&f_group_id=4&f_region_reference_id=1&f_parameter_id=SO2&f_sub_region_id=1&f_output=screen&action=exceedance3&go=Submit).

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

by a lack of data. In locations where the installation of more expensive AURN-style stations might not be deemed a worthwhile investment, low-cost sensors can be strategically deployed to fill data gaps. These sensors, due to their minimal cost, can be implemented in various locations, enhancing spatial coverage.

Conversely, the model can contribute to overcoming the challenges associated with low-cost sensors, which are often less robust than AURN stations. By leveraging the model, missing data points can be backfilled temporally and spatially, ensuring the generation of a more complete dataset. This collaborative approach is particularly valuable in scenarios where the less frequent deployment of AURN stations results in gaps in the feature vector, as discussed in Section 4.2. An additional advantage is the model's ability to handle temporally messy data commonly encountered in citizen science initiatives. Unlike AURN monitoring stations that provide data at regular intervals, citizen science datasets may exhibit irregular time stamps (e.g., 10:14, 10:47, and 11:46). The model enables a more sophisticated estimation of air pollution at specific times (e.g., 10:00, 11:00, 12:00), facilitating further analysis such as legislation compliance or integration with other forecasting systems.

Importantly, the model method offers substantial benefits compared to other approaches for filling missing data, such as interpolation, as it takes into account the nuanced patterns present in air pollution concentration time series datasets.

Although the proposed model demonstrates significant advantages, there is a clear pathway to extracting even more benefits from the model framework, given its inherent scalability in both spatial and temporal dimensions. The extension of the method to make predictions at a minute temporal-level is straightforward, and similarly, increasing the spatial resolution grid size to 100 m^2 is feasible. This scalable approach empowers researchers by providing the desired data without being constrained by limitations in financial resources for monitoring station placement.

Furthermore, the encoding of the temporal aspect into a tabular format facilitates a substantial acceleration through parallelization. Each timestep and grid within the estimation is independent of one another, enabling the simultaneous calculation of all timesteps and grids. This approach yields a significant speedup over traditional forecasting methods, whether machine-learning or physics-based, that rely on lags from previous timesteps.

From a performance standpoint, the capacity to parallelize estimations becomes pivotal when combined with the scalability of the approach. This combination forms the basis for a computationally effective method of estimating air pollution concentrations at a global level. Future work could extend the experiment conducted in Section 5.3, where air pollution concentrations at one station were estimated using data from other monitoring stations, to a study that analyzes the feasibility of estimating air pollution between countries and their respective air pollution monitoring networks. The potential benefit of this analysis is to help reduce inequalities between countries concerning monitoring stations, enabling the design of interventions based on air pollution without the need for high-cost, dedicated monitoring station networks to be implemented by a country's government.

While the datasets employed in this study successfully estimated air pollution concentrations under a variety of conditions, there remains room for improvement in the input feature vectors. Presently, the model does not consider variations associated with specific days, such as distinct travel patterns on bank holidays compared to regular weekdays. Incorporating local knowledge into the model, such as categorizing whether a day is a bank holiday or another national holiday, would enhance the model's understanding of unique circumstances on special days, such as Bonfire Night in the United Kingdom, known to have considerable impact on air pollution concentrations (Adams et al., 2020). Additionally, some feature vectors used in the model will improve over time as technology advances, enabling improved model performance. For example, remote sensing of trace gases over Europe will improve with the Sentinel-4 missions, which are currently scheduled for launch in 2024, on the MTG-s Satellite (ESA, 2024). Sentinel-4 will provide hourly temporal resolution, with a spatial resolution of 8km for much of northern Europe for O_3, NO_2, SO_2, and Aerosol Optical Depth (AOD) (EUMETSAT, 2024), which has been used in the literature to estimate air pollution concentrations (Ranjan et al., 2021). Further it is possible that other model outputs could be used as inputs to the model framework proposed here, such as outputs of chemical transport models as has been used before in the literature (Gariazzo et al., 2020).

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

In addition to incorporating additional knowledge into the model, a thorough analysis of the training data used in the study is crucial to ensure comprehensive coverage of all scenarios, minimizing the need for extrapolation during model estimations. For instance, as discussed in Section 4.2, there are environmental conditions where no air pollution concentration measurements are available. Future work could focus on analyzing missing scenarios in the training data, identifying locations where additional air pollution monitoring stations should be placed. When combined with low-cost sensors, this approach could form the basis for a dynamic mobile monitoring network to identify areas where the model predictions are most uncertain.

