Full length article

Europe-wide air pollution modeling from 2000 to 2019 using geographically weighted regression

Youchen Shen [a,*], Kees de Hoogh [b,c], Oliver Schmitz [d], Nicholas Clinton [e],
Karin Tuxen-Bettman [e], Jørgen Brandt [f], Jesper H. Christensen [f], Lise M. Frohn [f], Camilla Geels [f],
Derek Karssenberg [d], Roel Vermeulen [a,g], Gerard Hoek [a]

[a] *Division of Environmental Epidemiology, Institute for Risk Assessment Sciences, Utrecht University, Utrecht, the Netherlands*
[b] *Swiss Tropical and Public Health Institute, Basel, Switzerland*
[c] *University of Basel, Basel, Switzerland*
[d] *Department of Physical Geography, Faculty of Geosciences, Utrecht University, Utrecht, the Netherlands*
[e] *Google, Inc, Mountain View, CA, United States*
[f] *Department of Environmental Science, Aarhus University, Roskilde, Denmark*
[g] *Julius Centre for Health Sciences and Primary Care, University Medical Centre, Utrecht University, Utrecht, the Netherlands*

ARTICLE INFO

ABSTRACT

Previous European land-use regression (LUR) models assumed fixed linear relationships between air pollution concentrations and predictors such as traffic and land use. We evaluated whether including spatially-varying relationships could improve European LUR models by using geographically weighted regression (GWR) and random forest (RF). We built separate LUR models for each year from 2000 to 2019 for NO_2, O_3, $PM_{2.5}$ and PM_{10} using annual average monitoring observations across Europe. Potential predictors included satellite retrievals, chemical transport model estimates and land-use variables. Supervised linear regression (SLR) was used to select predictors, and then GWR estimated the potentially spatially-varying coefficients. We developed multi-year models using geographically and temporally weighted regression (GTWR). Five-fold cross-validation per year showed that GWR and GTWR explained similar spatial variations in annual average concentrations (average R^2 = NO_2: 0.66; O_3: 0.58; PM_{10}: 0.62; $PM_{2.5}$: 0.77), which are better than SLR (average R^2 = NO_2: 0.61; O_3: 0.46; PM_{10}: 0.51; $PM_{2.5}$: 0.75) and RF (average R^2 = NO_2: 0.64; O_3: 0.53; PM_{10}: 0.56; $PM_{2.5}$: 0.67). The GTWR predictions and a previously-used method of back-extrapolating 2010 model predictions using CTM were overall highly correlated ($R^2 > 0.8$) for all pollutants. Including spatially-varying relationships using GWR modestly improved European air pollution annual LUR models, allowing time-varying exposure-health risk models.

1. Introduction

Ambient air pollution contributes to around 7 million deaths mainly from non-communicable diseases (World Health Organization, 2021). To better understand the health effects of air pollution exposure, cohort studies are increasingly conducted in large study areas (Brauer et al., 2019; Di et al., 2017; Stafoggia et al., 2022). These cohort studies had follow-up periods of 10–20 years, with recruitment dating back to mid 1990s. To allow time-varying exposure analyses, these studies require annual exposure estimates harmonized for the study area during the follow-up period.

For the ELAPSE project (Effects of Low-Level Air Pollution: A Study

in Europe) (de Hoogh et al., 2018a), we developed Europe-wide land use regression (LUR) models for nitrogen dioxide (NO_2), ozone (O_3), particulate matter less than 2.5 μm ($PM_{2.5}$), and black carbon for a single year (2010) using supervised linear regression (SLR). Chemical transport model (CTM) estimates at a large spatial resolution (50 km × 50 km) for multiple years were used to back- and forward-extrapolate estimates from the LUR model built for the year 2010 to earlier and later years where LUR models were unavailable.

