medRxiv preprint doi: https://doi.org/10.1101/2020.07.19.20157396; this version posted September 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

A satellite-based spatio-temporal machine learning model to reconstruct daily $PM_{2.5}$ concentrations across Great Britain

Rochelle Schneider dos Santos[1,2,3]*, Ana M. Vicedo-Cabrera[4,5], Francesco Sera[1], Pierre Masselot[1], Massimo Stafoggia[6], Kees de Hoogh[7,8], Itai Kloog[9] , Stefan Reis[10,11], Massimo Vieno[10], and Antonio Gasparrini[1,2,12]

[1] Department of Public Health, Environments and Society, London School of Hygiene & Tropical Medicine, London WC1H 9SH, UK
[2] The Centre on Climate Change and Planetary Health, London School of Hygiene and Tropical Medicine, London WC1H 9SH, UK
[3] Forecast Department, European Centre for Medium-Range Weather Forecast, Reading, UK
[4] Institute of Social and Preventive Medicine, University of Bern, 3012 Bern , Switzerland
[5] Oeschger Center for Climate Change Research, University of Bern, 3012 Bern, Switzerland
[6] Department of Epidemiology, Lazio Regional Health Service, Rome 00147, Italy
[7] Swiss Tropical and Public Health Institute, Basel, Switzerland
[8] University of Basel, Basel, Switzerland
[9] Department of Geography and Environmental Development, Ben-Gurion University of the Negev, P.O.B. 653 Beer Sheva, Israel
[10] UK Centre for Ecology & Hydrology, Bush Estate, Penicuik, Edinburgh, Midlothian, EH26 0QB, UK
[11] University of Exeter Medical School, Knowledge Spa, Truro, TR1 3HD, United Kingdom
[12] Centre for Statistical Methodology, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK
* Correspondence: rochelle.schneider@lshtm.ac.uk

Abstract: Epidemiological studies on health effects of air pollution usually rely on measurements from fixed ground monitors, which provide limited spatio-temporal coverage. Data from satellites, reanalysis and chemical transport models offer additional information used to reconstruct pollution concentrations at high spatio-temporal resolution. The aim of this study is to develop a multi-stage satellite-based machine learning model to estimate daily fine particulate matter ($PM_{2.5}$) levels across Great Britain during 2008-2018. This high-resolution model consists of random forest (RF) algorithms applied in four stages. Stage-1 augments monitor-$PM_{2.5}$ series using co-located PM_{10} measures. Stage-2 imputes missing satellite aerosol optical depth observations using atmospheric reanalysis models. Stage-3 integrates the output from previous stages with spatial and spatio-temporal variables to build a prediction model for $PM_{2.5}$. Stage-4 applies Stage-3 models to estimate daily $PM_{2.5}$ concentrations over a 1 km grid. The RF architecture performed well in all stages, with results from Stage-3 showing an average cross-validated R^2 of 0.767 and minimal bias. The model performed better over the temporal scale when compared to the spatial component, but both presented good accuracy with an R^2 of 0.795 and 0.658, respectively. The high spatio-temporal resolution and relatively high precision allows this dataset (approximately 950 million points) to be used in epidemiological analyses to assess health risks associated with both short- and long-term exposures to $PM_{2.5}$.

Keywords: Fine Particulate Matter; Aerosol Optical Depth; Satellite; Reanalysis; Machine Learning; Random Forest

1. Introduction

The World Health Organization estimates in 7 million the global deaths associated with air pollution (both outdoor and household) every year, emphasising that exposure to particulate matter (PM) is among the greatest cause of concern [WHO,2020]. Fine particles with aerodynamic diameter smaller than 2.5 μm ($PM_{2.5}$) can penetrate into the human circulatory system through the lungs and

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.19.20157396; this version posted September 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

provoke multiple adverse health outcomes, including mortality [Liu et al., 2019] hospital admissions [Basagaña, et al., 2015], lung dysfunction [Raaschou-Nielsen et al., 2016], cardiovascular diseases [Lavigne, et al., 2019b], and allergic reactions [Lavigne et al., 2019a]. Usually, epidemiological studies collect air quality (AQ) data from ground monitors to quantify both short-term and long-term $PM_{2.5}$ exposures associated with acute and chronic health effects, respectively. The limitation in this health assessment approach is the lack of continuous temporal records of $PM_{2.5}$ and the limited spatial distribution of the monitors. Great Britain is an example of countries with very limited spatio-temporal coverage of $PM_{2.5}$, whereby the monitoring network is densely located only in major cities and widespread measurements of $PM_{2.5}$ only started from 2010.

Remote sensing observations of aerosol optical depth (AOD) obtained from satellites, which measures how much direct sunlight has been scattered and absorbed by aerosols particles suspended in the atmosphere, has recently been proposed as an alternative to measure PM variability for epidemiological purposes [NASA Earth Observations, 2020]. However, while offering the advantage of global coverage and relatively high spatio-temporal resolution, the use of AOD for $PM_{2.5}$ exposure assessments present limitations, for instance the fact that it represents the total atmospheric column concentration of aerosol rather than surface values [van Donkelaar, 2010]. Unsurprisingly, early studies based only on satellite-AOD achieved very low performances in predicting $PM_{2.5}$ [Koelemeijer et al., 2006]. Recent studies have proposed more sophisticated approaches, combining AOD measures with information from other satellite products, reanalysis data, chemical transport models, and geospatial features to improve the prediction of $PM_{2.5}$. Such studies used various analytical methods, including multiple linear regression [Gupta and Christopher, 2009], land-use regression [Beckerman et al., 2013; Vienneau et al., 2010], and mixed effect models [de Hoogh et al., 2018; Kloog et al., 2011; Kloog et al., 2015; Lee et al., 2011; Staffogia et al., 2016]. The last development in this research area is represented by the application of machine learning (ML) algorithms, including various architectures such as random forests [Chen et al., 2018; Di et al., 2019; Staffogia et al., 2019; Yazdi, et al., 2020], neural network [Di et al., 2019; Yazdi, et al., 2020], and gradient boosting [Chen et al., 2019; Di et al., 2019; Just et al., 2018; Zan et al., 2017]. These have demonstrated higher performances, linked with the ability to model any kind of predictor(s)-response association and to deal better with potentially complex relationship between $PM_{2.5}$, spatial, and spatio-temporal predictors [Di, et al., 2019; Polley et al., 2010].

The aim of this study is to develop and apply a multi-stage satellite-based ML model to estimate daily concentrations of $PM_{2.5}$ over a 1 km grid across Great Britain in the period 2008-2018. The analysis is based on a dataset with synchronised information from various data sources, such as several remote sensing satellite products, multiple climate and atmospheric reanalysis databases, chemical transport models, and spatial and spatio-temporal variables. The model is assessed through measures of predictive performance, error, and bias, obtained through cross-validation.

2. Materials and Methods

2.1. Study area and period

Great Britain is an island with an extension of 229,462 km^2 surrounded by the Atlantic Ocean, Irish Sea, North Sea, and English Channel. It comprises the countries of England, Scotland, and Wales with a total population in 2018 of almost 65 million [ONS, 2020]. According to the Köppen climate classification, the United Kingdom (Great Britain and Northern Ireland) is defined as having a warm temperate climate, fully humid with mostly warm summer (cold summer for some parts of Scotland and England) [Kottel et al., 2006]. The study area included 234.429 1 km grid cells (containing a unique identification code, cell-ID) from the original 1 km Great Britain National Grid Squares [Digimap, 2020] for a period between 1st January 2008 and 31st December 2018.

2.2. $PM_{2.5}$ and PM_{10} observed data

Daily PM_{10} and $PM_{2.5}$ (μg/m^3) measurements in the study period were obtained from five monitoring network sources through the R package *openair* [Openair, 2020]: Automatic Urban and

medRxiv preprint doi: https://doi.org/10.1101/2020.07.19.20157396; this version posted September 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Rural Network (AURN), Air Quality England (AQE), King College London (KCL), Scotland Air Quality Network (SAQN), and Wales Air Quality Network (WAQN). When monitors from different sources showed exactly the same temporal distributions (*i.e.*, correlation equal to 1) and were located at approximately the same coordinates, only the AURN monitor was kept. Monitors with less than 18 hours of $PM_{2.5}$ records per day as well as less than 30 days by year were removed. The final set includes 581 and 183 monitors measuring PM_{10} and $PM_{2.5}$ along the study period, respectively. Figure 1 shows the locations of these monitors across Great Britain, illustrating how the network coverage is densely located in major cities, leaving several small cities and rural areas with only a few or none AQ records. Each monitor was indexed using the cell-ID of the 1 km grid cell that contains it.

