Environmental and Ecological Statistics (2024) 31:1085–1108
https://doi.org/10.1007/s10651-024-00635-5

A comparison of statistical and machine learning models for spatio-temporal prediction of ambient air pollutant concentrations in Scotland

Qiangqiang Zhu[1] · **Duncan Lee**[1] · **Oliver Stoner**[1]

Received: 23 April 2024 / Revised: 4 October 2024 / Accepted: 15 October 2024 /
Published online: 13 November 2024

Abstract

The spatio-temporal prediction of air pollutant concentrations is vital for assessing regulatory compliance and for producing exposure estimates in epidemiological studies. Numerous approaches have been utilised for making such predictions, including land use regression models, additive models, spatio-temporal smoothing models and machine learning prediction algorithms. However, relatively few studies have compared the predictive performance of these models thoroughly, which is one of the novel contributions of this paper. For the specific challenge of predicting monthly average concentrations of NO_2, PM_{10} and $PM_{2.5}$ in Scotland, we find that random forests typically outperform (or are as good as) more traditional statistical prediction approaches. Additionally, we utilise the best performing model to provide a new data resource, namely, predictions of monthly average concentrations (with uncertainty quantification) of the above pollutants on a regular 1 km^2 grid for all of Scotland between 2016 and 2020.

Keywords Ambient air pollutants · Data product · Model comparison · Spatio-temporal prediction

Handling Editor: Luiz Duczmal.

Duncan Lee and Oliver Stoner contributed equally to this work.

✉ Qiangqiang Zhu
q.zhu.1@research.gla.ac.uk

Duncan Lee
Duncan.Lee@glasgow.ac.uk

Oliver Stoner
Oliver.Stoner@glasgow.ac.uk

[1] School of Mathematics and Statistics, University of Glasgow, University Place, Glasgow G12 8QQ, UK

1 Introduction

Air pollution is a complex mixture of different components including nitrogen dioxide (NO_2), ozone (O_3), and particulate matter (PM), the latter being measured by particles ≤ 10 μm (PM_{10}), and ≤ 2.5 μm ($PM_{2.5}$) in aerodynamic diameter. Long-term exposure to these pollutants has been associated with a range of adverse health outcomes, including respiratory diseases (Bălă et al., 2021), cardiovascular diseases (Rajagopalan et al., 2018) and mental ill health (Gu et al., 2020). A recent summary of the evidence is given by Chief Medical Officer (2022). Globally, it is estimated that 99% of the population are exposed to concentrations that exceed the World Health Organisation's guideline limits (World Health Organization, 2021), with the burden disproportionately affecting low- and middle-income countries (https://www.who.int/health-topics/air-pollution). In Scotland, the focus of this study, air pollution management is supported by a robust legislative framework, including the UK Air Quality Strategy in 2007, the Air Quality Standards (Scotland) Regulations in 2010 that enacted the European Union 2008 Ambient Air Quality Directive (2008/50/EC), and the UK Environment Act in 2021. These regulations establish a set of air quality standards, objectives and targets, with a summary being available at https://www.scottishairquality.scot/air-quality/standards.

Comprehensive monitoring of air pollutant concentrations is thus essential for a number of reasons, including the assessment of whether the above targets are being met, as well for producing exposure estimates for epidemiological studies (e.g., Dibben and Clemens, 2015). Ideally, high-resolution air pollution maps should be produced based on data from a dense network of air pollution monitors, but as these monitors are expensive to install and run they are spatially sparse, with only about 100 currently active in Scotland (https://www.scottishairquality.scot/latest/summary). They are predominantly located in the major urban centres of Aberdeen, Dundee, Edinburgh and Glasgow, which leaves vast swathes of the south and north of Scotland with no monitors at all. To overcome this, numerical air quality models such as the Pollution Climate Mapping (PCM) model (https://uk-air.defra.gov.uk/data/pcm-data) have been developed, which provides annual average estimated concentrations on 1 km^2 grid squares with complete spatial coverage of the country. However, these estimates lack the accuracy of the monitoring data, and are only available at a coarse yearly resolution. This prevents, for instance, assessing impacts of seasonal patterns in pollution exposure on human health, or impacts of interactions between air pollution and other risk factors that vary on a seasonal-scale.

The goal of this study is to predict monthly average concentrations of NO_2, PM_{10} and $PM_{2.5}$ between 2016 and 2020 on a 1 km^2 resolution for the whole of Scotland, which will be used to produce exposure estimates for future epidemiological studies. However, as neither the spatially sparse and irregular measured data nor the annual PCM estimates are sufficient for estimating monthly average pollution concentrations across the entire country, this study will utilise a fusion approach for spatial prediction that incorporates both data sources. In essence, this approach "downscales" the PCM output to a higher temporal resolution and calibrates it against the monitoring data.

1.1 Existing methods for predicting air pollution

A range of methodological approaches have been proposed for predicting air pollution concentrations at unmeasured locations, including chemical transport models such as the Community Multiscale Air Quality model (CMAQ, https://www.epa.gov/cmaq) and EMEP4UK (http://www.emep4uk.ceh.ac.uk/), as well as fully data-driven approaches based on statistical or machine learning methods. The latter are the focus here, and one of the simplest is land use regression (LUR). LUR is typically set within a linear modelling framework, and describes the relationship between air pollution levels and related predictors, including satellite data, meteorological factors, land cover, land use, geography, traffic features and population density (Larkin et al., 2023). For example, Novotny et al. (2011) utilised LUR to estimate NO_2 concentrations across the United States in 2006, while Brauer et al. (2016) used it to calibrate satellite-driven and chemical transport models against ground measurements. On a larger scale, Larkin et al. (2017) utilised global LUR models to predict NO_2 concentrations across 58 countries using data from 5,200 monitoring sites.

While LUR models offer advantages such as simplicity, efficiency, stability, and interpretability, they have limitations in their ability to capture non-linear covariate-response relationships, residual spatial autocorrelation, and spatial heterogeneity. To address these limitations, various more complex modelling approaches have been proposed. For example, generalised additive models (GAMs) can capture unknown shaped non-linear relationships between air pollutant concentrations and predictors via smooth functions (Li et al., 2012; Zou et al., 2016; Hou and Xu, 2022; Gao et al., 2023), while hierarchical spatio-temporal models can accommodate the complex spatial and temporal correlations inherent in air pollution data (Banerjee et al., 2014; Cressie and Wikle, 2015; Saez and Barceló, 2022). More recently, flexible machine learning (ML) algorithms have been utilised for air pollution prediction, including random forests (Hu et al., 2017; Zhan et al., 2018; Guo et al., 2021), support vector regression (Hu et al., 2017; Castelli et al., 2020), and deep learning approaches (Eren et al., 2023; Niu et al., 2023). These approaches have exhibited improved predictive accuracy and computational efficiency compared to simpler linear models, due to their ability to uncover complex non-linear patterns in the data. However, these models typically ignore the spatio-temporal correlation in air pollution data, and also require careful hyperparameter tuning and validation schemes to prevent overfitting and achieve robust results (Meyer et al., 2018).

Therefore despite the advantages of the above modelling approaches, there is no clear consensus as to which produces the most accurate predictions. Comparative air pollution prediction studies have been carried out by Chen et al. (2019), Berrocal et al. (2020) and Ren et al. (2020), but their studies focused on different study regions (USA or Europe), temporal resolutions (daily or annual), pollutants (NO_2, O_3, or $PM_{2.5}$), and prediction techniques. Their results have thus been heterogeneous, with Chen et al. (2019) and Ren et al. (2020) finding that machine learning methods generally performed best, while Berrocal et al. (2020) suggested that spatio-temporal smoothing models are preferable. Therefore this study builds on these previous works, by providing a comprehensive comparison of statistical and

machine learning prediction paradigms for multiple air pollutants in a new geographical context. Specifically, we focus on predicting monthly average concentrations of NO_2, $PM_{2.5}$ and PM_{10} in Scotland between 2016 and 2020, and our predictive comparison study includes linear models, additive models, additive models of location, scale, and shape (Rigby and Stasinopoulos, 2005), hierarchical spatio-temporal models, and random forests. Additionally, we provide a new data resource for others to use, namely predictions (with uncertainty quantification) of monthly average concentrations of the above pollutants from the best performing models between 2016 and 2020 at a 1 km^2 resolution for the entire country. Finally, we use the models to provide new insight into: (i) the overall temporal trends and seasonal spatial patterns in air pollution concentrations; (ii) locations where concentrations have reduced the most during the study period; and (iii) the level of uncertainty in the predictions. The data and study region are summarised in the next section, while the set of models compared are outlined in Sect. 3. The predictive model comparison study is described in Sect. 4, while predictions from the best performing model for each pollutant are presented in Sect. 5. Finally, Sect. 6 presents key conclusions and areas for future work.