An important assumption, in the Europe-wide model (de Hoogh et al., 2018a) and other large-scale models (Bechle et al., 2015; Chen et al., 2019; Knibbs et al., 2014; Larkin et al., 2017; Lu et al., 2020), is that the relationship between air pollution and predictors such as traffic,

population density, and meteorology can be modeled with fixed co-efficients. However, the relationships between air pollution concentrations and predictors could differ across countries. For example, the same road density might be associated with higher NO_2 concentrations in southern Europe than in northern Europe, because the road density only serves as a proxy for traffic intensity and/or traffic-related emission sources, which can be varying across countries/regions. Geographically Weighted Regression (GWR) is a technique that allows for spatially-varying relationships between the predictors and air pollution concentrations. In previous work in North America (Van Donkelaar et al., 2015) and the globe (Hammer et al., 2020), GWR was used to adjust for spatial heterogeneity in the bias between surface concentrations calculated from satellite-derived observations and a large-scale chemical transport model (geophysical $PM_{2.5}$ estimate) and ground-based $PM_{2.5}$ observations. These studies have shown that GWR greatly improved the five-fold cross-validation (CV) accuracy compared to unadjusted predictions. Building on these studies, we evaluate whether it would improve air pollution predictions by incorporating spatial heterogeneity directly in the relationship between ground-based observations and several spatial variables in Europe.

Air pollution concentrations have generally been reduced in the past decades in Europe (Ortiz and Guerreiro, 2020). To obtain air pollution predictions for multiple years, explicit models for these years would likely improve upon back-extrapolating a single-year model with modeled spatio-temporal trends. It is however not clear whether developing specific models for every single year (single-year model) or developing models for groups of multiple years (multi-year model) is the most effective approach. We explored the application of geographically and temporally weighted regression (GTWR) (Fotheringham et al., 2015; Huang et al., 2010; Wu et al., 2014) to develop models for multiple years. GTWR is an extension of GWR, allowing both spatially- and temporally-varying relationships between air pollution and predictors. Incorporating temporally-varying relationships is important as we would avoid assuming that the quantitative relationship between a predictor and air pollution concentrations remains fixed over decades. Previously, GTWR was used at a regional scale in China to estimate ground-level daily $PM_{2.5}$ concentrations using satellite-derived aerosol optical depth (AOD) values and other spatial variables (Bai et al., 2016; He and Huang, 2018). They showed that GTWR performed better than GWR at a daily scale. Building on these studies, we evaluate whether model performance is also improved in modeling annual average concentrations for a long (20-year) time period at the continental (European) scale.

Recently, machine-learning methods, including random forest (RF), have been applied in developing LUR models (Chen et al., 2019; Kerckhoffs et al., 2019; Lu et al., 2020). RF models relax the linear assumption in linear regression methods and allow interactions between variables. Potentially, RF could therefore also include spatial heterogeneity, e.g. by using kriging to explain its residuals (Zhan et al., 2018) or by including climate zones or geographical coordinates as a predictor (Hengl et al., 2018).

We built annual LUR models in Europe from 2000 to 2019 for four regulated air pollutants: NO_2, O_3, particulate matter < 10 μm (PM_{10}), and $PM_{2.5}$. Our first aim was to evaluate potential improvements by including spatially-varying relationships in Europe-wide LUR models. Our second aim was to compare single-year and multi-year modeling approaches. Our third aim was to compare our annual LUR models with our previous back-extrapolation methods (de Hoogh et al., 2018a; Gulliver et al., 2013). The novelty of our work is to evaluate model performance of GWR and GTWR in modeling relationships between ground-based observations and spatial predictors; to include multiple major pollutants with different spatio-temporal patterns; to compare with a machine-learning method; and the spatial resolution of 25 m at the continental European scale.

2. Materials and methods

Our models extended our previous Europe-wide models (de Hoogh et al., 2018a) by evaluating the spatial variations in the coefficients of the LUR model; by developing models for a large number of years; by improving spatial resolution from 100 m × 100 m to 25 m × 25 m; by extending the domain to include Eastern European countries; by incorporating time-varying predictors when available; and by incorporating more predictors such as meteorology.