In summary, we believe this work holds significant importance for a broad audience, addressing critical challenges outlined in the United Nations (UN) Sustainable Development Goals (SDGs). The work presented empowers decision-makers with high-quality data for crucial indicators (UN SDG 3.9.1, 11.6.2) for essential goals such as Good Health and Well-being (SDG 3) and Sustainable Cities and Communities (SDG 11). The contribution to SDG 3 is evident in reallocating resources from monitoring air pollution to clean air initiatives, providing estimates in all regions, not just those with monitoring stations. Simultaneously, the research contributes to SDG 11 by advancing the understanding of the relationship between urban and rural air pollution.

Open peer review. To view the open peer review materials for this article, please visit http://doi.org/10.1017/eds.2025.9.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/eds.2025.9.

Author contributions. Conceptualisation: LJB, RM; Methodology: LJB, LSN, HJB, BRE, RM; Software: LJB; Validation: LJB, LSN, HJB, BRE, RM; Formal Analysis: LJB, BRE; Investigation, LJB, RM; Resource: LJB, LSN, HJB, BRE, RM; Data Curation: LJB, LSN, BRE; Writing - Original Draft: LJB; Writing - Review and Editing: LJB, LSN, HJB, BRE, RM; Visualisation: LJB; Supervision: RM.

Competing interest. The authors declare none.

Data availability statement. The air pollution concentration dataset is accessed via OpenAir[9]. Land Use data is accessed via UKCEH Land Cover Map 2015.[10]. Sentinel 5P data is accessed via Google Earth Engine Sentinel Catalogue[11]. Meteorology data is accessed via ECMWF ERA5[12]. Transportation Network data is accessed via OpenStreetMaps[13]. Transport Network use data is accessed via the Department of Transport Road Traffic Statistics[14]. The UK Time Use Survey data is available via the UK Data Service[15]. Emissions data is accessed via the National Atmospheric Emissions Inventory[16].

Ethics statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country

Funding statement. Liam Berrisford is supported by a UKRI Studentship at the Center for Doctoral Training for Environmental Intelligence at the University of Exeter.

References

Adams MP, **Tarn MD**, **Sanchez-Marroquin A**, **Porter GC**, **O'Sullivan D**, **Harrison AD**, **Cui Z**, **Vergara-Temprado J**, **Carotenuto F**, **Holden MA**, et al. (2020) A major combustion aerosol event had a negligible impact on the atmospheric ice-nucleating particle population. . *Journal of Geophysical Research: Atmospheres 125*(22), e2020JD032938.

AEA Technology. (2006, August) https://uk-air.defra.gov.uk/assets/documents/reports/cat06/0608141644-386_Purchasing_Guide_for_AQ_Monitoring_Equipment_Version2.pdf. (accessed on August 2022 [Online]).

[9] Clickable Link: OpenAir Homepage (https://github.com/openair-project/openair).

[10] Clickable Link: UK CEH Catalogue (https://catalogue.ceh.ac.uk/documents/6c6c9203-7333-4d96-88ab-78925e7a4e73).

[11] Clickable Link: Google Earth Engine Sentinel 5p Catalogue (https://developers.google.com/earth-engine/datasets/catalog/sentinel-5p).

[12] Clickable Link: ECMWF ERA5 Data Repository (https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5).

[13] Clickable Link: OpenStreetMaps Data Repository (https://www.geofabrik.de/data/download.html).

[14] Clickable Link: Department of Transport Road Traffic Statistics Data Repository (https://roadtraffic.dft.gov.uk/about).

[15] Clickable Link: UK Time Use Survey Data Repository (https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=2000054).

[16] Clickable Link: National Atmospheric Emissions Inventory Homepage (https://naei.beis.gov.uk/).

Amato F, **Querol X**, **Johansson C**, **Nagl C and Alastuey A** (2010) A review on the effectiveness of street sweeping, washing and dust suppressants as urban PM control methods. *Science of the Total Environment 408*(16), 3070–3084.

Arnfield AJ (1990) Street design and urban canyon solar access. *Energy and Buildings 14*(2), 117–131.

Atkinson RW, **Analitis A**, **Samoli E**, **Fuller GW**, **Green DC**, **Mudway IS**, **Anderson HR and Kelly FJ** (2016) Short-term exposure to traffic-related air pollution and daily mortality in London, Uk. *Journal of Exposure Science & Environmental Epidemiology 26*(2), 125–132.