2 Description of the data and the study

The study region is Scotland, United Kingdom, and we focus on predicting monthly average concentrations (in $\mu g/m^3$) of NO_2, PM_{10} and $PM_{2.5}$ at a 1 km^2 spatial resolution for the 5-year period spanning January 2016–December 2020. The choice of a 1 km^2 spatial resolution at a monthly temporal scale is driven by a number of factors. The first is the goal of the study, which is to produce monthly spatially resolved pollution estimates for use in a subsequent small-area population-level epidemiological study quantifying the long-term effects of air pollution. As the spatial resolution of these small areas is typically greater than 1 km^2, a smaller spatial resolution for the pollution modelling is not necessary. Similarly, as the study will examine the long-term effects of pollution, a finer temporal frequency is not necessary. The second reason for this resolution is the availability and resolution of the predictor variables, which are mostly at the 1 km^2 level at an annual scale. Finally, moving to a finer spatio-temporal scale would be computationally challenging given the resulting number of prediction points.

2.1 Measured pollution concentrations

Daily mean concentrations of the three pollutants were obtained from 106 monitoring sites across Scotland between 1st January 2016 and 31st December 2020 from the Scottish Air Quality website (https://www.scottishairquality.scot/). These monitoring site locations are displayed in Fig. 1, and are colour-coded by the type of local environment in which they reside which comprises kerbside (7 monitors, 6.6% of sites), roadside (78, 73.6%), rural (5, 4.7%), suburban (3, 2.8%), urban background (10, 9.4%) and urban industrial (3, 2.8%). Definitions of these site types can

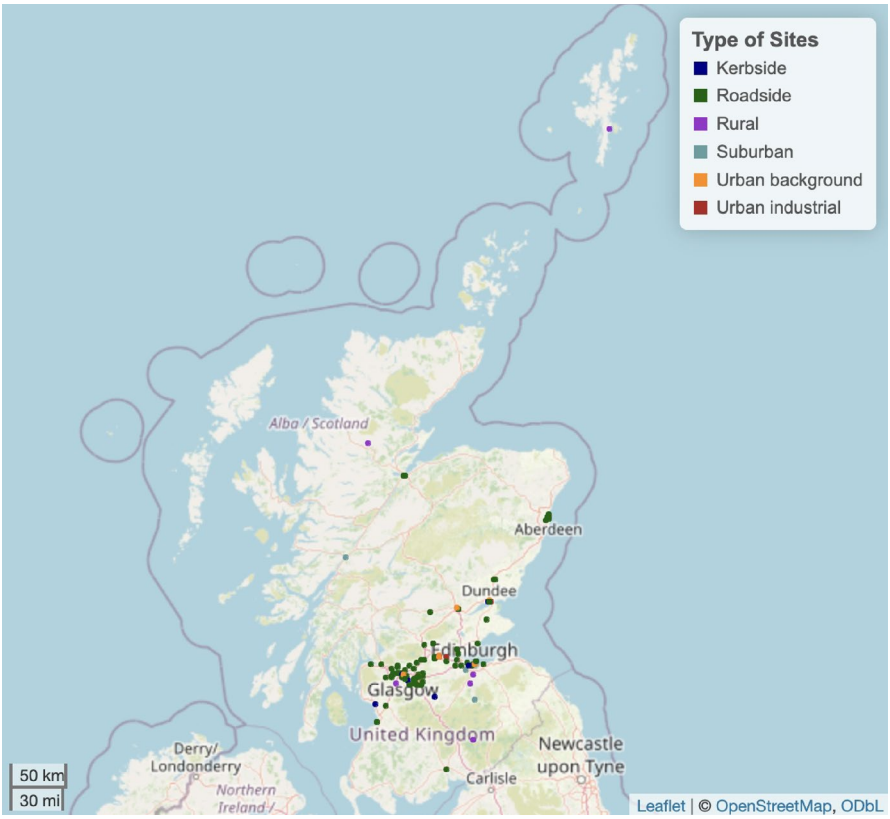


Fig. 1 Map showing the air pollution monitoring sites in Scotland, which are colour-coded according to the type of local environment in which they reside

be found in Section 1.1 of the supplementary material. The majority of the sites are next to main roads (`Kerbside` and `Roadside`), which will naturally have higher concentrations than nearby background (`Rural`, `Suburban`, and `Urban`) locations. Moreover, the sites are very unevenly spread across the country, with the vast majority in the central belt containing Glasgow in the west and Edinburgh in the east, while Northern and Southern Scotland contain hardly any monitoring sites.

These monitors capture different numbers of daily measurements for each month, pollutant and site, which results from equipment malfunction, the fact that not all monitors measured all pollutants, and that not all monitors were in place during the entire 5-year study duration. Details of these data irregularities are given in Section 1.1 of the supplementary material. These daily averages were then aggregated to a monthly temporal resolution by averaging (mean), as long as there was one or more observations recorded within the month. Additionally, we recorded the numbers of days that measurements were available in each month, which will be used to allow for heteroscedasticity in the additive models for location, scale, and shape (see Sect. 3.2).

Figure 2 shows box plots of the temporal trends in these monthly mean concentrations for all three pollutants, with dark blue lines denoting the median values across the set of available monitoring sites. NO_2 concentrations exhibit clear seasonality, which peaks in winter due to increased vehicular use and heating needs, and declines in the summer when increased sunlight converts it to ozone (O_3) (Air Quality Expert Group, 2004). Conversely, temporal trend and seasonality are less clear for PM_{10} and $PM_{2.5}$, although notable spikes are apparent in February and April 2019 that align with two high particulate pollution events in continental Europe (Department for Environment, Food and Rural Affairs, 2020).

2.2 Modelled pollution concentrations

As the monitoring sites are spatially sparse, we also utilise annual average modelled concentrations on a 1 km^2 grid from the Pollution Climate Mapping (PCM, https://uk-air.defra.gov.uk/data/pcm-data) model. These PCM concentrations will be used as annual predictors in the regression models for the monthly monitoring site concentrations, i.e., all 12 months of the same calendar year will have the same value. Thus, the statistical/machine learning models presented in this paper essentially downscale these PCM data to a monthly resolution. These modelled concentrations are generated by integrating various data sources, such as road traffic data,

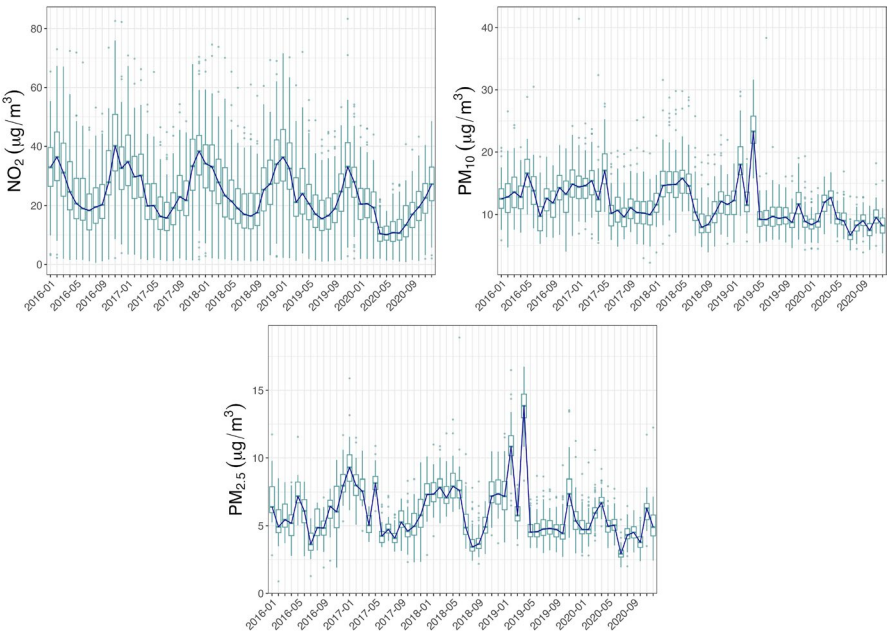


Fig. 2 Monthly temporal trends in the monitored concentrations of NO_2, PM_{10} and $PM_{2.5}$ via box plots and medians (across sites, dark blue lines)

meteorological data and air dispersion model outputs, and the average modelled concentrations over the 5-year study period are presented in Fig. 3.

For NO_2 a discernible urban-rural difference is visible, with clearly elevated concentrations in the four largest cities of Aberdeen, Dundee, Edinburgh and Glasgow. This pattern is primarily due to vehicular emissions and heating combustion, which are more prevalent in urban areas due to their increased population density. In contrast, the spatial distribution of PM_{10} concentrations reveals elevated levels in both urban areas and eastern regions, with the former again being driven by dense populations while the latter is partially driven by the influence of long-range transboundary pollution from continental Europe (Department for Environment Food & Rural Affairs, 2023). Finally, $PM_{2.5}$ concentrations exhibit a similar pattern to PM_{10}, which is unsurprising given that the former is a subset of the latter. Further exploratory analysis of the modelled and monitoring data is presented in Section 1.2 of the supplementary material.