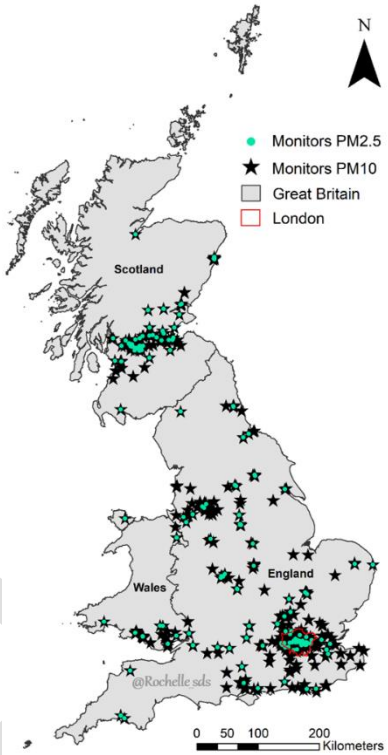

Figure 1: Spatial distribution of 581 PM_{10} (black star) and 183 $PM_{2.5}$ (turquoise dots) monitors across Great Britain during the study period.

2.3. *Spatially-lagged and nearest monitor $PM_{2.5}$ variables*

Four spatio-temporal variables were generated from the monitor series of $PM_{2.5}$ to represent spatially-lagged annual average concentrations. Monitor types were grouped into two classes (background (urban, suburban, and rural) and hotspots (traffic and industrial) to compute the annual averages values of nearby monitors by class using inverse-distance weighted leave-one-out cross-validated (IDW-LOOCV) approach. Two different weights were applied, namely the inverse distance and the inverse squared distance (in km). The former assigns relatively more weight to distant monitors, and therefore can represent a *regional* background, while the latter captures *local* differences. Therefore, these four IDW-LOOCV variables were named as follow: (i) Spatially–lagged hotspot-$PM_{2.5}$ regional, (ii) Spatially–lagged background-$PM_{2.5}$ regional, (iii) Spatially–lagged hotspot-$PM_{2.5}$ local, and (iv) Spatially–lagged background-$PM_{2.5}$ local. To improve the model performance in the spatial domain, two additional spatial variables were generated based on the closest Euclidean distance for each monitor class, named as: (i) Nearest hotspot monitor distance and (ii) Nearest background monitor distance. These six variables were used as additional predictors to capture the heterogeneity across monitors and exploit their spatial autocorrelation, and thus help the model to better categorise differences in the spatial patterns of measured-$PM_{2.5}$ series.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.19.20157396; this version posted September 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

2.4. AOD data: satellite and atmospheric reanalysis models

Daily satellite-AOD was obtained from the Collection 6 Level-2 gridded product (MCD19A2). These data are generated at a 1 km grid through the Multi-angle Implementation of Atmospheric Correction (MAIAC) algorithm using data from Moderate Resolution Imaging Spectroradiometer (MODIS) sensor on board both Terra and Aqua earth observation satellites [Lyapustin and Wang, 2018]. Four layers from MCD19A2 product were extracted: (i) AOD Blue band (0.47μm), (ii) AOD Green band (0.55μm), (iii) AOD uncertainty (i.e. level of uncertainty based on blue-band surface brightness [reflectance]), (iv) AOD_QA (quality assurance flags to retrieve only the best quality AOD). These layers are generated for each passing time of Terra and Aqua satellites over the area of study and combined by day. The layers AOD-0.47μm and AOD-0.55μm were used as the outcome variables, after their values were filtered using AOD uncertainty and AOD_QA layers to guarantee high product quality.

This calibration process, together with pixels covered by clouds, removed a large sample of AOD grid cells over Great Britain. To fill the satellite-AOD gaps, modelled-AOD total column was used from Copernicus Atmosphere Monitoring Service (CAMS) reanalysis provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) [Bozzo, et al., 2017]. CAMS reanalysis provides every 3-hourly modelled-AOD at five different wavelengths (0.47μm, 0.55μm, 0.67μm, 0.87μm, and 1.24μm) with a spatial resolution of approximately 80 km but the data was downloaded at 10 km, based on an interpolation performed through the ECMWF's API request. The satellite-AOD and CAMS modelled-AOD were indexed to the closest 1 km grid cell from their pixel centroid.

2.5. Other spatio-temporal predictors

2.5.1. Modelled $PM_{2.5}$ from chemical transport models

Atmospheric chemistry transport models (ACTMs) incorporate anthropogenic and natural sources of emission, land use and meteorological conditions to simulate the atmospheric compositions and deposition of various air pollutants (trace gases and particles). Based on the European Modelling and Evaluation Programme (EMEP) ACTM, EMEP4UK, has been developed to represent the UK hourly atmospheric composition at a spatial resolution of 5 km [EMEP4UK, 2020]. The description of EMEP4UK model framework and setup can be found elsewhere [Vieno et al., 2010]. Daily EMEP4UK simulations of $PM_{2.5}$ ($\mu g/m^3$) concentrations at surface-level were included to represent ground-level contributions, in contrast to AOD products that refer to the total column of aerosols concentrations. Each EMEP4UK 5 km pixel was linked to the closest 1 km grid cell centroid.

2.5.2. Meteorological variables from climate reanalysis models

Meteorological variables were retrieved from the ECMWF's climate reanalysis models with the highest spatial resolution available during 2008-2018 and at two sub-day times (0:00 and 12:00). Sea pressure and the boundary layer height (BLH) were downloaded from the ERA 5 global reanalysis with a spatial resolution of approximately 30 km [ERA5, 2020]. Air temperature at 2m height and total precipitation were obtained from the ERA 5 Land global reanalysis with a spatial resolution of approximately 9 km [ERA5 Land, 2020]. Relative humidity, wind direction, and wind speed were downloaded from UERRA regional reanalysis at 5.5 km for the MESCAN-SURFEX system [UERRA, 2020]. All meteorological variables were indexed to the closest 1 km grid cell to their centroid.

2.5.3. Normalized difference vegetation index

Monthly Normalized difference vegetation index (NDVI) layer is used to quantify vegetation presence and it ranges from -1 to 1. The NDVI 1 km grid was obtained from MOD13A3 Version 6 Level 3, Terra-MODIS product [Didan et al., 2015]. Each NDVI pixel was indexed to the closest 1 km grid cell to its pixel centroid and the NDVI values repeated for the days inside the corresponded month.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.19.20157396; this version posted September 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

2.6. Spatial predictors

2.6.1. Land variables and night-time light data from earth observation satellites

Three land predictors were collected from the Copernicus Land Monitoring Service (CLMS) [CLMS, 2020] database: elevation, land cover, and impervious surfaces. Elevation data was obtained from the 2011 European Digital Elevation Model (EU-DEM) version 1.1 with a spatial resolution of 25 m. The elevation values were obtained from the mean of all 25 m-pixels values located inside each 1 km grid cells. Land cover data was obtained from 2012 CORINE Land Cover (CLC) inventory. It was derived from high-resolution ortho-rectified satellites images that mapped all land elements at a spatial resolution ranging from 5 m to 60 m and aggregated into 100 m. Nine predictors were defined by grouping the original 44 CLC classes and each predictor represents its group proportion inside each 1 km Great Britain Grid cell. Imperviousness degree is a binary raster product at a spatial resolution of 100 m, where the value 0 represents natural land cover or water surface and value 1 represents entirely artificial surfaces (i.e. built-up areas). The amount of impervious surfaces was estimated by the proportion of artificial surfaces inside each 1 km Great Britain Grid cell.

Night-time lights data was provided by the visible Infrared Imaging Radiometer Suite (VIIRS) sensor aboard the Suomi-National Polar-orbiting Partnership (Suomi-NPP) satellite. The VIIRS Day/Night band collects cloud-free average radiance values at annual and monthly composites in spatial resolution of 750m [EOG, 2020]. The 2015 night-time lights annual mean was computed based on the weighted average of VIRRS pixels inside each 1 km Great Britain Grid cell.

2.6.2. Population density

The resident population counts data from 2011 were collected from the Office for National Statistics (England and Wales) and the National Records of Scotland. The smallest geographic unit of the UK census is output areas (OA), with a total of 227.769 OAs polygons for Great Britain. The proportion of each OA's area inside each 1 km Great Britain Grid cell was extracted to estimate the weighted average population by 1 km grid cell.

2.6.3. Road density and distance

Road density and length predictors were derived from the Ordnance Survey (OS) [2020] Open Roads product, which offers a geospatial representation of Great Britain's Road network. Three density predictors were defined by grouping the original eight OS road types, where each predictor was computed as the sum of all roads length inside each 1 km Great Britain Grid cell by group (highway, secondary, and local). Three distance predictors were defined by computing the inverse distance of each 1 km grid centroid from the closest road group (highway, secondary, and local).

2.6.4. Inverse distance from airports and seashore

Information on location and size of airports was derived from the Civil Aviation Authority (CAA) [2020], which collects monthly statistics about air traffic movements for more than 60 UK Airports, including aviation activities for terminal passengers, commercial flights, and cargo tonnage. A total of 19 airports across Great Britain were selected based on a minimum of 1% for the annual percentage of passengers at the airport during 2015-2018. For each cell of the 1 km Great Britain grid the inverse distance from the closest airport was calculated.