Beirle S, **Platt U**, **Wenig M and Wagner T** (2003) Weekly cycle of no 2 by gome measurements: A signature of anthropogenic sources. *Atmospheric Chemistry and Physics 3*(6), 2225–2232.

Belgian Interregional Environment Agency (2024) Why Are Ozone Concentrations Higher in Rural Areas Than in Cities? Available: https://www.irceline.be/en/documentation/faq/why-are-ozone-concentrations-higher-in-rural-areas-than-in-cities (accessed 05 January 2024. [Online]).

Bennett J (2010) *OpenStreetMap*. Packt Publishing Ltd.

Berrisford LJ, **Ribeiro E and Menezes R** (2022) Estimating ambient air pollution using structural properties of road networks. Preprint, arXiv:2207.14335.

Bloomer BJ, **Stehr JW**, **Piety CA**, **Salawitch RJ and Dickerson RR** (2009) Observed relationships of ozone air pollution with temperature and emissions. *Geophysical Research Letters 36*(9).

Byun DW, **Lacser A**, **Yamartino R and Zannetti P** (2003) *Chapter 10 Eulerian Dispersion Models*.

Carslaw DC and Ropkins K (2012) Openair – An R package for air quality data analysis. *Environmental Modelling and Software 27–28*(0), 52–61.

Castell N, **Dauge FR**, **Schneider P**, **Vogt M**, **Lerner U**, **Fishbain B**, **Broday D and Bartonova A** (2017, February) Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment International 99*, 293–302. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0160412016309989

Chaaban F, **Mezher T and Ouwayjan M** (2004) Options for emissions reduction from power plants: An economic evaluation. *International Journal of Electrical Power & Energy Systems 26*(1), 57–63.

Chen C-C, **Wang Y-R**, **Yeh H-Y**, **Lin T-H**, **Huang C-S and Wu C-F** (2021) Estimating monthly pm2. 5 concentrations from satellite remote sensing data, meteorological variables, and land use data using ensemble statistical modeling and a random forest approach. *Environmental Pollution 291*, 118159.

Cichowicz R, **Wielgosinski G and Fetter W** (2017) Dispersion of atmospheric air pollution in summer and winter season. *Environmental Monitoring and Assessment 189*, 1–10.

Concas F, **Mineraud J**, **Lagerspetz E**, **Varjonen S**, **Liu X**, **Puolamäki K**, **Nurmi P and Tarkoma S** (2021) Lowcost outdoor air quality monitoring and sensor calibration: A survey and critical analysis. *ACM Transactions on Sensor Networks (TOSN) 17*(2), 1–44.

Corbett JJ and Fischbeck P (1997) Emissions from ships. *Science 278*(5339), 823–824.

Crosby T (1994) How to Detect and Handle Outliers.

Davies F, **Middleton D and Bozier K** (2007) Urban air pollution modelling and measurements of boundary layer height. *Atmospheric Environment 41*(19), 4040–4049.

DEFRA, Department for Environment Food and Rural Affairs (2017) The Air Quality Data Validation and Ratification Process. Available: https://uk-air.defra.gov.uk/assets/documents/Data_Validation_and_Ratification_Process_Apr_2017.pdf (accessed 29 November 2023. [Online]).

DEFRA, Department for Environment Food and Rural Affairs (2023) Air Quality Statistics in the UK, 1987 to 2022 - Background. Available: https://www.gov.uk/government/statistics/air-quality-statistics/background (accessed 29 November 2023. [Online]).

Department of Transport, UK Government (2023) Road Traffic Stations – About. Available: https://roadtraffic.dft.gov.uk/about (accessed 29 November 2023. [Online]).

E. Assessment (1992) Guidelines for exposure assessment. *Federal Register 57*(104), 888–938.

Eliassen A (1984) Aspects of lagrangian air pollution modelling. In *Air Pollution Modeling and Its Application III*. Springer, pp. 3–21.

ESA (2024) Introducing MTG. Available: https://www.esa.int/Applications/Observing_the_Earth/Meteorological_missions/meteosat_third_generation/Introducing_MTG (accessed 04 January 2024. [Online]).

EUMETSAT (2024) Sentinel-4. Available: https://www.eumetsat.int/sentinel-4 (accessed 04 January 2024. [Online]).