2.3 Predictors of air pollution

We also collected data on a range of other predictors that are likely to be helpful in predicting air pollution concentrations. Firstly, Liu et al. (2022) have illustrated the influence of meteorological factors on pollutant concentrations, and we therefore obtained data on: (i) average temperature (°C); (ii) average relative humidity (%); (iii) total sunshine hours; (iv) total rainfall (mm); (v) average wind speed at 10 ms (m/s); and (vi) average sea level pressure (hPa). These data were obtained from HadUK-Grid (https://www.metoffice.gov.uk) at a monthly 1 km^2 gridded resolution. Secondly, we collected data on the Normalised Difference Vegetation Index (NDVI) at a 1 km^2 resolution to measure land cover, which was obtained from the Terra Moderate Resolution Spectroradiometer (MODIS) Vegetation Indices (MOD13A3) Version 6 data (https://lpdaac.usgs.gov/products/mod13a3v006/).

Thirdly, we collected data on population density and road networks, as both are known to influence local pollution concentrations. For the former we obtained mid-year population estimates from the National Records of Scotland (https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/

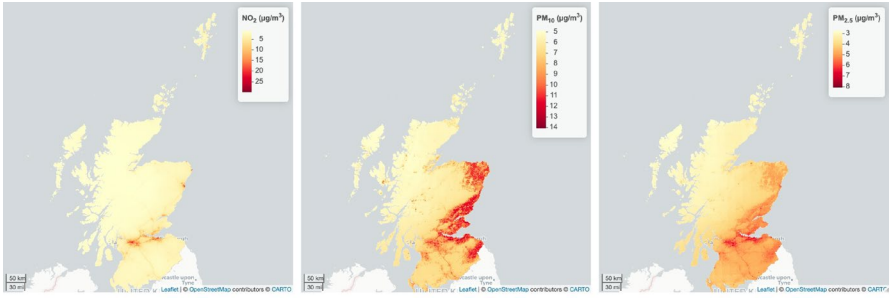


Fig. 3 Spatial distributions of the temporal mean (between 2006 and 2020) of the 1 km^2 estimates of NO_2, PM_{10} and $PM_{2.5}$ from the Pollution Climate Mapping model

population-estimates), which are available at a small areal spatial resolution called Data Zones (there are 6,976 in Scotland). These population density data were spatially re-aligned to our 1 km^2 grid square resolution by area weighted averaging. We also calculated the distance to the nearest major road for each monitoring site and 1 km^2 grid square centroid using OpenStreetMap data (https://download.geofabrik.de/europe/united-kingdom/scotland.html). Finally, we considered a measure of the urban-rural nature of each small-area in Scotland, which was obtained from the 6-fold urban rural classification produced by Rural and Environment Science and Analytical Services Division (2022). Here we simplified this into the following three categories for each monitoring site and 1 km^2 grid square: `Rural` – areas with populations under 9,999; `Urban` – areas with populations between 10,000 and 124,999; and `Large urban` – areas with populations over 125,000.

3 Methods for spatio-temporal pollution prediction

The set of prediction models compared in this study are summarised below, and in each case are fitted to data at locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ for $t \in \{1, \ldots, 60\}$ months, before being used to predict concentrations for those months at unmeasured spatial locations. In what follows, $Y(\mathbf{s}_i, t)$ denotes the monthly average concentration of a single air pollutant (one of NO_2, PM_{10} or $PM_{2.5}$), at location $\mathbf{s}_i$ in month t, while $\mathbf{x}(\mathbf{s}_i, t)$ denotes a row of the design matrix of p predictor variables described in Sect. 2 (including the modelled concentrations).

3.1 Linear models

Normal linear models (LM) are often used to predict air pollution concentrations within an LUR context, and the general model form is given by

$$Y(\mathbf{s}_i, t) \sim \text{ Normal } \left(\alpha + \mathbf{x}(\mathbf{s}_i, t)^\top \boldsymbol{\beta}, \sigma^2\right). \tag{1}$$

Here, α is the intercept term, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is the coefficient vector for the predictor variables, and σ^2 is the error variance. This model assumes that the elements of $\mathbf{x}(\mathbf{s}_i, t)$ are linearly related to the response, and that the *model errors* are independent and identically normally distributed. Details of parameter estimation and prediction, including uncertainty intervals, are given by Montgomery et al. (2021), and follow maximum likelihood methods and standard normal distribution theory.

3.2 Additive models

Additive models (AMs, Hastie and Tibshirani, 1986) extend linear models to allow non-linear predictor effects on the response, where the shape of the effect is determined by the data. The general model form is given by

$$Y(\mathbf{s}_i, t) \sim \text{Normal}\left(\alpha + \mathbf{x}^{(1)}(\mathbf{s}_i, t)^\top \boldsymbol{\beta} + \sum_{j=p_1+1}^{p} f_j\{x_j(\mathbf{s}_i, t)\}, \sigma^2\right), \tag{2}$$

where $\mathbf{x}^{(1)}(\mathbf{s}_i, t) = (x_1(\mathbf{s}_i, t), \ldots, x_{p_1}(\mathbf{s}_i, t))^\top$ is a subset of p_1 predictors from $\mathbf{x}(\mathbf{s}_i, t)$ that are assumed to have linear relationships with the response, while $\{x_j(\mathbf{s}_i, t)\}$ for $j = p_1 + 1, \ldots, p$ are allowed to have non-linear predictor-response relationships. The shapes of these latter relationships $\{f_j(\cdot)\}$ are estimated from the data, using penalised regression splines or related techniques. Parameter estimation is achieved using restricted maximum likelihood estimation (REML), and further details of this and predictions (with uncertainty intervals) using standard normal distribution theory are given by Wood (2017).

A limitation of AMs in this study is that they assume the model errors have a constant variance, where as heteroscedasticity could be likely because the monthly average concentrations are aggregated from different numbers of data points due to the incompleteness of the data (see Sect. 2). Therefore, we also compare additive models of location, scale and shape (AMLSS, Rigby and Stasinopoulos, 2005) here, which extend (2) by allowing the error variance σ^2 to vary over space and time, i.e., become $\sigma^2(\mathbf{s}_i, t)$, as a function of predictors. Specifically, the variance structure is given by

$$\log(\sigma(\mathbf{s}_i, t) - b) = \alpha_\sigma + \mathbf{x}_\sigma(\mathbf{s}_i, t)^\top \boldsymbol{\beta}_\sigma + \sum_{k=q_1+1}^{q} f_k(x_{\sigma_k}(\mathbf{s}_i, t)) + \nu(\mathbf{s}_i, t). \tag{3}$$

Here, the natural log of the error variance $\log(\sigma(\mathbf{s}_i, t) - b)$ can depend linearly and/or non-linearly on predictors $\{x_{\sigma_k}(\mathbf{s}_i, t)\}$, which thus models it in an analogous way to the mean function in Equation (2). In this specification $\nu(\cdot)$ is an independent white-noise process, while b pre-specifies a minimum value for $\sigma(\mathbf{s}_i, t)$ to avoid singularities in the model likelihood (Wood et al., 2016). Details of parameter estimation and prediction can again be found in Wood (2017).

3.3 Hierarchical spatio-temporal models

Spatio-temporal data typically exhibit short-range autocorrelations that the above models do not account for, which is why hierarchical spatio-temporal smoothing models are often used for air pollution prediction (e.g., Sahu et al., 2006). One general class of such models (Bakar and Sahu, 2015) assumes that the data $\{Y(\mathbf{s}_i, t)\}$ are an error-prone estimate of the true underlying spatio-temporal pollution surface $\{O(\mathbf{s}_i, t)\}$, yielding a first-level model of the form:

$$Y(\mathbf{s}_i, t) \mid O(\mathbf{s}_i, t) \sim \text{Normal}\left(O(\mathbf{s}_i, t), \sigma^2\right). \tag{4}$$

One option within this general framework is the Gaussian process (GP) model given by:

$$O(\mathbf{s}_i, t) = \alpha + \mathbf{x}(\mathbf{s}_i, t)^\top \boldsymbol{\beta} + w(\mathbf{s}_i, t). \tag{5}$$

Here, the true unobserved value $O(\mathbf{s}_i, t)$ is modelled by predictors and a latent spatio-temporal process $w(\mathbf{s}_i, t)$. The latter is denoted by $\boldsymbol{w}(t) = (w(\mathbf{s}_1, t), \ldots, w(\mathbf{s}_n, t))^\top$ for all sites at time t, and is assumed to be temporally independent but spatially autocorrelated. Specifically, this model assumes that $\boldsymbol{w}(t) \sim \text{N}(\mathbf{0}, \sigma_w^2 \boldsymbol{S}_w)$, where σ_w^2 controls the amount of spatially smooth variation and $\boldsymbol{S}_w$ is the spatial autocorrelation matrix defined by an exponential autocorrelation function. These GP models thus allow for spatial autocorrelation but assume temporal independence in $\{O(\mathbf{s}_i, t)\}$. An extension is the autoregressive (AR) model, which accounts for temporal dependence by replacing the linear predictor in Equation (5) with:

$$O(\mathbf{s}_i, t) = \rho O(\mathbf{s}_i, t-1) + \mathbf{x}(\mathbf{s}_i, t)^\top \boldsymbol{\beta} + w(\mathbf{s}_i, t). \tag{6}$$

Here ρ is the temporal autocorrelation parameter assumed to be within the interval $(-1, 1)$, such that when $\rho = 0$ the AR model reduces to the GP model. These two spatio-temporal smoothing models are chosen for our study from the myriad of other spatio-temporal models that have been developed because of the availability of software for implementation and their previous use in an air pollution context (Mukhopadhyay and Sahu, 2017). Both models are set within a Bayesian framework with inference using Markov chain Monte Carlo (MCMC) simulation, and they can be implemented in `R` using `spTimer` (Bakar and Sahu, 2015), where further details about prediction and uncertainty intervals are available. Specifically, predictions are based on the posterior median from the MCMC samples, while 95% prediction intervals are obtained as the (2.5%, 97.5%) percentiles of these samples.