The inverse distance from seashore was computed for each 1 km grid cell using the geographical information about the boundaries of England, Wales, and Scotland provided by the UK Data Service [2020].

2.7. Statistical methods

A four-stage model was developed to obtain daily $PM_{2.5}$ concentrations for all 234.429 grid cells covering Great Britain. Each stage is described in detail below. Briefly, Stage-1 applies a random forest (RF) algorithm to predict $PM_{2.5}$ concentrations in monitors with only PM_{10} records. Stage-2 uses

medRxiv preprint doi: https://doi.org/10.1101/2020.07.19.20157396; this version posted September 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

RF models to impute missing satellite–AOD from Terra and Aqua satellites, using modelled-AOD from CAMS. Stage-3 combines the output from Stage-1 and Stage-2 with a list of spatial and spatio-temporal synchronised predictors to estimate $PM_{2.5}$ concentrations at the locations of the monitors. Stage-4 uses the Stage-3 model to predict daily $PM_{2.5}$ across the whole Great Britain.

2.7.1. Random forest algorithm

RF is a supervised tree-based design ML algorithm which trains an ensemble of independent decision trees (or forest) in parallel. The final model accuracy is estimated by the performance average of all decision trees. There are two main advantages of the RF architecture (known as bagging ensemble method): first, it controls the bias-variance trade-off by feeding each tree model with two-thirds of the training set while one-third is left out for validation (*i.e.*, out-of-bag (OOB)) [Schneider dos Santos, 2020]. When all decision trees receive the same amount and list of predictors, they become highly correlated, not solving the variance problem. Therefore, the second positive aspect from RF is that the number of predictors on each tree is less than the full list available and they are randomly selected. This approach changes the predictor placed on the top of the tree, generating different splits and internal nodes. The algorithm is then able to estimate an importance ranking by quantifying the amount of error decreased due to a split of a specific predictor [James et al., 2013].

The performance of the models in each stage was assessed using statistics based on OOB samples and then from a 10-fold cross-validation (CV) procedure based on monitors. In the latter, ten random groups of monitors were defined, and the complete outcome series in each group were predicted using a model fitted in the other nine. This procedure offers a measure of true predictive ability of the RF model in locations where no monitor is available. Measures of performance were generated by regressing the OOB or cross-validated predicted values on the observed series, and computing the R^2, the root mean square error (RMSE), and intercept and slope of the prediction. These statistics were computed overall using the whole series, and then separated in spatial and temporal contributions. The former was computed using the averages of predicted and observed values across the series, and it offers a measure of performance in capturing long-term average $PM_{2.5}$ concentrations. The latter was computed as daily deviations from the averages, and it quantifies the temporal variability explained by the model.

2.7.2. Stage-1: increasing $PM_{2.5}$, measurements using co-located PM_{10} monitors

The number of $PM_{2.5}$ monitors across Great Britain at the beginning of the study period was relatively low, and even if the quantity has increased substantially after 2010, most of the monitors were mostly installed in major cities. Stage-1 aims to increase the number of spatio-temporal ground-level $PM_{2.5}$ references using observations from co-located monitors measuring PM_{10}, which are available in higher number of locations and better distributed across Great Britain. Specifically, in this stage we fitted a RF model in locations with both $PM_{2.5}$ and PM_{10} measurements, using data from all years (2008-2018). The RF model for each year y is defined as:

$$PM^{y}_{2.5_{i,t}} = f(PM_{10_{i,t}},\ mon.type_i,\ month_t,\ dow_t,\ lat_i\ ,\ lon_i) \qquad (1)$$

where: $PM^{y}_{2.5_{i,t}}$ and $PM_{10_{i,t}}$ are the target variable and main predictor, respectively, measured in year y at monitor i on day t; $mon.type_i$ is a categorical variable classifying monitor i (traffic, industrial, urban, suburban, and rural); $month_t$ and dow_t are categorical variables representing the months and day of the week of day t; and lat_i and lon_i define the coordinates of monitor i. The RF model was defined based on the best parameters setting. The optimised parameters were 500 decision trees as the RF ensemble (Ntree=500) and four variables randomly selected to be used on each tree (mtry=4). This model was eventually used to predict $PM_{2.5}$ measurements in locations/days in which only PM_{10} was measured.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.19.20157396; this version posted September 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

2.7.3. Stage-2: imputing missing satellite-AOD from CAMS modelled-AOD

The percentage of missing satellite-AOD measurements in Great Britain ranged between 87% and 94% during 2008-2018 with the greatest portion during autumn and winter, near to the coast, and at the North of Scotland. Stage-2 imputes satellite-AOD missing for every day and 1 km grid based on an optimised RF model (Ntree=50 and mtry=20) and satellite-AOD wavelength (0.47μm and 0.55μm), separately in each year within the study period. These RF models were built for each year y as follows:

$$\text{Satellite-}AOD_{i,t}^{(z,y)} = f(CAMS.AOD_{1,1_{i,t}}, \dots, CAMS.AOD_{5,1_{i,t}}, CAMS.AOD_{1,2_{i,t}}, \dots, CAMS.AOD_{5,7_{i,t}}, yday_t, lat_i, lon_i) \quad (2)$$

where: Satellite-$AOD_{i,t}^{z,y}$ is the target variable representing satellite-AOD (wavelength z, year y) estimates at grid cell i on day t; CAMS.AOD is the main predictor representing CAMS modelled-AOD estimates at grid cell i, on day t, at five wavelengths (0.47μm, 0.55μm, 0.67μm, 0.865μm, and 1.24μm), and at seven sub-day times (3h, 6h, 9h, 12h, 15h, 18h, and 21h); $yday_t$ defines the sequence of days in a year from 1 to 365 (366, for leap years); lat_i and lon_i represent the coordinates of grid cell centroid i. This model was eventually used to predict missing satellite-AOD measurements.

2.7.4. Stage-3: estimating $PM_{2.5}$ concentrations using spatial and spatio-temporal variables

Stage-3 aims to build a predictive model for daily $PM_{2.5}$ concentrations, using as target variable the combined set of $PM_{2.5}$ directly measured from monitors or predicted from Stage-1, AOD measured from satellite instruments or predicted from Stage-2, together with all the spatially and spatio-temporally synchronized predictors described in the previous section. The RF models were fit separately by year y. The optimization process for this stage selected Ntree=500 and mtry=20 as the best set of parameter values. In this stage, $PM_{2.5}$ concentrations were log-transformed to ensure the prediction of non-negative values. The Stage-3 model for each year y is defined as:

$$\log\left(PM_{2.5}{}_{i,t}^{y}\right) = f(SPT_{1_{i,t}}, \dots SPT_{15_{i,t}}, SP_{1_i}, \dots, SP_{27_i}) \quad (3)$$

where: $\log(PM_{2.5}{}_{i,t}^{y})$ is the target variable representing the log-$PM_{2.5}$ concentrations in year y at the monitor located in grid cell i on day t, while $SPT_{i,t}$ and SP_i represent the spatio-temporal and spatial predictors, respectively.
Specifically, $SPT_{i,t}$ included: surface-level log-$PM_{2.5}$ values from the EMEP4UK model; Stage-2 AOD (at 0.47μm, 0.55μm); meteorological variables (air temperature, sea pressure, relative humidity, total precipitation, wind speed and direction); BLH (at 12:00 and 24:00 hours); monthly-averaged NDVI; and day of the year, month, and week. SP_i included: spatially–lagged regional and local log-$PM_{2.5}$ average concentrations from background and hotspot groups; nearest background and hotspot monitor distance; land variables (elevation, land cover in 9 groups, and % of impervious surface); night-time light; road density and inverse distance (each for highway, secondary, and local); population density; and inverse distance from closest airport and seashore.

2.7.5. Stage-4: reconstructing $PM_{2.5}$ time-series at 1 km grid

Using the RF models developed by year in Stage-3, daily $PM_{2.5}$ concentrations for each 1 km grid cell were reconstructed across Great Britain for the whole study period (2008-2018).

medRxiv preprint doi: https://doi.org/10.1101/2020.07.19.20157396; this version posted September 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

3. Results

3.1. Stage-1 results

Not all monitors in each network provided the full series of daily PM concentrations during 2008-2018. However, PM_{10} was more often available, especially in the years 2008-2009, when there were less than 80 $PM_{2.5}$ monitors across Great Britain. The imputation process in Stage-1 enabled the expansion from 46 to 269 in 2008 and from 65 to 278 in 2009. Table 1 shows the Stage-1 model performance (*i.e.*, a separate RF model for each year) reported by two CV methods: (i) OOB - the original RF CV algorithm, which presented better accuracy (R^2 average of 0.932), and (ii) 10-Fold CV - a targeted sampling, which leaves-out monitors with the full set of observations (R^2 average of 0.855). As expected, the 10-Fold CV approach resulted in a slightly lower predictive accuracy although still showing a good performance. There is some indication of bias in the 10-Fold CV, with a slope lower than the expected value of 1 and a slightly negative intercept. The results for 10-Fold CV spatial and temporal domains are in Table A1 in Appendix A.