We collected routine-monitoring data of air pollution and spatio-temporal predictors to estimate air pollution concentrations across Europe from 2000 to 2019. We developed models to estimate annual average air pollution exposure because we aimed to apply the model predictions in studying health effects of long-term air pollution exposure. As the health studies started recruitment in the past, the most important objective was to develop models for current and historical long-term exposure at a fine spatial resolution. Some recent studies have developed daily air pollution models for multiple years with good performance (de Hoogh et al., 2019, 2018b; Di et al., 2019; Liu et al., 2020; Requia et al., 2020; Shtein et al., 2018) at a spatial of 1 km × 1 km for the US studies and about 12 km × 12 km for the Chinese studies. As our objective was to develop long-term exposure at a fine spatial resolution and as we do not need daily maps, we did not develop daily models.

The data served as input to LUR models built by four algorithms in two temporal settings: 1) SLR for single-year and multi-year, 2) GWR for single-year, GTWR for multi-year, and 3) RF for single-year and multi-year. We refer to single-year as training a single model per year using data from that specific year and refer to multi-year as training a single model for consecutive years using all data from 2000 to 2019. We also evaluated shorter multi-year periods, specifically 3-to-6-year periods. Model performance was evaluated by 5-fold cross-validation (CV) and by an external validation dataset obtained in the ESCAPE project (European Study of Cohorts for Air pollution Effects) collected across Europe in 2010 (Cyrys et al., 2012; Eeftens et al., 2012).

We built the models from 2000 to 2019, because of two reasons. Firstly, some potential predictors were only available after 2000 such as satellite-derived data. Secondly, the monitoring observations were highly clustered in specific countries and were limited in quantity before 2000 especially for $PM_{2.5}$ (with less than 10 observations) and PM_{10} (with less than 450 observations), as shown in Figs. S1 & S2.

2.1. Air pollution monitoring observations

We collected routine monitoring data for NO_2, O_3, PM_{10}, and $PM_{2.5}$ from the European Environmental Agency (EEA). We collected the observations before 2012 from the Airbase database v8 (European Environmental Agency, 2020a) and after 2012 from the Air Quality e-Reporting database (European Environmental Agency, 2020b).

For PM_{10} and $PM_{2.5}$, we aggregated daily observations to yearly averages, because most observations for PM were only available at a daily scale (integrated filter-based methods or continuous methods with insufficient hourly precision). NO_2 and O_3 were measured with continuous methods and available as hourly averages. For NO_2, we aggregated hourly observations to annual averages. For O_3, we calculated the daily maximum 8-hour mean for each day from hourly observations and then aggregated the daily values to annual averages. This statistic of daily maximum 8-hour mean was chosen because it is used in regulatory guidelines (WHO, 2021). All annual statistics were only used if more than 75% of the daily or hourly observations were valid as defined by the EEA. We did not train the single-year model if the annual average observations were available from less than 200 monitoring sites across Europe for that specific year. This only applied to the $PM_{2.5}$ observations before 2006. From 2000 to 2019, the number of monitoring sites with more than 75% annual validity grows from 1533 to 3176 for NO_2, from 1243 to 1954 for O_3, from 445 to 2937 for PM_{10}, and from 11 to 1433 for $PM_{2.5}$ (Table S2). PM monitoring in the earlier years was also limited to

specific countries such as Germany, the Netherlands, and UK (Figs. S1 & S2). After around 2010, monitoring sites were spread sufficiently and widely for all pollutants (Figs. S1 & S2). The observed air pollution concentrations have decreased substantially in the past decade (Table S2), except for O_3. The decreasing trend in annual averages is not linear for PM, because the complex interactions between emission sources and meteorological conditions might affect the trend in annual averages across Europe. Especially for $PM_{2.5}$ the trend could also be affected by the different set of monitors included in different years.