3.4 Random forests

Random forests (RF, Breiman 2001) are one of the most popular general-purpose prediction algorithms, and have been used in an air pollution context by Guo et al. (2021). They are based on a bootstrap resampling strategy, where n_{tree} copies of the data (of the same size) are created by randomly resampling the data with replacement. A decision tree model is then fitted to each bootstrapped sample through recursive binary partitioning of the predictor space, resulting in a tree-like structure with a root at the top and a set of branches and nodes that split the data into subsets. The nodes at the bottom of the tree are known as leaves, with each containing a subset of the data with similar values. The number of leaves determines the complexity of the tree, and the tree stops growing when making a further split would either reduce the minimal node size below a threshold *min.size*, or increase the number of levels in the tree above *max.depth*. The set of trees constructed in this manner tend to be correlated, so to reduce between tree correlation Breiman (2001) proposed only considering a subset of m_{try} predictors for use in each split of each decision tree.

The final random forest consists of the n_{tree} decision tree models applied to the bootstrapped data sets, and predictions are obtained by averaging the predictions from the n_{tree} fitted decision tree models. However unlike classical statistical models, random forests generally only provide point predictions, and do not quantify uncertainty in these predictions. To overcome this, quantile regression forests (QRF)

proposed by Meinshausen and Ridgeway (2006) can be used within the RF framework to provide 95% prediction intervals, which are based on the lower and upper quantiles of the empirical conditional distribution based on the value of all observations in the leaves. Further details of their implementation using the `ranger` package in `R` are given by Wright and Ziegler (2017).

4 Study 1 – Comparison of the different models' predictive performance

This section presents our predictive comparison study, including a description of the specific models compared, the study design and finally the results.

4.1 Specific models compared

Prior to modelling both the monitoring and modelled pollution concentrations are log transformed, because they are non-negative quantities with right-skew distributions. However, in what follows all predictions, uncertainty intervals, and model evaluation metrics are computed on the original scale for interpretability. The Gaussian assumption made by most of the above models means that all point predictions are backtransformed by $\exp(\mu(\mathbf{s}_i, t) + \frac{1}{2}\sigma^2)$, because it corresponds to the expectation of a log-normal distribution.

All of the general classes of models outlined in Sect. 3 are compared in this study, and all of the predictors outlined in Sect. 2 are included in each model. Two specific linear models (see Sect. 3.1) are considered in this study, which exhibit different temporal trends. The first, LM_{cs}, assumes the same monthly seasonal pattern occurs for each of the five years, by including factor variables for `Year` and `Month` to the model. The second, LM_{vs}, allows the monthly seasonality to vary by year, by including a factor variable with one level for each of the 60 months of the study. The effect of including spatial coordinates (longitude and latitude) as linear trends was assessed, but it did not improve the predictive performance. Two additive models (see Sect. 3.2) are also compared in this study, with the model denoted by AM allowing: (i) all continuous predictors to have non-linear relationships with the response via univariate *P*-splines if appropriate; and (ii) a Gaussian process (GP) smooth spline for the overall temporal trend and a cyclic *P*-spline for seasonality. These latter elements allow for different seasonal patterns for each year, and hence directly extend LM_{vs}. The second model denoted AM_{sp} additionally includes longitude and latitude as a bivariate smooth term to allow for a non-linear spatial trend. These two additive models are also extended to allow for heteroscedasticity, by allowing the error variance to depend non-linearly (via a *P*-spline) on the number of days with missing records for each month. These two AMLSS variants are denoted by AMLSS and AMLSS_{sp} respectively.

The two hierarchical Bayesian spatio-temporal models outlined in Sect. 3.3 are also assessed here, which are the Gaussian process model denoted by SP_{gp} and the autoregressive model denoted by SP_{ar}. In both cases inference is based on four

parallel Markov chains that are burnt-in for 2,000 iterations before being used to generate a further 8,000 samples each. For each model convergence is assessed using the Gelman-Rubin diagnostic (Gelman and Rubin, 1992). Finally, random forests (Sect. 3.4) are applied to the data on both the original (denoted RF_{oc}) and log scales (denoted RF_{lc}). The former is included because random forests do not make distributional assumptions and hence can have non-Gaussian residuals, and the latter to ensure a fair comparison with the other models included in this study that are applied on the log pollutant scale. Additionally, we re-fit both random forest models including longitude and latitude as additional features (denoted here by RF_{oc_sp} and RF_{lc_sp}), because this allows non-linear spatial trends and heterogeneity to be captured by the model.

4.2 Study design

The out-of-sample predictive performance of each model for each pollutant is assessed using a spatial validation experiment. For each air pollutant we split the monitoring sites measuring that pollutant into an 80% training set and a 20% test set. This process is repeated 10 times to prevent a single training-test split from adversely affecting the outcomes. All models are fitted to each training set, before being used to make out-of-sample predictions for the corresponding test set. The predictive performance metrics outlined below are then computed, and are averaged over the 10 training and test splits in the results that follow. As the random forests contain a number of tuning parameters, we choose the optimal combination by applying a 10-fold cross validation procedure to each training set. Specifically, the training set is further split at random into 10 folds, and the model is fitted to nine of these folds and used to predict the tenth for each tuning parameter combination. This process is repeated 10 times, and the optimal tuning parameter combination is the one that minimises the root mean square prediction error (see below). In all, 480 combinations of tuning parameters are considered, which includes all combinations of $\langle n_{tree} = \{100, 200, 500, 1000\}, m_{try} = \{4, 6, 8, 10\}, min.size = \{1, 3, 5, 10, 15\}, max.depth = \{1, 5, 10, 15, 20, 30\}\rangle$ The optimal values of these tuning parameters chosen by the above approach are presented in Section 2 of the supplementary material. Once the tuning parameters have been selected the random forest is re-fitted to the entire training set with these values, and is used to make predictions in the test set. Thus, for all models only the training set is used to optimise any tuning parameters and fit the models, meaning that all of the predictions made for the test set are completely out of sample.

The predictive accuracy and precision of each model are compared using the following metrics, and in what follows $(Y(\mathbf{s}_i, t), \hat{Y}(\mathbf{s}_i, t))$ respectively denote the observation and prediction at location $\mathbf{s}_i$ and month t, while all averages are taken over all T observations in the test set.

Bias $= \frac{1}{T}\sum(\hat{Y}(\boldsymbol{s}_i, t) - Y(\boldsymbol{s}_i, t))$, which is the mean value of the prediction errors and should be close to zero.

RMSE $= \sqrt{\frac{1}{T}\sum(Y(\boldsymbol{s}_i, t) - \hat{Y}(\boldsymbol{s}_i, t))^2}$, which measures the overall size of the prediction error and should be as small as possible.

MAE = Median $\{|Y(\boldsymbol{s}_i, t) - \hat{Y}(\boldsymbol{s}_i, t)|\}$, which again measures the overall size of the prediction error and should be as small as possible.
CVG: Coverage of the 95% prediction intervals, i.e., the proportion of the 95% prediction intervals for $Y(\mathbf{s}_i, t)$ that contain the true value.
AIW: The mean width of the 95% prediction intervals for $Y(\mathbf{s}_i, t)$.

The RMSE and MAE metrics both quantify the accuracy of the point-predictions, with the RMSE being more sensitive to larger errors than the MAE, while bias quantifies any systematic over or under-prediction. Predictive uncertainty is characterised by CVG and AIW, with the former being close to 0.95 for appropriate uncertainty quantification. If CVG is much more than 0.95 then the intervals are too wide (the model is under-confident), while a value much less than 0.95 means the intervals are too narrow (the model is overly-confident). For two models with similar CVG, the one with the lower AIW produces the most precise predictions. This situation can occur when a model is both more accurate in terms of its point estimate and has an appropriate coverage level.