Table 1. Predicted-$PM_{2.5}$ concentrations obtained from Stage-1 RF models were regressed against measured-$PM_{2.5}$ concentrations in a linear regression model. The performance was evaluated using two CV methods (OOB and 10-Fold) together with RMSE (measure of the model error, μg/m³), intercept (μg/m³), and slope (μg/m³).

Stage-1

	OOB-CV				10-Fold CV			
	R2	RMSE	Inter.	Slope	R2	RMSE	Inter.	Slope
2008	0.918	1.196	-0.417	1.033	0.707	4.954	0.803	0.886
2009	0.921	1.110	-0.390	1.030	0.791	3.996	0.410	0.937
2010	0.919	1.122	-0.401	1.029	0.843	3.496	0.043	0.983
2011	0.949	1.124	-0.266	1.019	0.902	3.439	-0.087	0.997
2012	0.942	1.058	-0.274	1.021	0.889	3.218	-0.035	0.986
2013	0.929	1.087	-0.368	1.028	0.847	3.584	0.218	0.972
2014	0.944	0.963	-0.267	1.022	0.891	3.003	-0.007	0.995
2015	0.933	0.865	-0.265	1.026	0.871	2.662	-0.003	0.983
2016	0.935	0.896	-0.251	1.025	0.885	2.654	-0.050	0.996
2017	0.939	0.828	-0.196	1.022	0.895	2.430	0.010	0.985
2018	0.928	0.791	-0.248	1.028	0.886	2.235	-0.019	0.993
Mean	0.932	1.003	-0.304	1.026	0.855	3.243	0.117	0.974

3.2. Stage-2 results

The Stage-2 procedure is illustrated in Figure 2, showing missing satellite-AOD, imputed through a combination of multiple modelled-AOD wavelengths and sub-day times. Table 2 shows the performance of Stage-2 models validated using the OOB CV method. The results presented consistently high R^2 (ranging from 0.963 to 0.989), low RMSE (varying between 0.006 to 0.011), and almost no bias (intercept zero and slope close to 1 in almost all years and wavelengths).

medRxiv preprint doi: https://doi.org/10.1101/2020.07.19.20157396; this version posted September 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

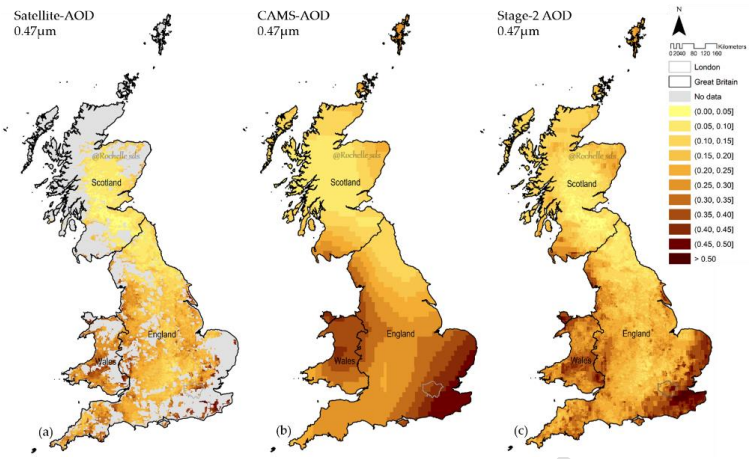

Figure 2: Satellite-AOD 0.47µm values are represented in Figure 2a (mean of all Terra- and Aqua-MODIS passing times). The missing values in grey were imputed through a combination of multiple modelled-AOD wavelengths and sub-day times, represented in Figure 2b by modelled-AOD 0.47µm at 12:00. Stage-2 output is shown in Figure 2c, illustrating the full coverage of AOD 0.47µm based on the combination of measurements and estimations across Great Britain. The maps correspond to values measured or reconstructed on 6th July 2018.

Table 2. Predicted-AOD 0.47µm and 0.55µm obtained from Stage-2 RF models were regressed against measured Satellite-AOD 0.47µm and 0.55µm in a linear regression model. The performance was evaluated using OOB CV together with RMSE (µg/m^3), intercept (µg/m^3), and slope (µg/m^3).

Stage-2 OOB CV

	Predicted-AOD 0.47µm				Predicted-AOD 0.55µm			
	R2	RMSE	Inter.	Slope	R2	RMSE	Inter.	Slope
2008	0.977	0.010	-0.001	1.009	0.977	0.007	-0.001	1.009
2009	0.976	0.010	-0.001	1.010	0.976	0.007	-0.001	1.010
2010	0.968	0.009	-0.001	1.013	0.968	0.007	-0.001	1.013
2011	0.988	0.010	-0.001	1.005	0.988	0.007	0.000	1.005
2012	0.980	0.010	-0.001	1.008	0.981	0.007	-0.001	1.008
2013	0.984	0.010	-0.001	1.007	0.984	0.007	-0.001	1.006
2014	0.970	0.009	-0.001	1.012	0.970	0.007	-0.001	1.012
2015	0.972	0.009	-0.001	1.011	0.973	0.007	-0.001	1.011
2016	0.975	0.009	-0.001	1.010	0.975	0.007	-0.001	1.010
2017	0.963	0.009	-0.001	1.015	0.963	0.007	-0.001	1.014
2018	0.969	0.010	-0.001	1.013	0.969	0.007	-0.001	1.013
Mean	0.978	0.010	-0.001	1.009	0.978	0.007	-0.001	1.009

3.3. Stage-3 results

Stage-3, the main step of the satellite-based machine learning framework, combines the output of Stage-1 and Stage-2 with a list of spatial and spatio-temporal predictors to estimate $PM_{2.5}$ at the locations of the monitors. The relative importance of the predictors for the Stage-3 RF models are ranked in Table 3. The list with the top-15 predictors demonstrates the larger contribution of the more informative spatio-temporal variables (EMEP4UK $PM_{2.5}$, meteorological parameters, BLH, Stage-2 AOD, and NDVI). All the proposed spatially–lagged $PM_{2.5}$ variables were classified as highly important and their ranking positions varied slightly across the displayed years. This suggests the presence of spatial correlations in $PM_{2.5}$ values that are not entirely captured by the other variables.

Table 4 shows the results of the 10-Fold CV Stage-3 RF models by year. The results indicate a good predictive performance of the model throughout the study period. Overall cross-validated R^2 ranged from 0.704 (2008) to 0.821 (2011), with an average of 0.767. The average prediction error is 4.042 µg/m^3, with negligible bias in intercept and slope. The inspection of the spatial and temporal

medRxiv preprint doi: https://doi.org/10.1101/2020.07.19.20157396; this version posted September 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

contributions shows that the model performs well generally across the two components, displaying a spatial performance drop only for 2008 (0.486) and 2015 (0.579). The cross-validated spatial R^2 ranges from 0.486 (2008) to 0.746 (2017), while the cross-validated temporal R^2 from 0.760 (2008) to 0.843 (2011). The two components have an average of 0.658 and 0.795, respectively. The high spatial R^2 performance across the years demonstrates that the Stage-3 RF models were able to predict the spatial variation of long-term $PM_{2.5}$ across Great Britain with good accuracy. In Appendix B, Table B1 shows the Stage-3 results by season, demonstrating that the seasonal patterns were well described by RF models, although with lower accuracy for the temporal domain in summer, characterized by a lower 10-Fold R^2. This drop in performance during summer was also seen by other studies [Stafoggia et al., 2019; Stafoggia et al., 2016].

Table 3. Relative importance (%) of the predictors in Stage-3 for the first, middle and last years.

Stage-3 Predictors	2008	2013	2018
EMEP4UK $PM_{2.5}$	32.41	32.83	36.74
Spatially–lagged hotspot-$PM_{2.5}$ regional	2.55	2.49	6.77
Wind direction	6.35	7.33	5.34
Spatially–lagged background-$PM_{2.5}$ regional	1.14	6.06	4.93
Day of the year	3.22	4.33	3.73
Spatially–lagged hotspot-$PM_{2.5}$ local	3.97	1.65	3.66
Precipitation	6.63	2.42	3.25
BLH 0h	2.28	2.91	2.79
Spatially–lagged background-$PM_{2.5}$ local	0.94	3.07	2.75
Month	1.76	2.72	2.68
2m Air temperature	2.60	2.93	2.65
Wind speed	3.08	3.75	2.53
Sea pressure	3.09	2.60	2.49
Relative humidity	1.77	1.56	1.81
Nearest background monitor distance	3.10	2.30	1.78

Note: The top 15 predictors in the RF importance ranking order is determined by the year 2018. The importance is measured by the amount of error reduced due to splits of a given predictor over all trees used in the RF ensemble.