2.2. Predictor variables

We calculated several road, land-use, satellite retrievals, and chemical transport model estimates to use them as predictors in the LUR models (Table S1). Predictors were similar to the ones used in the ELAPSE study (de Hoogh et al., 2018a), supplemented with extra time-varying spatial predictors such as meteorology chemical transport model estimates, and satellite retrievals. These time-varying spatial predictors could help capturing regional changes in annual average concentrations over time (Gulliver et al., 2013). We did not include industrial emission data as predictors, because emission data was already included in the chemical transport models (Brandt et al., 2012; Marécal et al., 2015) and we used the model estimates as a predictor.

To capture the dispersion of the air pollution, several variables were calculated in several circular buffer sizes (ranging from 50 m to 10,000 m), such as land use data and road data (Table S1). After the buffered values of land cover data and road data were calculated, all predictor maps were regridded to a 25-m resolution using nearest neighbor resampling. Most of these predictors were processed in Google Earth Engine (GEE) (Gorelick et al., 2017) except for road data.

We calculated road predictors from OpenStreetMap data (OpenStreetMap contributors, 2020). OSM is a crowd-sourced project, and the robustness of the OSM data highly depends on the data contributed by the users. We obtained the OSM road data in 2020 because the robustness and validity of OSM show large spatial discrepancies back in time (more than 5–8 years ago) (Girres and Touya, 2010; Neis et al., 2011). Although road network changes over time, we focused on capturing the representative of the road network instead of the temporal changes in the network, as most of the network stays unchanged in most developed countries in Europe. From OSM, we extracted motorways, primary, and local roads. Each road segment in each road type was reprojected to ETRS89-LAEA and then was intersected with 25-m grids. Road lengths per grid cell were calculated and then summed to obtain the buffered road predictors. We finally calculated predictors in buffer sizes ranging from 50 m to 10,000 m (Table S1). Other detailed descriptions of all predictor variables are in Table S1.

2.3. Algorithms for training LUR models

We used four algorithms (i.e., SLR, GWR, GTWR, RF) to train our annual LUR models. Details on the technical implementation of GWR and GTWR are in the Supporting Information.

2.3.1. Supervised linear regression (SLR)

SLR is a standardized approach for including the most informative variables stepwise and it has been widely used for LUR modeling (Chen et al., 2021; de Hoogh et al., 2018a; Eeftens et al., 2012a). We followed the ESCAPE protocol (Eeftens et al., 2012a) to train SLR. In short, firstly, a linear regression model was built using the predictor variable that explains the most variation in the concentrations (highest R^2 value) among all variables. In subsequent steps, additional predictor variables were allowed to enter the model if they improved the adjusted R^2; if the sign of the coefficient met the predefined direction of effect (stated in Table S1); and if the coefficient value was statistically significant (p < 0.1). Finally, a predictor variable was excluded if its variance inflation factor (VIF) was larger than 3 to avoid multicollinearity and this

exclusion step stopped when no variables had VIF larger than 3. We built the single-year SLR model for each year, and therefore the model for each year could have different variables selected. We built one multi-year SLR model using data from 2000 to 2019, and after variables were selected, year was included as a predictor to vary the intercept value of the regression model for each year. Year was only included as a predictor after the variables were selected because year would not necessarily be selected in the step-wise procedure in SLR.

2.3.2. Geographically weighted regression (GWR)

GWR (Brunsdon et al., 1996) includes spatial heterogeneity by using spatially-varying coefficient values. The coefficient values were estimated by using a weight function. The weight function gives higher weights to observation points closer to the estimate points than points further away. The function decays with spatial Euclidean distance according to a predefined kernel function and bandwidth. The kernel function determines the shape of the relation between distance and weight, and the bandwidth controls how fast the kernel function decays in space. For the kernel function, we used an exponential function to give an abrupt decrease in weight with increasing distance. Due to the heterogeneous spatial distribution of the air pollution observations across Europe, we chose an adaptive bandwidth (nngb) to ensure sufficient local information is included. We used 200 km × 200 km grids for the spatially-varying coefficient values of the linear regression. We used functions in the *GWmodel* package version 2.2–4 (Gollini et al., 2015; Lu et al., 2014) in R (R Core Team, 2020) to train GWR. GWR can tackle the spatial heterogeneity in the relationships between predictors and air pollution concentrations. The detailed technical implementation of GWR is in the Supporting Information.