4.3 Results of the study

The results of this predictive model comparison study are displayed in Table 1, which presents the model comparison metrics for each pollutant and model. For NO_2, the random forest model applied on the original scale (RF_{oc}) appears to be the best performing model, having the lowest RMSE (7.70), second lowest bias in absolute terms (0.32), and a coverage probability close to the nominal 95% (94.97%) level. In contrast, the models with spatial smoothing components (AM_{sp}, $AMLSS_{sp}$, SP_{gp} and SP_{ar}) typically perform worst in terms of RMSE and MAE, suggesting that such smoothness is not appropriate for this pollutant. For PM_{10}, LM_{vs} is the best performing model because it exhibits the lowest RMSE and MAE values and has close to the nominal coverage level (93.81%), although in common with the results for NO_2 the differences between the best performing models are not large. Finally, for $PM_{2.5}$ the set of four random forest models perform best, having the lowest RMSE and MAE values and appropriate uncertainty quantification. Of these RF_{lc} performs best as it has a slightly lower MAE value. Section 2 of the supplementary material presents additional results from this predictive assessment study, including scatter plots of the observed versus the predicted concentrations for each model and pollutant. In addition, line plots showing the temporal variation in the RMSEs are presented illustrating how model performance varies over the 60-month study duration.

5 Study 2 – Understanding spatio-temporal patterns in monthly air pollution concentrations across Scotland

This section summarises the pollutant predictions and their uncertainties from the best performing models identified in Sect. 4.3, which are: NO_2 – random forest on the original scale (RF_{oc}); PM_{10} – linear model with varying seasonality (LM_{vs}); and

Table 1 Results of the predictive model comparison study for each of the three pollutants. The models and metrics are summarised in Sects. 4.1 and 4.2

Pollutants	Models	RMSE	MAE	Bias	Coverage	AIW
NO_2	LM_{cs}	7.97	4.23	0.78	92.85%	29.51
	LM_{vs}	7.82	4.16	0.75	92.28%	28.42
	AM	9.71	4.94	0.25	84.02%	25.37
	AM_{sp}	12.29	5.92	1.51	68.64%	23.48
	AMLSS	9.55	4.79	−0.44	82.87%	24.80
	$AMLSS_{sp}$	12.31	5.84	1.06	67.26%	22.80
	SP_{gp}	12.14	6.01	1.45	99.46%	98.23
	SP_{ar}	12.15	5.99	1.37	97.98%	134.49
	RF_{oc}	7.70	4.36	0.32	94.97%	32.26
	RF_{lc}	7.98	4.19	−0.82	95.02%	32.67
	RF_{oc_sp}	7.91	4.49	0.38	95.27%	33.52
	RF_{lc_sp}	8.15	4.30	−0.85	95.20%	34.03
PM_{10}	LM_{cs}	3.09	1.71	0.08	94.04%	10.55
	LM_{vs}	2.73	1.44	0.15	93.81%	9.32
	AM	3.01	1.57	0.19	89.44%	9.02
	AM_{sp}	3.81	1.74	0.06	83.21%	8.53
	AMLSS	2.99	1.56	−0.03	88.37%	8.74
	$AMLSS_{sp}$	4.01	1.79	−0.12	80.94%	8.22
	SP_{gp}	2.75	1.47	−0.11	99.86%	25.30
	SP_{ar}	2.80	1.47	−0.17	99.93%	35.29
	RF_{oc}	2.78	1.52	0.19	93.89%	10.99
	RF_{lc}	2.76	1.49	0.04	94.11%	11.00
	RF_{oc_sp}	2.77	1.51	0.22	93.99%	11.17
	RF_{lc_sp}	2.75	1.47	0.04	94.07%	11.16
$PM_{2.5}$	LM_{cs}	1.56	0.81	−0.00	94.19%	5.30
	LM_{vs}	1.18	0.57	0.02	94.11%	4.12
	AM	1.30	0.66	0.08	90.79%	3.95
	AM_{sp}	1.27	0.69	0.12	89.66%	3.70
	AMLSS	1.31	0.62	−0.02	88.24%	3.69
	$AMLSS_{sp}$	1.30	0.67	0.04	85.24%	3.41
	SP_{gp}	1.15	0.54	−0.12	99.98%	24.43
	SP_{ar}	1.20	0.61	0.02	99.49%	11.54
	RF_{oc}	1.15	0.56	0.08	94.67%	4.69
	RF_{lc}	1.15	0.54	0.03	94.58%	4.75
	RF_{oc_sp}	1.15	0.57	0.09	94.81%	4.74
	RF_{lc_sp}	1.15	0.55	0.03	94.88%	4.85

$PM_{2.5}$ – random forest on the log scale (RF_{lc}). These models are re-fitted to the entire data, and are subsequently used to produce predictions and 95% prediction intervals at a monthly temporal resolution between January 2016 and December 2020 on a regular 1 km^2 grid across all of Scotland. These predictions are provided as a data product for others to use, and are available to download from https://github.

com/Qiangqiang-Zhu/air-pollution-prediction. In this section our analysis focuses on the three motivating questions outlined in Sect. 1, namely the: (i) overall spatial and temporal trends; (ii) locations where pollution concentrations have reduced the most; and (iii) the precision of the predictions. Additional results summarising the relative importance of the individual predictors are presented in Section 3 of the supplementary material.

5.1 Spatio-temporal pollution trends

Figure 4 presents the seasonal patterns in the average (over all 1 km^2 grid cells in Scotland) predictions of NO_2 (top left), PM_{10} (top right) and $PM_{2.5}$ (bottom) between 2016 and 2020, where each year is represented by a separate line. For NO_2, a clear consistent seasonal pattern is evident, where concentrations are approximately twice as high on average in the winter compared to the summer, with a minimum around July and a maximum around December-January. The plot shows the spatially averaged concentrations over all 1 km^2 grid squares in Scotland, which as they are mostly rural explains the low average concentrations. Across the 5-year study period NO_2 concentrations exhibit a general downward trend, which appears to be particularly pronounced in 2020. The latter is likely to be due in part to the impact of national lockdowns during the Covid-19 pandemic, which resulted in reduced vehicular and industrial emissions. However, Covid-19 lockdowns could only explain part of the

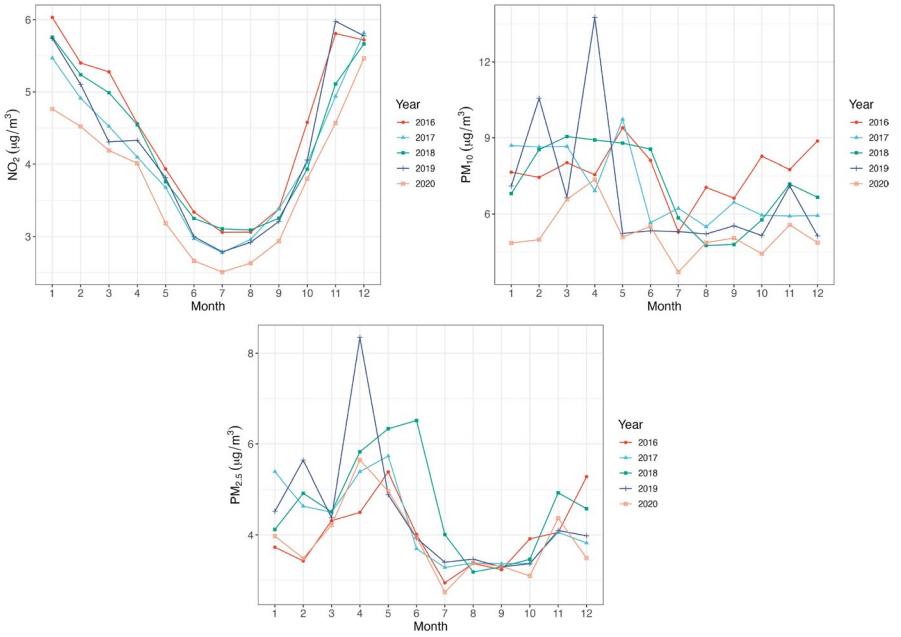
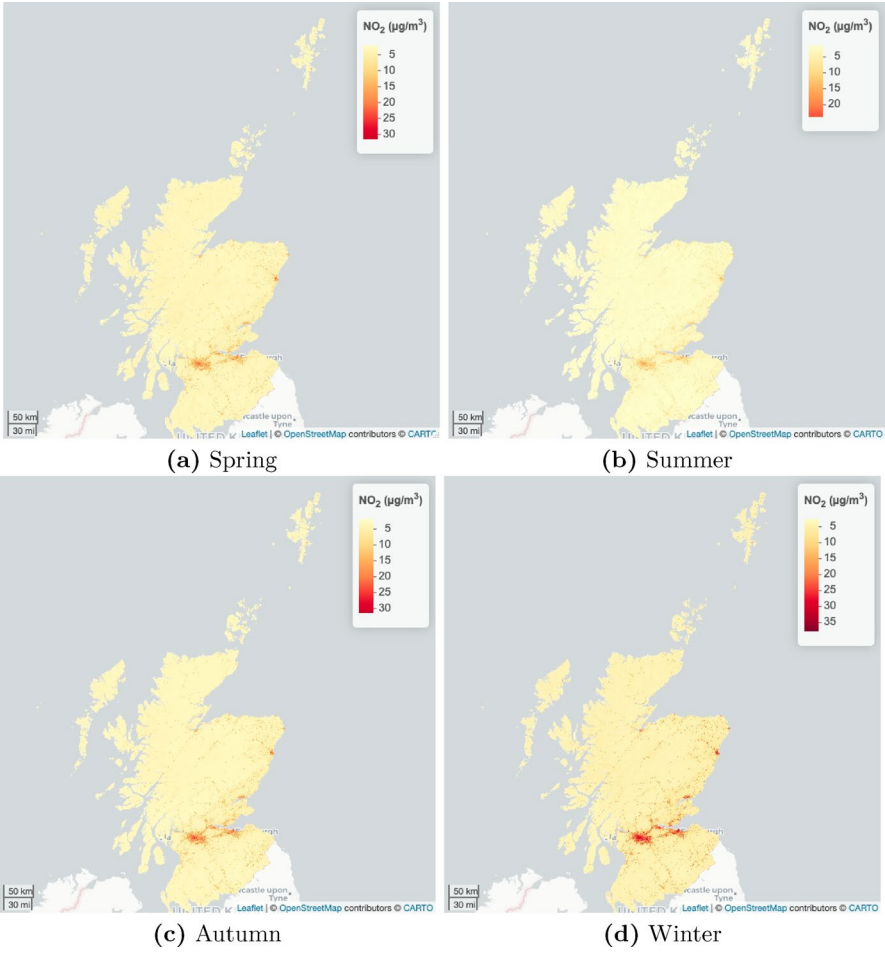