European Space Agency - Copernicus (2017) Sentinel-5 Precursor Calibration and Validation Plan for the Operational Phase. Available: https://sentinels.copernicus.eu/documents/247904/2474724/Sentinel-5P-Calibration-and-Validation-Plan.pdf (accessed 29 November 2023. [Online]).

European Space Agency - Copernicus (2023) Sentinel, Missions, Sentinel 5P - Orbit. Available: https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-5p/orbit (Accessed 29 November 2023. [Online]).

Eustice G and **Lord Goldsmith of Richmond Park** (2021) Environment Bill. Available: https://bills.parliament.uk/bills/2593 (accessed 29 November 2023. [Online]).

Feng X, **Li Q**, **Zhu Y**, **Wang J**, **Liang H and Xu R** (2014) Formation and dominant factors of haze pollution over Beijing and its peripheral areas in winter. *Atmospheric Pollution Research 5*(3), 528–538.

Finlayson-Pitts BJ and Pitts Jr JN (1986) *Atmospheric Chemistry. Fundamentals and Experimental Techniques*.

Freeman BS, **Taylor G**, **Gharabaghi B and Thé J** (2018) Forecasting air quality time series using deep learning. *Journal of the Air & Waste Management Association 68*(8), 866–886.

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

Gariazzo C, **Carlino G**, **Silibello C**, **Renzi M**, **Finardi S**, **Pepe N**, **Radice P**, **Forastiere F**, **Michelozzi P**, **Viegi G**, et al. (2020) A multi-city air pollution population exposure study: Combined use of chemical-transport and random-forest models with dynamic population data. *Science of the Total Environment 724*, 138102.

Garland J and Derwent R (1979) Destruction at the ground and the diurnal cycle of concentration of ozone and other gases. *Quarterly Journal of the Royal Meteorological Society 105*(443), 169–183.

Gietl JK and Klemm O (2009) Analysis of traffic and meteorology on airborne particulate matter in Münster, Northwest Germany. *Journal of the Air & Waste Management Association 59*(7), 809–818.

Goldberg DL, **Anenberg SC**, **Kerr GH**, **Mohegh A**, **Lu Z and Streets DG** (2021) Tropomi NO2 in the United States: A detailed look at the annual averages, weekly cycles, effects of temperature, and correlation with surface NO2 concentrations. *Earth's Future 9*(4), e2020EF001665.

Gu B, **Zhang L**, **Van Dingenen R**, **Vieno M**, **Van Grinsven HJ**, **Zhang X**, **Zhang S**, **Chen Y**, **Wang S**, **Ren C**, et al. (2021) Abating ammonia is more cost-effective than nitrogen oxides for mitigating $PM_{2.5}$ air pollution. *Science 374*(6568), 758–762.

Haagen-Smit A (1959) Urban air pollution. *Advances in Geophysics 6*, 1–18.

Harishkumar K, **Yogesh K**, **Gad I**, et al. (2020) Forecasting air pollution particulate matter (pm2. 5) using machine learning regression models. *Procedia Computer Science 171*, 2057–2066.

He Q, **Ye T**, **Zhang M and Yuan Y** (2023) Enhancing the reliability of hindcast modeling for air pollution using history-informed machine learning and satellite remote sensing in China. *Atmospheric Environment*, 119994, 2023.

Henze DK, **Hakami A and Seinfeld JH** (2007) Development of the adjoint of geos-chem. . *Atmospheric Chemistry and Physics*, *7*(9), 2413–2433. [Online]. Available: https://acp.copernicus.org/articles/7/2413/2007/

Hersbach H (2016) The ERA5 atmospheric reanalysis. *AGU Fall Meeting Abstracts 2016*, NG33D–01.

Hippler H, **Rahn R and Troe J** (1990) Temperature and pressure dependence of ozone formation rates in the range 1–1000 bar and 90–370 k. *The Journal of Chemical Physics 93*(9), 6560–6569.

Hodson TO (2022) Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development 15*(14), 5481–5487.

Hoek G, **Beelen R**, **De Hoogh K**, **Vienneau D**, **Gulliver J**, **Fischer P and Briggs D** (2008) A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment 42*(33), 7561–7578.

Hoerl AE and Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Jolliet O and Hauschild M (2005) Modeling the influence of intermittent rain events on long-term fate and transport of organic air pollutants. *Environmental Science & Technology 39*(12), 4513–4522.