Table 4. Predicted-$PM_{2.5}$ concentrations obtained from Stage-3 RF models were regressed against Stage-1 measured/predicted-$PM_{2.5}$ concentrations in a linear regression model. The CV–R^2 (how well the model described the $PM_{2.5}$ variability in new locations) described in three different patterns (overall, spatial, and temporal), RMSE (measure of the model error, μg/m³), intercept (μg/m³), and slope (μg/m³).

Stage-3

	Overall				Spatial				Temporal			
	R2	RMSE	Inter.	Slope	R2	RMSE	Inter.	Slope	R2	RMSE	Inter.	Slope
2008	0.704	4.547	-1.251	1.064	0.486	2.698	-0.749	1.026	0.760	3.677	0.000	1.074
2009	0.742	4.247	-1.104	1.042	0.680	2.255	-0.203	0.982	0.762	3.593	0.000	1.055
2010	0.709	4.330	-1.424	1.075	0.627	2.342	0.137	0.972	0.738	3.628	0.000	1.102
2011	0.821	4.421	-0.898	1.029	0.733	2.280	-0.509	1.003	0.843	3.756	0.000	1.035
2012	0.786	4.354	-0.749	1.027	0.661	2.527	0.073	0.966	0.823	3.552	0.000	1.043
2013	0.764	4.305	-1.093	1.047	0.637	2.616	-0.565	1.013	0.791	3.604	0.000	1.061
2014	0.784	4.140	-1.044	1.051	0.632	2.292	-0.145	0.983	0.815	3.478	0.000	1.062
2015	0.736	3.792	-1.194	1.072	0.579	2.139	-0.026	0.969	0.776	3.127	0.000	1.095
2016	0.781	3.702	-0.980	1.050	0.725	1.964	-0.532	1.010	0.796	3.149	0.000	1.061
2017	0.816	3.343	-0.933	1.041	0.746	1.720	-0.406	0.994	0.834	2.860	0.000	1.055
2018	0.790	3.275	-1.030	1.046	0.726	1.776	-0.745	1.015	0.807	2.775	0.000	1.056
Mean	0.767	4.042	-1.064	1.049	0.658	2.237	-0.334	0.994	0.795	3.382	0.000	1.063

DRAFT

medRxiv preprint doi: https://doi.org/10.1101/2020.07.19.20157396; this version posted September 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

3.4. Stage-4 results

Stage-4 provides the prediction of daily $PM_{2.5}$ concentrations for each of the 234.429 1 km grid covering Great Britain. Results indicate an annual $PM_{2.5}$ average of 9.41 μg/m³ for 2008, 10.17 μg/m³ for 2013, 8.05 μg/m³ for 2018, and 8.84 μg/m³ for 2008-2018 but with a strong spatial and temporal variation. The spatial distribution of annual average $PM_{2.5}$ concentrations for 2008, 2013, and 2018 are shown in Figure 3, revealing a decrease of pollution levels in recent years across the whole territory, although slightly stronger in England. Table C1 in Appendix C provides the same figures for all the years, confirming the decreasing trend. The spatial comparison suggests that $PM_{2.5}$ concentrations are lower in Scotland and Wales compared to the more populated southern regions of England, with hotspots located in urban areas such as Liverpool, Manchester, Birmingham, and Greater London. At the bottom of Figure 3, the maps display the corresponding annual average of $PM_{2.5}$ levels in London, demonstrating the precision of the multi-stage ML model in reconstructing $PM_{2.5}$ concentrations in 1 km grid cells within urban areas. The maps show local hotspots of high pollution, with a spatial distribution that however changes along the study period.

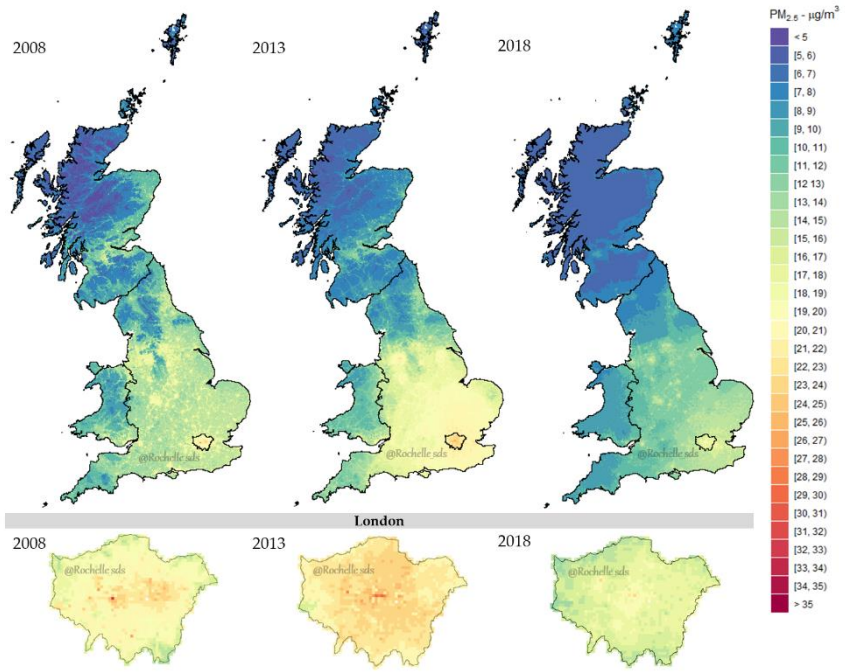

Figure 3: Stage-4 predicted $PM_{2.5}$ concentrations across Great Britain (Top) and London (Bottom) for 2008, 2013, and 2018 aggregated by annual means. All plots were built under the same colour scale.

The greatest contribution of this study is the ability of the satellite-based ML models to reconstruct daily levels of $PM_{2.5}$ over a 1 km grid across a wide geographical domain. Figure 4 displays the spatial distribution of $PM_{2.5}$ estimations across Great Britain (Top) and London (Bottom) for specific days within the study period. It is interesting to note the wide variation in $PM_{2.5}$ concentrations between days in the same area and between areas in different days. For instance, the maps in the left panels represent a day with almost no variation and generally low concentrations; the maps in the mid panels display a strong north/south split likely linked to weather conditions; the right panels show a more complex pattern with wider range of $PM_{2.5}$ values, a large area with very high pollution concentrations located in east London, and hotspots in highly-populated urban areas across England.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.19.20157396; this version posted September 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

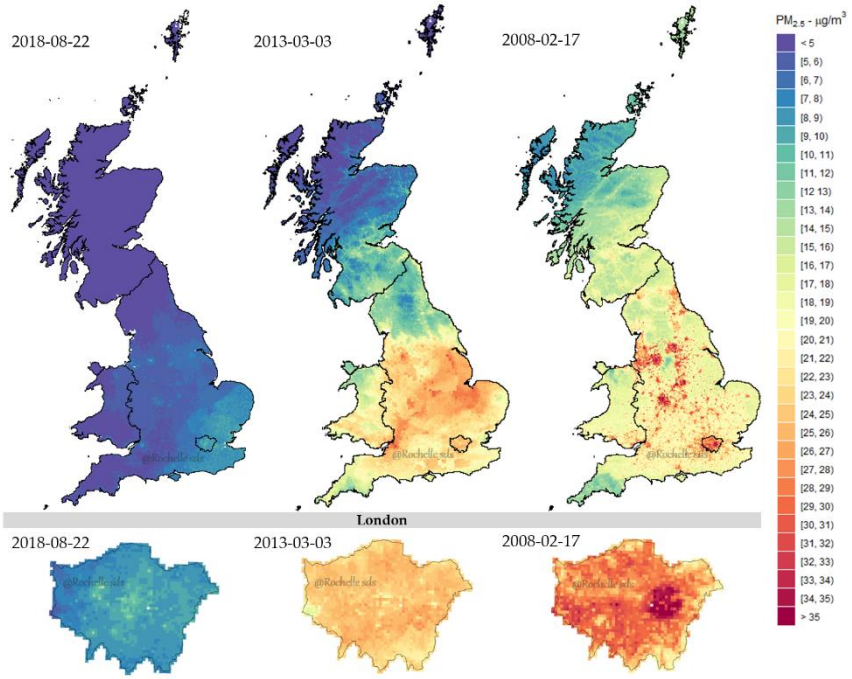

Figure 4: Stage-4 day-specific $PM_{2.5}$ estimations across Great Britain (Top) and London (Bottom).

4. Discussion

This study presents the first application of satellite-based spatio-temporal ML methods to reconstruct levels of pollution across Great Britain, providing estimates of daily $PM_{2.5}$ concentrations over a 1 km grid during 2008-2018. The multi-stage ML framework provided significant advantages, allowing the combination of information from multiple data sources, such as air quality monitoring networks, remote sensing satellite products, chemical dispersion models, reanalysis databases, administrative census data, among others.