Fig. 4 Line plots of the average (over all 1 km^2 grid cells in Scotland) monthly predicted concentrations between 2016 and 2020. Top left – NO_2; top right – PM_{10}; and bottom – $PM_{2.5}$

reduction in NO_2 concentrations, because this reduction is evident in January 2020 while the lockdowns only started in March 2020. For PM_{10} and $PM_{2.5}$ the concentrations usually reach their lowest levels around July each year, although there is some variation in this from year to year. In addition, the pollutant levels in 2020 are generally lower than most previous years, which may also be partly a result of the Covid-19 lockdowns. For these pollutants Fig. 4 is dominated by the two high pollution events in February and April 2019, which were discussed in Sect. 2.1.

As a result of this seasonality in the predictions, Figs. 5, 6 display spatial maps of the predicted seasonal mean concentrations for NO_2 (see Fig. 5) and $PM_{2.5}$ (see Fig. 6) over the period 2016–2020. Here, Spring – {March, April, May}; Summer – {June, July, August}; Autumn – {September, October, November}; and Winter



(a) Spring (b) Summer

(c) Autumn (d) Winter

Fig. 5 Spatial maps of the predicted seasonal mean NO_2 concentrations over the period 2016–2020. Here, **a** Spring – { March, April, May }; **b** Summer – { June, July, August }; **c** Autumn – { September, October, November }; and **d** Winter – { December, January, February }

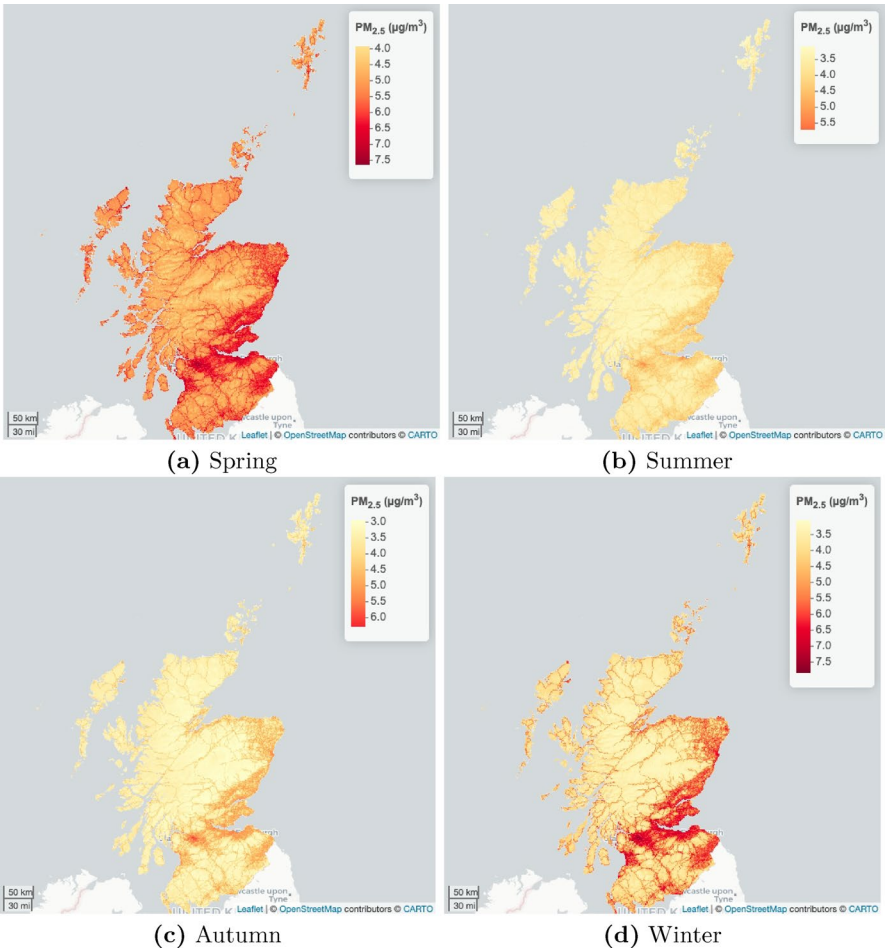


Fig. 6 Spatial maps of the predicted seasonal mean $PM_{2.5}$ concentrations over the period 2016–2020. Here, **a** Spring – { March, April, May }; **b** Summer – { June, July, August }; **c** Autumn – { September, October, November }; and **d** Winter – { December, January, February }

– {December, January, February}. For brevity, the seasonal maps for PM_{10} are presented in Section 3 of the supplementary material. Figure 5 shows that as expected there are higher mean NO_2 concentrations across Scotland in winter compared to summer, which re-enforces the results from Fig. 4 above. Urban areas appear to exhibit the biggest seasonal changes, with increases from around 20 $\mu g/m^3$ in summer to 30 $\mu g/m^3$ in winter. In contrast, rural areas exhibit minimal seasonal variation, with most grid squares changing by less than 3 $\mu g/m^3$ on average between summer and winter.

For $PM_{2.5}$ (see Fig. 6) the spatial patterns are again fairly consistent for all four seasons, with the highest concentrations in the urban centres and along the east coast (partially driven by trans-boundary pollution). The figure suggests that the average

concentrations are higher in spring than in the other seasons, but this is likely in part to be influenced by the high air pollution episode in April 2019. An additional reason for a spring peak in concentrations is the elevated nitrate concentrations from agricultural operations across the UK and continental Europe, causing higher average particulate matter levels than in winter (Department for Environment Food & Rural Affairs, 2023) months. Finally, $PM_{2.5}$ concentrations show the least regional variation in summer, with a spatial standard deviation of 0.354 $\mu g/m^3$ compared to 0.924 $\mu g/m^3$ in winter. Thus, the spatial inequality in pollution exposure is typically highest when concentrations are highest, which is in winter for NO_2 and in winter and spring for $PM_{2.5}$.

5.2 Long-term changes in concentrations

We examine how average concentrations have increased or decreased over the whole study period in different parts of Scotland, by regressing the yearly average concentrations at each 1 km^2 grid square separately against calendar year (scaled to $t = 1, \ldots, 5$). Figure 7 presents the estimated regression slopes from these models for each 1 km^2 grid square. It shows that NO_2 concentrations have typically reduced by between 1 $\mu g/m^3$ and 3 $\mu g/m^3$ in urban areas, but have remained largely constant or even increased slightly by 1 $\mu g/m^3$ in the north Highlands. In contrast, PM_{10} and $PM_{2.5}$ concentrations have almost universally reduced across Scotland over the five year period, with reductions of between 0.1 $\mu g/m^3$ and 1.0 $\mu g/m^3$ for PM_{10} and up to -0.4 $\mu g/m^3$ for $PM_{2.5}$. These reductions are typically largest in the urban areas and on the eastern coast, with the former possibly being due to the fact that urban concentrations are highest and hence have greater scope for reductions.