Jurado X, **Reiminger N**, **Vazquez J and Wemmert C** (2021) On the minimal wind directions required to assess mean annual air pollution concentration based on CFD results. *Sustainable Cities and Society 71*, 102920. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210670721002079

Kang Y, **Aye L**, **Ngo TD and Zhou J** (2022) Performance evaluation of low-cost air quality sensors: A review. *Science of the Total Environment 818*, 151769.

Ke G, **Meng Q**, **Finley T**, **Wang T**, **Chen W**, **Ma W**, **Ye Q and Liu T-Y** (2017) Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems 30*.

King's Printer of Acts of Parliament (2010) The Air Quality Standards Regulations 2010. Available: https://www.legislation.gov.uk/uksi/2010/1001/contents/made (accessed 29 November 2023. [Online]).

Konstantinoudis G, **Padellini T**, **Bennett J**, **Davies B**, **Ezzati M and Blangiardo M** (2021) Long-term exposure to air-pollution and covid-19 mortality in England: A hierarchical spatial analysis. *Environment International 146*, 106316.

Li J, **Zhang H**, **Chao C-Y**, **Chien C-H**, **Wu C-Y**, **Luo CH**, **Chen L-J and Biswas P** (2020) Integrating low-cost air quality sensor networks with fixed and satellite monitoring systems to study ground-level pm2. 5. *Atmospheric Environment 223*, 117293.

Li Q, **Zhang H**, **Jin X**, **Cai X and Song Y** (2022) Mechanism of haze pollution in summer and its difference with winter in the North China plain. *Science of the Total Environment 806*, 150625. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S004896972105703X

Lundberg SM and Lee SI (2017) A unified approach to interpreting model predictions. In Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S and Garnett R (eds.), *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Matthaios VN, **Lawrence J**, **Martins MA**, **Ferguson ST**, **Wolfson JM**, **Harrison RM and Koutrakis P** (2022) Quantifying factors affecting contributions of roadway exhaust and non-exhaust emissions to ambient $PM_{10–2.5}$ and $PM_{2.\ 5–0.2}$ particles. *Science of the Total Environment 835*, 155368.

Meng K, **Xu X**, **Cheng X**, **Xu X**, **Qu X**, **Zhu W**, **Ma C**, **Yang Y and Zhao Y** (2018) Spatio-temporal variations in SO2 and NO2 emissions caused by heating over the Beijing-Tianjin-Hebei region constrained by an adaptive nudging method with OMI data. *Science of the Total Environment 642*, 543–552.

Microsoft. (2023) LightGBM - Parameters. Accessed on: 29/11/2023. [Online]. Available: https://lightgbm.readthedocs.io/en/latest/Parameters.html

Mishra S (2023) Hyper Parameter Optimization - Suggested Parameter Grid ISSUE #695 Microsoft/LIGHTGBM. Available: https://github.com/microsoft/LightGBM/issues/695 (accessed 29 November 2023. [Online]).

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

Molnár VÉ, **Simon E**, **Tóthmérész B**, **Ninsawat S and Szabó S** (2020) Air pollution induced vegetation stress– The air pollution tolerance index as a quick tool for city health evaluation. *Ecological Indicators 113*, 106234.

Nair AT, **Senthilnathan J and Nagendra SS** (2019) Emerging perspectives on VOC emissions from landfill sites: Impact on tropospheric chemistry and local air quality. *Process Safety and Environmental Protection 121*, 143–154.

Ning G, **Wang S**, **Yim SHL**, **Li J**, **Hu Y**, **Shang Z**, **Wang J and Wang J** (2018) Impact of low-pressure systems on winter heavy air pollution in the Northwest Sichuan basin, China. *Atmospheric Chemistry and Physics 18*(18), 13601–13615.

Nowak DJ, **Crane DE and Stevens JC** (2006) Air pollution removal by urban trees and shrubs in the United States. *Urban Forestry & Urban Greening 4*(3–4), 115–123.

Nowak DJ et al. (2002) The effects of urban trees on air quality. *USDA Forest Service*, 96–102.