2.3.3. Geographically and temporally weighted regression (GTWR)

GTWR (Fotheringham et al., 2015; Huang et al., 2010; Wu et al., 2014) is an extension of GWR, allowing both spatial and temporal heterogeneity of the relationships between air pollution and predictors. GTWR uses a weight function to estimate the spatiotemporally-varying coefficient values of the regression model, and the weight function includes both spatial and temporal distances. In short, it assumes that observation points closer in the spatiotemporal distance have a higher impact on local coefficient estimates than observation points further away, and one-meter spatial distance has a different impact compared to one-year temporal distance. We used functions in the *GWmodel* package version 2.2–4 (Gollini et al., 2015; Lu et al., 2014) in R (R Core Team, 2020) to train GTWR. We first used SLR to select the informative variables and then used GTWR to estimate the potential spatially- and temporally-varying coefficient values. Some of the original settings in GTWR are unsuitable for air pollution modeling (Wu et al., 2014) and therefore we adjusted some parameter settings. First, the original GTWR only uses the observations from prior time steps to estimate the GTWR coefficients for the current time step. But we used observations from both prior time steps and subsequent time steps to train GTWR for the current time step, because we assumed that observations from the past and subsequent years could improve the concentration predictions. Second, we included more flexibility in modeling spatial and temporal distances by adding a conversion factor. We tuned the parameters using 5-fold CV. The detailed technical implementation of GTWR is in the Supporting Information.

2.3.4. Random forest (RF)

RF is an ensemble tree-based machine learning method (Breiman, 2001). Previous studies showed that RF gave similar or superior model performance for air pollution compared to linear regression or spatial interpolation methods (Chen et al., 2020, 2019; Kerckhoffs et al., 2021, 2019; Li et al., 2011; Lu et al., 2020). RF can deal with highly correlated variables by randomly selecting a subset of variables in each split node of a tree. The regression trees are built using bootstrapping training data (i.e., by sampling training data with replacement), and thus not all

training data is used in each regression tree. The unused data is often referred to as out-of-bag (OOB) data. We used the overall OOB root mean squared error (RMSE) to optimize RF hyperparameters.

The optimized hyperparameters that led to the least OOB RMSE were used in our RF. We optimized the hyperparameters using a grid search method. We only optimized the number of trees (*ntree*) and the number of variables being split at each node (*mtry*) due to their higher effect on model performance than other hyperparameters (Lu et al., 2020). We optimized the hyperparameters for every year (single-year) or between 2000 and 2019 (multi-year) using the grid searching process with *ntree* ranging from 200 to 800 with an increment of 200 and *mtry* ranging from the squared number of the predictors to the total number of the predictors with an increment of 20. We used functions in the *ranger* package version 0.12.1 (Wright and Ziegler, 2017) in R (R Core Team, 2020) to train RF.

Because RF can tackle highly correlated variables, we used all variables (shown in Table S1) in the RF and evaluated the variable importance. Variable importance was measured by averaging a variable's total decrease in the remaining mean square errors (MSE) left after the variable was used as the node split. The variable importance was calculated using the *ranger* function (Wright and Ziegler, 2017) by setting the variable importance mode as 'impurity'.