5.3 Predictive uncertainty quantification

Figure 8 presents maps of the average widths of the monthly 95% prediction intervals for each of the three pollutants, to showcase the levels of uncertainty in our model predictions. It shows that predictive uncertainty is highest where concentrations are highest for both NO_2 and PM_{10}, which is likely to be due to the approximate

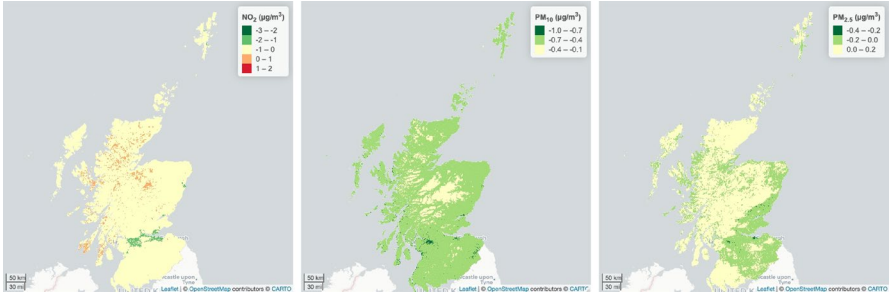


Fig. 7 Maps of the estimated slopes of the temporal trends obtained from regressing the annual mean concentrations of each pollutant against a normalised calendar year variable for NO_2, PM_{10} and $PM_{2.5}$

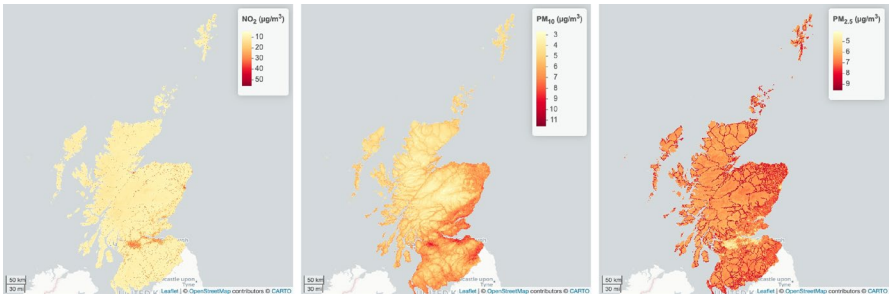


Fig. 8 Spatial distributions of the average widths of the monthly 95% prediction intervals over the period 2016–2020 for NO_2, PM_{10} and $PM_{2.5}$

log-normality of the data that causes the variance to increase with the mean. In contrast, predictive uncertainty for $PM_{2.5}$ is generally larger in the rural areas where concentrations are lower but the monitoring sites are very sparse. In this respect the spatial sparseness of the data is more dominant for increasing predictive uncertainty than the average level of pollutant concentrations, which is probably because the latter has a relatively small variation across the country (see Fig. 6).

6 Discussion

This paper has presented a new study comparing statistical and machine learning prediction methodologies in a Scottish air pollution context, and has provided a new data resource for others to use comprising monthly average predictions and 95% prediction intervals of NO_2, PM_{10} and $PM_{2.5}$ at a 1 km^2 resolution for all of Scotland between 2016 and 2020. We have compared normal linear models that are commonly used in a land use regression context against more complex approaches, with the latter collectively allowing for nonlinear predictor-response relationships and spatio-temporal autocorrelations. These models were used in a temporal downscaling and data-fusion context (Berrocal et al., 2010), because they integrated measured and modelled pollutant concentration data, climate model outputs and geospatial inputs (such as the road network) with different spatio-temporal resolutions.

Our results show that, overall, random forests have the best prediction performance, having optimal values of RMSE for NO_2 and $PM_{2.5}$ and near optimal values for PM_{10}. Linear models with temporally varying seasonality also yield strong predictive capabilities, with RMSE and MAE values that are similar to those from the random forests. These methods also provide appropriate uncertainty quantification, as their 95% prediction intervals are close to the nominal coverage levels. In contrast, the models with spatio-temporal smoothing components give comparatively poor results, either exhibiting much larger RMSE/MAE values of having inappropriate coverage probabilities. This is particularly evident for NO_2, which is highly traffic dependent and hence is more localised spatially with very short-range spatial autocorrelations. The poor general performance of spatial smoothing models could

also be due to the availability of informative spatially-structured predictors, whose inclusion in the models will have made the residuals close to independent and hence the spatial smoothing components largely irrelevant. Moreover, given that the monitoring sites are sparsely and unevenly distributed across Scotland, the quality of the predictions based on geographical location from spatial smoothing terms are not based on a large amount of spatial information, leading to their poorer performance.

Comparing our results to those from existing studies is somewhat difficult given the differences in the study designs and regions, which include the temporal scale of measurements, the volume of data available, the choice of validation strategies, and the average pollutant concentrations. For example, Chen et al. (2019) compared various regression and machine learning techniques for predicting $PM_{2.5}$ and NO_2 concentrations in Europe, and reported cross validation RMSEs of around 3 $\mu g/m^3$ for $PM_{2.5}$ and 9 $\mu g/m^3$ for NO_2 from both random forest and linear regression models. Additionally, Larkin et al. (2017) used global LUR models with data from 5,200 monitoring sites in 58 countries to predict NO_2 concentrations, achieving an RMSE of around 4.5 $\mu g/m^3$. Similarly, de Hoogh et al. (2016) developed Europe-wide LUR models for NO_2 and $PM_{2.5}$, incorporating satellite-derived and chemical transport modelling data. Our models' performance metrics are in the middle of those reported by the above studies, with RMSEs of around 1.15 $\mu g/m^3$ for $PM_{2.5}$ and 8 $\mu g/m^3$ for NO_2. However, as mentioned above these results are not directly comparable. Furthermore, it is also common that different study designs and datasets for the same region may lead to very different conclusions. For instance, in the United States, Ren et al. (2020) compared 13 spatio-temporal modelling algorithms, and found that nonlinear machine learning methods achieved higher prediction accuracy than statistical models. Conversely, Berrocal et al. (2020) found that spatial statistical models outperformed machine learning algorithms when predicting $PM_{2.5}$ concentrations. Clearly, it is impossible to tell for certain why these differences arise, but the most likely cause is that the random forests and the spatio-temporal smoothing models utilise different aspects of the data to make the predictions. Random forests are essentially non-spatial and hence use the information in the covariates to make the predictions, while the spatio-temporal smoothing models utilise a combination of the covariates and the residual (after covariate adjustment) spatio-temporal autocorrelation in the pollutant concentrations to make the predictions. Thus, if the main predictive facet of the data comes from the covariates then one would expect random forests to do better, while if residual spatio-temporal correlation is the dominant component then one would expect the spatio-temporal Kriging style models to do better. There are however many other differences between the studies listed above (and ours) that could be the reason for the different optimal prediction models, including differences in the quality and number of covariates, the temporal frequency of the data, the number and spatial configuration of the monitoring sites, and the type of cross-validation approach utilised.

The results from the final prediction models show that elevated concentrations of NO_2 are highly localised to the city centres, while particulate matter also exhibits elevated concentrations on the east coast due to trans-boundary pollution imported from

Europe. During the study period the predicted 5-year average concentrations of NO_2 across Scotland range between 4.46 and 30.9 in urban areas and between 2.12 and 22.7 in rural ones, which compare to targets for annual mean concentrations (not 5-year averages) of 40 $\mu g/m^3$ from the National Air Quality Strategy (NAQS) and 10 $\mu g/m^3$ from the more recent World Health Organisation (WHO) guidelines that were enacted after the study period in 2021. For PM_{10} these targets are 18 $\mu g/m^3$ (NAQS) and 15 $\mu g/m^3$ (WHO) respectively, and the 5-year average concentrations during the study were below these levels across the country (see Figure 10 in the supplementary material). Finally, for $PM_{2.5}$ these targets are 10 $\mu g/m^3$ (NAQS) and 5 $\mu g/m^3$ (WHO), and while the former was achieved for all of Scotland on average over the study period, the latter is not going to be easy to achieve as 14.14% of the grid squares did not meet this target albeit as a 5-year rather than an annual average.

Figure 4 highlights that NO_2 and $PM_{2.5}$ concentrations generally decreased during the study period in Scotland, while more recent data from Air Quality in Scotland (https://www.scottishairquality.scot/data/trends) shows that these decreasing trends have continued in the subsequent years, with statistically significant decreasing trends being observed at most monitoring sites. However, further reducing pollutant concentrations in the future is going to be even more challenging, because one has to reduce both emissions from sources critical to human productivity and livelihoods such as traffic and heating, as well as influencing emissions from other countries that are imported to Scotland as trans-boundary pollution.