The beginning of the study period (2008 and 2009) had lower quantity of monitors measuring $PM_{2.5}$ across Great Britain, but this number increased considerably from 2010 after a new measuring network has been established in 2009 [DEFRA, 2012]. Therefore, Stage-1 was an extremely important step to extend the number of $PM_{2.5}$ measurements. The implementation of Stage-2 was also relevant to fill the gaps in MAIAC AOD retrievals from satellites, thus maximizing the available information. Stage-3 produced a ML prediction model based on a long list of informative predictors, accounting for potentially complex inter-relationships and functional forms. The Stage-3 ML algorithms offered an excellent performance, showing an average cross-validated R^2 of 0.767 across the period, with increased predictive ability in the last years. Stage-4 provided a single $PM_{2.5}$ estimation for each of the 234.429 1 km grid cells in each of 4,018 days, totaling around 950 million data points. The methodology provides a complete spatial coverage, high resolution, and relative small error of the predictions, coupled with and the ability to capture variations in $PM_{2.5}$ concentrations across both spatial and temporal domains. The model offers a prediction accuracy that makes the output suitable for application in epidemiological studies on short and long-term health effects of air pollution.

The output from the empirical ML model developed in this study complements existing databases of modelled $PM_{2.5}$ in the United Kingdom, with some advantages for applications in epidemiological studies. Country-wide maps generated by emission-dispersion models are usually available at coarser spatial or temporal resolution [DEFRA, 2020; EMEP4UK, 2020; Savage et al., 2013, and generally they show lower small-scale accuracy when tested against observed monitoring data [Hood et al., 2018; Lin et al., 2017]. The spatio-temporal ML models presented here demonstrated comparable predictive performance to similar methods applied in other countries, based either on single-learner ML models [de Hoogh et al., 2018; Stafoggia et al., 2019], ensemble ML models [Chen

medRxiv preprint doi: https://doi.org/10.1101/2020.07.19.20157396; this version posted September 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

et al., 2018; Di et al., 2019; Yazdi et al., 2020], or generalised additive models (GAM) [Kloog et al., 2015]. Yazdi et al. (2020) developed an ensemble ML model (composed by RF, Deep Neural Network, GAM, Gradient Boosting, K-nearest Neighbour) to estimate $PM_{2.5}$ for Greater London, reaching mean 2005-2013 CV spatial-R^2 of 0.396. Using 2008-2013 period, this study reached CV spatial-R^2 of 0.637 for the whole Great Britain. Modelling a larger area might have provided more information and higher spatial variability, improving substantially the CV spatial-R^2. The application of a single learner to model air pollution in both spatial and temporal domains for Great Britain achieved a satisfactory performance. Nonetheless, ensemble model formats can be used in future applications using the same area of study and variables to assess how much performance gain is reached compared to the RF-only learner.

In 2008, a new Air Quality Directive (AQD) came into force bringing considerable changes to the following UK annual air quality assessment in 2010, setting an annual mean target of 25 $\mu g/m^3$ [Brookes et al., 2011]. Several policy controls and emission reductions had been put in place aiming to reduce from 2010 the road traffic PM emission by 83%, off-road mobile machinery by 54%, and energy production by 32% until 2020 [AQEG, 2013]. Independently of the temporal aggregation (daily or annual), all maps shown in this study detected this considerable drop in the $PM_{2.5}$ concentrations from 2010 across Great Britain.

Some limitations must be acknowledged. First, the multi-stage model relies on the extension of the observed series of $PM_{2.5}$ by predicting values based on co-located PM_{10} in Stage-1. This step was necessary for the application of the method in early years characterized by sparse $PM_{2.5}$ monitoring, which is likely to have contributed to the lower predictive performance in this period. In addition, the cross-validation procedures revealed the presence of some bias in the Stage-1 predictions, particularly in the spatial domain, probably linked to limitations in modelling the relative distribution of the two PM components. Second, while the model displayed a good performance throughout the period, the accuracy is worse in the temporal domain in summer (Table B1), suggesting limitations in capturing the higher temporal variation in this season. Third, the generalization of the prediction model was dependent on the selected locations of the monitors, which may be not representative of the study domain. This can result in an underestimation of the error and potentially biases in the predictions in more remote and less represented areas, if structural spatial differences are not entirely captured by the model covariates.

Future research directions for improving the current definition of the spatio-temporal ML model and its output can be discussed as well. First, availability of predictors that capture spatio-temporal variations is critical for ensuring high predictive ability. Improvements in emission-dispersion models such as EMEP4UK, whose output was consistently selected as the most important predictor, or in remote sensing satellite products, which in contrast show limited contributions at present, can be key factors for increasing the model performance. For instance, new instruments available in more recent earth observation satellites (*e.g.*, Copernicus Sentinel-5 Precursor) are explicitly developed for applications in pollution measurements and environmental health, and can be integrated in future versions of the model [ESA, 2020]. Second, spatially-lagged $PM_{2.5}$ variables, computed using IDW-LOOCV, were selected within the top ten-best predictors, suggesting that the performance can be extended further if the underlying spatial correlation structure is appropriately modelled. This can be achieved by extending the current spatial definition to fully spatio-temporal variables, and by applying more sophisticated spatial methods to incorporate such correlation. Third, published articles have presented alternative ML algorithms in single-learner [Just et al., 2018; Zan et al., 2017] or ensemble models [Di et al., 2019], which can be tested and applied in alternative to the RF method proposed here. Fourth, finally, previous studies have developed methods to increase the resolution of the predictions beyond the 1 km grid, for instance using additional stages involving small-area variables [de Hoogh et al., 2018; Di et al., 2019; Stafoggia et al., 2019]. Similar approached can be applied here.

DRAFT

medRxiv preprint doi: https://doi.org/10.1101/2020.07.19.20157396; this version posted September 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

5. Conclusions

This study developed and applied a multi-stage ML model, combining data for multiple sources, including remote sensing satellite products, climate and atmospheric reanalysis models, chemical transport models, and geospatial features, to generate a complete map of daily $PM_{2.5}$ concentrations in 1 km grid across Great Britain in the period 2008-2018. The model showed good performance overall and in both spatial and temporal domains, with an accuracy that is compatible with the use of such reconstructed values as proxy for $PM_{2.5}$ exposures in epidemiological studies. In particular, the availability of high-resolution measures that can be linked as such or aggregated at different spatial and/or temporal scales makes the output suitable for investigations on both transient and chronic health risks associated to short and long-term exposures to $PM_{2.5}$, respectively.

Funding: This study was supported by the Medical Research Council-UK (Grant ID: MR/M022625/1), the Natural Environment Research Council UK (Grant ID: NE/R009384/1), and the European Union's Horizon 2020 Project Exhaustion (Grant ID: 820655). EMEP4UK Model results and contributions by S.R. and M.V. were supported by award number NE/R016429/1 as part of the UK-SCAPE programme delivering National Capability.

Acknowledgments: The authors are grateful for the technical support received from the European Centre for Medium-Range Weather Forecasts (ECMWF). The authors also would like to thank the European Space Research Institute (ESRIN) from the European Space Agency (ESA) for their feedback on this study. This research had free access to all data sources: (i) NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC), (ii) NOAA National Centers for Environmental Information (NCEI), (iii) Copernicus Land Monitoring Service (CLMS), (iv) Copernicus Atmosphere Data Store, (v) Copernicus Climate Data Store, (vi) EMEP4UK, (vii) UK Department for Environment, Food & Rural Affairs.

References

Air Quality Expert Group (AQEG). 2013. Mitigation of United Kingdom PM2.5 Concentrations. Available online: https://uk-air.defra.gov.uk/assets/documents/reports/cat11/1508060903_DEF-PB14161_Mitigation_of_UK_PM25.pdf (accessed on 6 July 2020).