Office for Health Improvement and Disparities (2022) Air Pollution: Applying All Our Health. Available: https://www.gov.uk/government/publications/air-pollution-applying-all-our-health (accessed 29 November 2023. [Online])

Office for National Statistics (2017) 2011 Census Aggregate Data. [Online]. Available: https://www.ons.gov.uk/census/2011census

Rai AC, **Kumar P**, **Pilla F**, **Skouloudis AN**, **Di Sabatino S**, **Ratti C**, **Yasar A and Rickerby D** (2017, December) End-user perspective of low-cost sensors for outdoor air pollution monitoring. *Science of the Total Environment 607–608*, 691–705. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0048969717316935

Ranjan AK, **Patra AK and Gorai A** (2021) A review on estimation of particulate matter from satellite-based aerosol optical depth: Data, methods, and challenges. *Asia-Pacific Journal of Atmospheric Sciences 57*, 679–699.

Rosofsky A, **Levy JI**, **Zanobetti A**, **Janulewicz P and Fabian MP** (2018) Temporal trends in air pollution exposure inequality in Massachusetts. *Environmental Research 161*, 76–86.

Rowland C, **Morton R**, **Carrasco L**, **McShane G**, **O'neil A and Wood C** (2017) Land cover map 2015 (vector, GB). *NERC Environmental Information Data Centre 10*.

Shi Q and Wu J (2021) Review on sulfur compounds in petroleum and its products: State-of-the-art and perspectives. *Energy & Fuels 35*(18), 14445–14461.

Stasiuk WN and Coffey PE (1974) Rural and urban ozone relationships in New York state. *Journal of the Air Pollution Control Association 24*(6), 564–568.

Su T, **Li Z and Kahn R** (2018) Relationships between the planetary boundary layer height and surface pollutants derived from lidar observations over China: Regional pattern and influencing factors," *Atmospheric Chemistry and Physics 18*(21), 15921–15935.

Sullivan J and Gershuny O (2023) United Kingdom Time Use Survey, 2014–2015. [Online]. Available: https://beta.ukdataservice.ac.uk/datacatalogue/doi/?id=8128#!#1

Tai AP, **Martin MV and Heald CL** (2014) Threat to future global food security from climate change and ozone air pollution. *Nature Climate Change 4*(9), 817–821.

Tao L, **Fairley D**, **Kleeman MJ and Harley RA** (2013) Effects of switching to lower sulfur marine fuel oil on air quality in the San Francisco bay area. *Environmental Science & Technology 47*(18), 10171–10178, 2013.

Tao Q, **Liu F**, **Li Y and Sidorov D** (2019) Air pollution forecasting using a deep learning model based on 1d convnets and bidirectional GRU. *IEEE Access 7*, 76690–76698.

UK National Atmospheric Emissions Inventory (NAEI) (2023) Pollutant Information: Non Methane VOC. Available: https://naei.beis.gov.uk/overview/pollutants?pollutant_id=9#:~:text=NMVOCs%20are%20emitted%20to%20air,over%20a%20large%20spatial%20scale (accessed 29 November 2023. [Online]).

UK-AIR (2019) Modelled Air Quality Data. Available: https://uk-air.defra.gov.uk/data/modelling-data (accessed 29 November 2023. [Online]).

UK-AIR (2023) Site Environment Types – Background Station. Available: https://uk-air.defra.gov.uk/networks/site-types (accessed 29 November 2023. [Online]).

UK-AIR, DEFRA (2021) 'Low-cost' Pollution Sensors - Understanding the Uncertainties. Available: https://uk-air.defra.gov.uk/research/aqeg/pollution-sensors/understanding-uncertainties.php (accessed 29 November 2023. [Online]).

UK-AIR, DEFRA (2023a) Site Information for Chesterfield Loundsley Green(UKA00604) - Defra, UK. Available: https://uk-air.defra.gov.uk/networks/site-info?site_id=CHLG (accessed 29 November 2023. [Online]).

UK-AIR, DEFRA (2023b) Closed AURN Monitoring Sites. Available: https://uk-air.defra.gov.uk/networks/aurn-sites (accessed 29 November 2023. [Online]).

UK-AIR, DEFRA (2023c) Interactive Monitoring Networks Map. Available: https://uk-air.defra.gov.uk/interactive-map. (accessed 29 November 2023. [Online]).

UK-AIR, DEFRA (2023d) FAQ – What Is the Definition of Running Annual Mean and Running 8hr Mean. Available: https://uk-air.defra.gov.uk/air-pollution/faq?question=20. (accessed 29 November 2023. [Online]).

United States Environmental Protection Agency (2023) Ground-level Ozone Basics. Available: https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics (accessed 29 November 2023. [Online]).