The results of this paper provide a baseline for future research developments, which on the methodological side include the development of multivariate statistical and machine learning prediction models that borrow strength when making predictions across multiple pollutants. Meanwhile, one of the main motivations for this study is the prediction of monthly average pollutant concentrations with uncertainty quantification at a fine spatial resolution across Scotland, which will be used as exposure estimates in a future epidemiological study. Here, we aim to utilise monthly data on prescription rates for respiratory diseases treated in primary care, which are available from https://www.opendata.nhs.scot/dataset/prescriptions-in-the-community. A key challenge in this work will be producing representative pollution estimates for the populations who attend each general practitioner (GP) surgery, given that the exact locations of the patient populations are partially unknown. This uncertainty will need to be fed into the pollution exposure estimates, by allowing them to be inherently uncertain when estimating their health effects.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10651-024-00635-5.

Acknowledgements The authors gratefully acknowledge the helpful comments from the editor and reviewers, which have improved both the content and presentation of the paper.

Funding The work of the first author, Qiangqiang Zhu, is funded by the China Scholarship Council (CSC, No. 202208060260) as part of his PhD.

References

Air Quality Expert Group (2004) Nitrogen Dioxide in the United Kingdom. Report, Department for Environment, Food and Rural Affairs. https://uk-air.defra.gov.uk/library/assets/documents/reports/aqeg/nitrogen_dioxide_in_the_UK-summary.pdf

Bakar KS, Sahu SK (2015) spTimer: Spatio-temporal Bayesian modeling using R. J Stat Softw 63(15):1–32. https://doi.org/10.18637/jss.v063.i15

Bălă GP, Râjnoveanu RM, Tudorache E et al (2021) Air pollution exposure-the (in) visible risk factor for respiratory diseases. Environ Sci Pollut Res 28:19615–19628

Banerjee S, Carlin BP, Gelfand AE (2014) Hierarchical modeling and analysis for spatial data, 2nd edn. Chapman and Hall/CRC, Boca Raton. https://doi.org/10.1201/b17115

Berrocal V, Gelfand A, Holland D (2010) Spatio-temporal downscaler for output from numerical models. J Agric Biol Environ Stat 15:176–197

Berrocal VJ, Guan Y, Muyskens A et al (2020) A comparison of statistical and machine learning methods for creating national daily maps of ambient PM2.5 concentration. Atmos Environ 222:117130

Brauer M, Freedman G, Frostad J et al (2016) Ambient air pollution exposure estimation for the global burden of disease 2013. Environ Sci Technol 50(1):79–88

Breiman L (2001) Random forests. Mach Learn 45:5–32

Castelli M, Clemente FM, Popovič A et al (2020) A machine learning approach to predict air quality in California. Complexity 1:8049504

Chen J, de Hoogh K, Gulliver J et al (2019) A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. Environ Intern 130:104934

Chief Medical Officer (2022) Chief Medical Officer's annual report 2022: air pollution. Report, Department of Health and Social Care. https://www.gov.uk/government/publications/chief-medical-officers-annual-report-2022-air-pollution

Cressie N, Wikle CK (2015) Statistics for spatio-temporal data. John Wiley & Sons, Hoboken

de Hoogh K, Gulliver J, van Donkelaar A et al (2016) Development of West-European PM2.5 and NO2 land use regression models incorporating satellite-derived and chemical transport modelling data. Environ Res 151:1–10. https://doi.org/10.1016/j.envres.2016.07.005

Department for Environment Food & Rural Affairs (2023) Air quality statistics in the UK, 1987 to 2022 - Particulate matter (PM10/PM2.5). https://www.gov.uk/government/statistics/air-quality-statistics/concentrations-of-particulate-matter-pm10-and-pm25, updated 27 April 2023

Department for Environment, Food and Rural Affairs (2020) Air pollution in the UK 2019. Report, Department for Environment, Food and Rural Affairs, London, United Kingdom. https://uk-air.defra.gov.uk/assets/documents/annualreport/air_pollution_uk_2019_issue_1.pdf

Dibben C, Clemens T (2015) Place of work and residential exposure to ambient air pollution and birth outcomes in Scotland, using geographically fine pollution climate mapping estimates. Environ Res 140:535–541

Eren B, Aksangür İ, Erden C (2023) Predicting next hour fine particulate matter (PM2.5) in the Istanbul Metropolitan City using deep learning algorithms with time windowing strategy. Urban Clim 48:101418

Gao Z, Ivey CE, Blanchard CL et al (2023) Emissions and meteorological impacts on PM2.5 species concentrations in Southern California using generalized additive modeling. Sci Total Environ 891:164464

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7(4):457–472

Gu H, Yan W, Elahi E et al (2020) Air pollution risks human mental health: an implication of two-stages least squares estimation of interaction effects. Environ Sci Pollut Res 27:2036–2043
Guo B, Zhang D, Pei L et al (2021) Estimating PM2.5 concentrations via random forest method using satellite, auxiliary, and ground-level station dataset at multiple temporal scales across China in 2017. Sci Total Environ 778:146288
Hastie T, Tibshirani R (1986) Generalized additive models. Stat Sci 1(3):297–310. https://doi.org/10.1214/ss/1177013604
Hou K, Xu X (2022) Evaluation of the influence between local meteorology and air quality in Beijing using generalized additive models. Atmosphere 13(1):24. https://doi.org/10.3390/atmos13010024
Hu K, Rahman A, Bhrugubanda H et al (2017) Hazeest: machine learning based metropolitan air pollution estimation from fixed and mobile sensors. IEEE Sens J 17(11):3517–3525
Hu X, Belle JH, Meng X et al (2017) Estimating PM2.5 concentrations in the conterminous United States using the random forest approach. Environ Sci Technol 51(12):6936–6944
Larkin A, Geddes JA, Martin RV et al (2017) Global land use regression model for nitrogen dioxide air pollution. Environ Sci Technol 51(12):6957–6964
Larkin A, Anenberg S, Goldberg DL et al (2023) A global spatial-temporal land use regression model for nitrogen dioxide air pollution. Front Environ Sci 11:1125979
Li L, Wu J, Wilhelm M et al (2012) Use of generalized additive models and cokriging of spatial residuals to improve land-use regression estimates of nitrogen oxides in Southern California (Oxford, England: 1994). Atmos. Environ. 55:220–228. https://doi.org/10.1016/j.atmosenv.2012.03.035
Liu Y, Wang P, Li Y et al (2022) Air quality prediction models based on meteorological factors and real-time data of industrial waste gas. Sci Rep 12(1):1–15
Meinshausen N, Ridgeway G (2006) Quantile regression forests. J Mach Learn Res 7(6):983–999
Meyer H, Reudenbach C, Hengl T et al (2018) Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environ Model Softw 101:1–9
Montgomery DC, Peck EA, Vining GG (2021) Introduction to linear regression analysis, 6th edn. John Wiley & Sons, Hoboken
Mukhopadhyay S, Sahu SK (2017) A Bayesian spatiotemporal model to estimate long-term exposure to outdoor air pollution at coarser administrative geographies in England and Wales. J Royal Stat Soc Ser A 181(2):465–486. https://doi.org/10.1111/rssa.12299
Niu M, Zhang Y, Ren Z (2023) Deep learning-based PM2.5 long time-series prediction by fusing multi-source data-A case study of Beijing. Atmosphere 14(2):340
Novotny EV, Bechle MJ, Millet DB et al (2011) National satellite-based land-use regression: NO2 in the United States. Environ Sci Technol 45(10):4407–4414
Rajagopalan S, Al-Kindi SG, Brook RD (2018) Air pollution and cardiovascular disease: JACC state-of-the-art review. J Am Coll Cardiol 72(17):2054–2070
Ren X, Mi Z, Georgopoulos PG (2020) Comparison of machine learning and land use regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States. Environ Intern 142:105827
Rigby RA, Stasinopoulos DM (2005) Generalized additive models for location, scale and shape. J Royal Stat Soc Ser C 54(3):507–554
Rural and Environment Science and Analytical Services Division (2022) Scottish Government Urban Rural Classification 2020. Report, Scottish Government. https://www.gov.scot/publications/scottish-government-urban-rural-classification-2020/documents/
Saez M, Barceló MA (2022) Spatial prediction of air pollution levels using a hierarchical Bayesian spatiotemporal model in Catalonia, Spain. Environ Model Softw 151:105369
Sahu SK, Gelfand AE, Holland DM (2006) Spatio-temporal modeling of fine particulate matter. J Agric Biol Environ Stat 11:61–86
Wood SN (2017) Generalized additive models: an introduction with R. CRC Press, Boca Raton
Wood SN, Pya N, Säfken B (2016) Smoothing parameter and model selection for general smooth models. J Am Stat Assoc 111(516):1548–1563. https://doi.org/10.1080/01621459.2016.1180986
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, Geneva
Wright MN, Ziegler A (2017) ranger: a fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw 77(1):1–17. https://doi.org/10.18637/jss.v077.i01

Zhan Y, Luo Y, Deng X et al (2018) Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. Environ Pollut 233:464–473

Zou B, Chen J, Zhai L et al (2016) Satellite based mapping of ground PM2.5 concentration using generalized additive modeling. Remote Sens 9(1):1