Basagaña, X. , Jacquemin, B. , Karanasiou, A. , Ostro, B. , Querol, X. , Agis, D. , Alessandrini, E. , Alguacil, J. , Artiñano, B. , Catrambone, M. , De La Rosa, J.D. , Díaz, J. , Faustini, A. , Ferrari, S. , Forastiere, F. , Katsouyanni, K. , Linares, C. , Perrino, C. , Ranzi, A. , Ricciardelli, I. , Samoli, E. , Zauli-Sajani, S. , Sunyer, J. , Stafoggia, M. , Alessandrini, E. , Angelini, P. , Berti, G. , Bisanti, L. , Cadum, E. , Catrambone, M. , Chiusolo, M. , Davoli, M. , De'Donato, F. , Demaria, M. , Gandini, M. , Grosa, M. , Faustini, A. , Ferrari, S. , Forastiere, F. , Pandolfi, P. , Pelosini, R. , Perrino, C. , Pietrodangelo, A. , Pizzi, L. , Poluzzi, V. , Priod, G. , Randi, G. , Ranzi, A. , Rowinski, M. , Scarinzi, C. , Stafoggia, M. , Stivanello, E. , Zauli-Sajani, S. , Dimakopoulou, K. , Elefteriadis, K. , Katsouyanni, K. , Kelessis, A. , Maggos, T. , Michalopoulos, N. , Pateraki, S. , Petrakakis, M. , Rodopoulou, S. , Samoli, E. , Sypsa, V. , Agis, D. , Alguacil, J. , Artiñano, B. , Barrera-Gómez, J. , Basagaña, X. , De La Rosa, J. , Diaz, J. , Fernandez, R. , Jacquemin, B. , Karanasiou, A. , Linares, C. , Ostro, B. , Perez, N. , Pey, J. , Querol, X. , Salvador, P. , Sanchez, A.M. , Tobias, A. , Bidondo, M. , Declercq, C. , Le Tertre, A. , Lozano, P. , Medina, S. , Pascal, L. and Pascal, M. Short-term effects of particulate matter constituents on daily hospitalizations and mortality in five South-European cities: Results from the MED-PARTICLES project. Environment International, 2015, volume75, pages 151-158. Doi: 10.1016/j.envint.2014.11.011.

Beckerman, B.S., Jerrett, M., Martin, R.V., van Donkelaar, A., Ross, Z., and Burnett, R.T. Application of the deletion/substitution/addition algorithm to selecting land use regression models for interpolating air pollution measurements in California. Atmos. Environ.,2013, Volume 77, Pages 172-177. Doi: 10.1016/j.atmosenv.2013.04.024

Brookes, D.M., Stedman, J.R., Grice, S.E., Kent, A.J., Walker, H.L., Cooke, S.L., Vincent, K.J., Lingard, J.J.N., Bush, T.J. and Abbott, J. (2011). UK air quality modelling under the Air Quality Directive (2008/50/EC) for 2010 covering the following air quality pollutants: SO2, NOx, NO2, PM10, PM2.5, lead, benzene, CO, and ozone. Report for the Department for Environment, Food and Rural Affairs (Defra), Welsh Government, Scottish Government and the Department of the Environment in Northern Ireland. AEA report. AEAT/ENV/R/3215 Issue 1. Available online: http://uk-air.defra.gov.uk/reports/cat09/1204301513_AQD2010mapsrep_master_v0.pdf (accessed on 6 July 2020).

medRxiv preprint doi: https://doi.org/10.1101/2020.07.19.20157396; this version posted September 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

DEFRA - Department for Environment, Food & Rural Affairs. 2012. Fine Particulate Matter (PM2.5) in the UK. Available online: https://www.gov.uk/government/publications/fine-particulate-matter-pm2-5-in-the-uk (Accessed on 25 May, 2020)

DEFRA. Modelled background pollution data. Available online: https://uk-air.defra.gov.uk/data/pcm-data (accessed on 13 July 2020).

Bozzo, A., S. Remy, A. Benedetti, J.Flemming, P. Bechtold, M.J. Rodwell, and J.-J. Morcrette, 2017: Implementation of a CAMS-based aerosol climatology in the IFSA. ECMWF Technical Memorandum 801, 33 pp, https://www.ecmwf.int/sites/default/files/elibrary/2017/17219-implementation-cams-based-aerosol-climatology-ifs.pdf.

Chen G., Li S., Knibbs L.D., Hamm N., Cao W., Li T., Guo J., Ren H., Abramson M.J., Guo Y. 2018. A machine learning method to estimate PM2.5 concentrations across China with remote sensing meteorological and land use information. Sci.TotalEnviron, 636, 52–60. https://doi.org/10.1016/j.scitotenv.2018.04.251.

Chen, Z.Y., Zhang, T.H., Zhang, R., Zhu, Z.M., Yang, J., Chen, P.Y., Ou, C.Q., Guo, Y. 2019. Extreme gradient boosting model to estimate PM2.5 concentrations with missing-filled satellite data in China. Atmospheric Environment, 202, pages 180-189. Doi: 10.1016/j.atmosenv.2019.01.027

Civil Aviation Authority (CAA). Available online: caa.co.uk/home (accessed on 29 May 2020).

Copernicus Land Monitoring Service (CLMS). Available online https://land.copernicus.eu/pan-european. (Acessed on 29 May, 2020)

de Hoogh, K., Héritier, H., Stafoggia, M., Künzli, N., and Kloog, I. Modelling daily PM2.5 concentrations at high spatio-temporal resolution across Switzerland. Environ., Pollut. 2018, 233, 1147–1154. Doi: 10.1016/j.envpol.2017.10.025

Didan, K. (2015). MOD13A3 MODIS/Terra vegetation Indices Monthly L3 Global 1km SIN Grid V006 [Data set]. NASA EOSDIS LP DAAC, doi: 10.5067/MODIS/MOD13A3.006.

Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M. B., Choirat, C., Koutrakis, P., Lyapustin, A., Wang, Y., Mickley, L. J., & Schwartz, J. (2019). An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. Environment international, 130, 104909. https://doi.org/10.1016/j.envint.2019.104909

Digimap. Available online: https://digimap.edina.ac.uk/webhelp/os/data_information/os_data_issues/grid_references.htm (Accessed on 1 April, 2020)

Di, Q.; Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M.B., Choirat, C., Koutrakis, P., Lyapustin, A., Wang, Y., Mickley, L.J., and Schwartz, J. An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. Environ. Int., 2019, 130, 104909. Doi: 10.1016/j.envint.2019.104909

Earth Observation Group (EOG). Available online: https://ngdc.noaa.gov/eog/viirs/download_dnb_composites.html (accessed on 1 July 2020)

ERA 5 global climate reanalysis. Available online: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview (Accessed on 28 May, 2020)

ERA 5 Land global climate reanalysis. Available online: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview (Accessed on 28 May, 2020)

European Modelling and Evaluation Programme for the UK (EMEP4UK). Available online: http://www.emep4uk.ceh.ac.uk/ (accessed on 13 July 2020)

European Space Agency. 2020. Copernicus Sentinel-5 Precursor mission. Available online: https://sentinel.esa.int/web/sentinel/missions/sentinel-5p. (Accessed on 13 July, 2020)

Gupta, P. and Christopher, S.A. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: multiple regression approach. J. Geophys. Res. Atmos., 2009, Volume114, Issue D14. Doi: 10.1029/2008JD011497

Hood, C., MacKenzie, I., Stocker, J., Johnson, K., Carruthers, D., Vieno, M., and Doherty, R.: Air quality simulations for London using a coupled regional-to-local modelling system. Atmos. Chem. Phys., 18, 11221–11245, https://doi.org/10.5194/acp-18-11221-2018, 2018.

Hu, X., Belle, J.H., Meng, X., Wildani, A., Waller, L.A., Strickland, M.J., and Liu, Y. Estimating PM2. 5 concentrations in the conterminous United States using the random forest approach. Environ. Sci. Technol. 2017, 51 (12), 6936–6944. Doi: 10.1021/acs.est.7b01210.

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning, vol 6. Springer, Berlin

DRAFT

medRxiv preprint doi: https://doi.org/10.1101/2020.07.19.20157396; this version posted September 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Just, A.C.; De Carli, M.M.; Shtein, A.; Dorman, M.; Lyapustin, A.; Kloog, I. Correcting Measurement Error in Satellite Aerosol Optical Depth with Machine Learning for Modeling PM2.5 in the Northeastern USA. Remote Sens. 2018, 10, 803. Doi: 10.3390/rs10050803

Kloog, I., Koutrakis, P., Coull, B.A., Lee, H.J., and Schwartz, J. Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements. Atmospheric Environment, 2011, Volume 45, Issue 35, Pages 6267-6275. Doi: 10.1016/j.atmosenv.2011.08.066

Kloog, I., Sorek-Hamer, M., Lyapustin, A., Coull, B., Wang, Y., Just, A.C., Schwartz, J., Broday, D.M. Estimating daily PM2.5 and PM10 across the complex geo-climate region of Israel using MAIAC satellite-based AOD data. Atmospheric Environment, 2015, 122, 409-416. Doi: 10.1016/j.atmosenv.2015.10.004

Koelemeijer, R., Homan, C., and Matthijsen, J. Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over Europe. Atmos. Environ., 2006, 40 (27), pp. 5304-5315. Doi: 10.1016/j.atmosenv.2006.04.044.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., and Rubel, F. (2006). World Map of the Köppen-Geiger climate classification updated. Meteorologische Zeitschrift, 15, 259-263, doi: 10.1127/0941-2948/2006/0130.