Van Roode S, **Ruiz-Aguilar J**, **González-Enrique J and Turias I** (2019) An artificial neural network ensemble approach to generate air pollution maps. *Environmental Monitoring and Assessment 191*, 1–15.

Veefkind JP, **Aben I**, **McMullan K**, **Förster H**, **De Vries J**, **Otter G**, **Claas J**, **Eskes H**, **De Haan J**, **Kleipool Q**, et al. (2012) Tropomi on the esa sentinel-5 precursor: A gmes mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications. *Remote Sensing of Environment 120*, 70–83.

Vitturi M, **Neri A**, **Ongaro TE**, **Savio SL and Boschi E** (2010) Lagrangian modeling of large volcanic particles: Application to vulcanian explosions. *Journal of Geophysical Research: Solid Earth 115*(B8).

https://doi.org/10.1017/eds.2025.9 Published online by Cambridge University Press

Vukovich FM (1979) A note on air quality in high pressure systems. *Atmospheric Environment 13*(2), 255–265

Wallace J, **Corr D and Kanaroglou P** (2010) Topographic and spatial impacts of temperature inversions on air quality using mobile air pollution surveys. *Science of the Total Environment 408*(21), 5086–5096.

Ward JH (1963) Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association 58*(301), 236–244.

Watkins L (1991) *Air Pollution from Road Vehicles*.

We˛glarczyk S (2018) Kernel density estimation and its application. In *ITM Web of Conferences*, vol. *23*. EDP Sciences, p. 00037.

World Health Organisation (2021) What Are the WHO Air Quality Guidelines? Available: https://www.who.int/news-room/feature-stories/detail/what-are-the-who-air-quality-guidelines (accessed 29 November 2023. [Online]).

World Health Organization (2021) *WHO Global Air Quality Guidelines: Particulate Matter (PM2. 5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*. World Health Organization.

Xiang Y, **Zhang T**, **Liu J**, **Lv L**, **Dong Y and Chen Z** (2019) Atmosphere boundary layer height and its effect on air pollutants in Beijing during winter heavy pollution. *Atmospheric Research 215*, 305–316.

Xu X, **Yu X**, **Bao L and Desai AR** (2019) Size distribution of particulate matter in runoff from different leaf surfaces during controlled rainfall processes. *Environmental Pollution 255*, 113234.

Yan F, **Winijkul E**, **Jung S**, **Bond TC and Streets DG** (2011) Global emission projections of particulate matter (PM): I. exhaust emissions from on-road vehicles. *Atmospheric Environment 45*(28), 4830–4844.

Yang L, **Yang J**, **Liu M**, **Sun X**, **Li T**, **Guo Y**, **Hu K**, **Bell ML**, **Cheng Q**, **Kan H**, et al. (2022) Nonlinear effect of air pollution on adult pneumonia hospital visits in the coastal city of Qingdao, China: A time-series analysis. *Environmental Research 209*, 112754.

Yassin MF (2011) Impact of height and shape of building roof on air quality in urban street canyons. *Atmospheric Environment 45*(29), 5220–5229.

Yuan Q, **Guerra HB and Kim Y** (2017) An investigation of the relationships between rainfall conditions and pollutant wash-off from the paved road. *Water 9*(4), 232.

Zhao B, **Wang S**, **Ding D**, **Wu W**, **Chang X**, **Wang J**, **Xing J**, **Jang C**, **Fu JS**, **Zhu Y**, et al. (2019) Nonlinear relationships between air pollutant emissions and pm2. 5-related health impacts in the Beijing-Tianjin-Hebei region. *Science of the Total Environment 661*, 375–385.

Zoogman P, **Liu X**, **Suleiman R**, **Pennington W**, **Flittner D**, **Al-Saadi J**, **Hilton B**, **Nicks D**, **Newchurch M**, **Carr J**, et al. (2017) Tropospheric emissions: Monitoring of pollution (tempo). *Journal of Quantitative Spectroscopy and Radiative Transfer 186*, 17–39.

Zou B, **Wilson JG**, **Zhan FB and Zeng Y** (2009) Air pollution exposure assessment methods utilized in epidemiological studies. *Journal of Environmental Monitoring 11*(3), 475–490.

Cite this article: Berrisford LJ, Neal LS, Buttery HJ, Evans BR and Menezes R (2025). A framework for scalable ambient air pollution concentration estimation. *Environmental Data Science*, 4: e17. doi:10.1017/eds.2025.9