Liu et al., 2019 - Liu C, Chen R, Sera F, Vicedo-Cabrera AM, Guo Y, Tong S, Coelho MSZS, Saldiva PHN, Lavigne E, Matus P, Valdes Ortega N, Osorio Garcia S, Pascal M, Stafoggia M, Scortichini M, Hashizume M, Honda Y, Hurtado-Diaz M, Cruz J, Nunes B, Teixeira JP, Kim H, Tobias A, Iniguez C, Forsberg B, Astrom C, Ragettli MS, Guo YL, Chen BY, Bell ML, Wright CY, Scovronick N, Garland RM, Milojevic A, Kysely J, Urban A, Orru H, Indermitte E, Jaakkola JJK, Ryti NRI, Katsouyanni K, Analitis A, Zanobetti A, Schwartz J, Chen J, Wu T, Cohen A, Gasparrini A, Kan H. Ambient particulate air pollution and daily mortality in 652 cities. New England Journal of Medicine. 2019;381(8):705-715. DOI: 10.1056/NEJMoa1817364

Lavigne E, Donelle J, Hatzopoulou M, Van Ryswyk K, van Donkelaar A, Martin RV, Chen H, Stieb DM, Gasparrini A, Crighton E, Yasseen III AS, Burnett RT, Walker M, Weichenthal S. Spatiotemporal variations in ambient ultrafine particles and the incidence of childhood asthma. American Journal of Respiratory and Critical Care Medicine. 2019a ;199(12):1487-1495. Doi: 10.1164/rccm.201810-1976OC

Lavigne E, Lima I, Hatzopoulou M, Van Ryswyk K, Decou ML, Luo W, van Donkelaar A, Martin RV, Chen H, Stieb DM, Crighton E, Gasparrini A, Elten M, Yasseen AS 3rd, Burnett RT, Walker M, Weichenthal S. Spatial variations in ambient ultrafine particle concentrations and risk of congenital heart defects. Environment International. 2019b, 130:104953. Doi: 10.1016/j.envint.2019.104953

Lee, H., Liu, Y., Coull, B., Schwartz, J., and Koutrakis, P. A novel calibration approach of MODIS AOD data to predict PM2. 5 concentrations. Atmos. Chem. Phys., 2011, 11, 7991–8002. Doi: 10.5194/acp-11-7991-2011

Lin, C., Heal, M. R., Vieno, M., MacKenzie, I. A., Armstrong, B. G., Butland, B. K., Milojevic, A., Chalabi, Z., Atkinson, R. W., Stevenson, D. S., Doherty, R. M., and Wilkinson, P.: Spatiotemporal evaluation of EMEP4UK-WRF v4.3 atmospheric chemistry transport simulations of health-related metrics for NO2, O3, PM10, and PM2. 5 for 2001–2010, Geosci. Model Dev., 10, 1767–1787, https://doi.org/10.5194/gmd-10-1767-2017, 2017.

Lyapustin, A., Y. Wang. MCD19A2 MODIS/Terra+Aqua Land Aerosol Optical Depth Daily L2G Global 1km SIN Grid V006. 2018, distributed by NASA EOSDIS Land Processes DAAC, https://doi.org/10.5067/MODIS/MCD19A2.006. Accessed 2020-05-28.

NASA Earth Observations. Available online: https://neo.sci.gsfc.nasa.gov/view.php?datasetId=MODAL2_M_AER_OD (accessed on 20 March 2020)

Office for National Statistics (ONS). Available online https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates (accessed on 1 April 2020)

Openair R package. Available online: https://cran.r-project.org/web/packages/openair/openair.pdf (accessed on 25 May 2020)

Ordnance Survey Open Roads. Available online: https://www.ordnancesurvey.co.uk/documents/os-open-roads-user-guide.pdf. (accessed on 29 May 2020)

Polley, E.C., Rose, S., and van der Laan, M.J. 2011. "Super Learning." In M.J. van der Laan and S. Rose, Targeted Learning: Causal Inference for Observational and Experimental Data, Chapter 3. New York, Springer.

Raaschou-Nielsen, O. and Beelen, R. and Wang, M. and Hoek, G. and Andersen, Z.J. and Hoffmann, B. and Stafoggia, M. and Samoli, E. and Weinmayr, G. and Dimakopoulou, K. and Nieuwenhuijsen, M. and Xun, W.W. and Fischer, P. and Eriksen, K.T. and S?rensen, M. and Tj?nneland, A. and Ricceri, F. and de Hoogh, K. and Key, T. and Eeftens, M. and Peeters, P.H. and Bueno-de-Mesquita, H.B. and Meliefste, K. and

medRxiv preprint doi: https://doi.org/10.1101/2020.07.19.20157396; this version posted September 17, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Oftedal, B. and Schwarze, P.E. and Nafstad, P. and Galassi, C. and Migliore, E. and Ranzi, A. and Cesaroni, G. and Badaloni, C. and Forastiere, F. and Penell, J. and De Faire, U. and Korek, M. and Pedersen, N. and Östenson, C.-G. and Pershagen, G. and Fratiglioni, L. and Concin, H. and Nagel, G. and Jaensch, A. and Ineichen, A. and Naccarati, A. and Katsoulis, M. and Trichpoulou, A. and Keuken, M. and Jedynska, A. and Kooter, I.M. and Kukkonen, J. and Brunekreef, B. and Sokhi, R.S. and Katsouyanni, K. and Vineis, P. Particulate matter air pollution components and risk for lung cancer. Environment International, 2016, volume87, pages 66-73. Doi: 10.1016/j.envint.2015.11.007.

Savage, N.H., Agnew, P., Davis, L.S., Ordonez, C., 2013. Air quality modelling using the Met Office Unified Model (AQUM OS24-26): model description and initial evaluation. Geosci. Model Dev. (GMD) 6, 353–372.https://doi.org/10.5194/gmd-6-353-2013.

Schneider dos Santos, R. Estimating spatio-temporal air temperature in London (UK) using machine learning and earth observation satellite data. International Journal of Applied Earth Observation and Geoinformation, Volume 88, June 2020, 102066. Doi: 10.1016/j.jag.2020.102066.

Stafoggia, M., Schwartz, J., Badaloni, C., Bellander, T., Alessandrini, E., Cattani, G., de' Donato, F., Gaeta, A., Leone, G., Lyapustin, A., Sorek-Hamer, M., de Hoogh, K., Di, Q., Forastiere, F, Kloog, I. 2016. Estimation of daily PM10 concentrations in Italy (2006–2012) using finely resolved satellite data, land use variables and meteorology. Environ. Int. 99, 234–244.

Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., de Hoogh, K., de' Donatoa, F., Gariazzo, C., Lyapustinf, A., Michelozzi, P., Renzi, M., Scortichini, M., Shtein A., Viegi, G., Kloog, I., and Schwartz, J. Estimation of daily PM10 and PM2.5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. Environment International, 2019, Volume 124, Pages 170-179. Doi: 10.1016/j.envint.2019.01.016

UERRA regional reanalysis. Available online: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview (accessed on 28 May, 2020)

UK Data Service. Available online: https://www.ukdataservice.ac.uk/ (accessed on 29 May 2020)

Van Donkelaar A., Martin R.V., Brauer M., Kahn R., Levy R., Verduzco C., and Villeneuve, P.J. 2010. Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application. Environ Health Perspect, 118(6), pp.847–55. Doi: 10.1289/ehp.0901623

Vienneau, D., de Hoogh, K., Beelen, R., Fischer, P., Hoek, G., and Briggs, D. Comparison of land-use regression models between Great Britain and the Netherlands. Atmos. Environ., 2010, Volume 44, issue 5, 688-696. Doi: doi.org/10.1016/j.atmosenv.2009.11.016

Vieno, M., Dore, A. J., Stevenson, D. S., Doherty, R., Heal, M. R., Reis, S., Hallsworth, S., Tarrason, L., Wind, P., Fowler, D., Simpson, D., and Sutton, M. A.: Modelling surface ozone during the 2003 heat-wave in the UK, Atmos. Chem. Phys., 10, 7963–7978, doi:10.5194/acp-10-7963-2010, 2010

Yazdi, M.D., Kuang, Z., Dimakopoulou, K., Barratt, B., Suel, E., Amini, H., Lyapustin, A., Katsouyanni, K., and Schwartz, J. Predicting Fine Particulate Matter (PM2.5) in the Greater London Area: An Ensemble Approach using Machine Learning Methods. Remote Sensing, 2020, 12, 914, 1-18. Doi:10.3390/rs12060914 www.mdpi

WHO, 2020 - Word Health Organization. Available online: https://www.who.int/health-topics/air-pollution#tab=tab_1 (accessed on 20 March 2020).

Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M.L., Shen, X., Zhu, L., Zhang, M., 2017. Spatiotemporal prediction of continuous daily PM2.5 concentrations across China using a spatially explicit machine learning algorithm. Atmos. Environ. 155, 129–139. https://doi.org/10.1016/j.atmosenv.2017.02.023.

DRAFT