2.4. Comparison with previous approach using back-extrapolation

We compared predictions from the newly developed annual LUR models with predictions from back-extrapolation methods used previously (de Hoogh et al., 2018a; Gulliver et al., 2013). We added this comparison to evaluate how large the reduction was in model performance of back-extrapolation compared to year-specific LUR. As previous studies have used the back-extrapolation methods, it is important for air pollution epidemiology to evaluate how different exposure predictions were from the new models and the previously applied back-extrapolation models. The back-extrapolation methods extrapolated predictions from the annual LUR model built for the year 2010 ($t_1 = 2010$) to another time (t_2) using the surfaces from the Danish Eulerian Hemispheric Model (DEHM) described in Table S1. The DEHM is a 3D long-range atmospheric chemical transport model (Brandt et al., 2012; Christensen, 1997). Following de Hoogh et al (2018) (de Hoogh et al., 2018a), we used two extrapolation methods: ratio (Eq (1)) and differencing (Eq (2)) (Gulliver et al., 2013):

$$y_{extrap}(t_2) = y_{LUR}(t_1) \times \frac{y_{DEHM}(t_2)}{y_{DEHM}(t_1)} \tag{1}$$

$$y_{extrap}(t_2) = y_{LUR}(t_1) + y_{DEHM}(t_2) - y_{DEHM}(t_1) \tag{2}$$

where y_{extrap} is the extrapolated value, y_{LUR} is the estimate from a LUR model, and y_{DEHM} is the DEHM estimate. The DEHM surfaces (50 km × 50 km) were resampled to the LUR surfaces (25 m × 25 m) using nearest neighbor resampling. The reference time was set to be 2010, the midpoint of our study period and the year in which our previous Europe-wide ELAPSE LUR model was built (de Hoogh et al., 2018a).

The comparison in model performance was done by using 5-fold CV and calculating correlations between extrapolated concentrations and the new annual LUR predictions. We calculated the correlations at random points. We generated 77,675 random points in populated areas (Fig. S3). The comparison was done for the full study domain and per NUTS1 area. The number of random points in each NUTS1 region was proportional to the population and each region had at least 300 points generated. We ensured these random points were randomly distributed in the populated areas defined by the impervious density data for the year 2018 and the population data for the year 2011 (see Table S1).

2.5. Model performance evaluation

For internal validation, we used 5-fold CV to evaluate the model performance in explaining the variability of observed concentrations for each year. For single-year modeling, the observation points were divided randomly into 5 folds stratified by the climate zone (see Fig. S4) and routine monitoring station type (i.e., background, industrial, traffic). For multi-year modeling, because some stations stopped or started monitoring during the period (2000–2019), we first obtained the number of annual averages available during the period for each station. Then, the observation points were divided randomly into 5 folds stratified by the number of annual averages, climate zone, and station type. For multi-year modeling, we ensured that the observations from the same station could only be in either the training fold or the validation fold across the years.

For external validation, we used ESCAPE measurements (Cyrys et al., 2012; Eeftens et al., 2012) to evaluate model performance in 2010 for NO_2, PM_{10}, and $PM_{2.5}$. The ESCAPE measurements consist of 1396 long-term NO_2 measurements and 415 PM_{10} and $PM_{2.5}$ measurements collected in specific clustered study areas across Europe in 2010 (see Fig. S5).

Two performance metrics were used: the coefficient of determination (R^2) and root mean square error (RMSE). We calculated the MSE-based R^2, affected by both systematic bias and random differences between model predictions and observations. The predictions from all five validation folds were combined to obtain one value of each performance metric described in Eq (3) and Eq (4):

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} \tag{3}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2} \tag{4}$$

where y_i is the observation of the annual average concentration at station i, $\widehat{y}_i$ is the estimate of the annual average concentration at station i, $\overline{y}$ is the average observations from all N stations.

3. Results and discussions

3.1. Model performance of algorithms with and without spatially-varying coefficients

Sections 3.1.1 to 3.1.3 document the performance of single and multi-year models. Section 3.1.4 illustrates the predictor variables in the different models. Section 3.1.5 compares our findings with previous studies and discusses the importance of including spatially-varying coefficients.

3.1.1. GWR: spatially-varying coefficients in single-year models

GWR modestly improved the explained variance R^2 values of single-year models compared to SLR and RF for all four pollutants (see columns in 'Difference in CV R^2 between single-year models' in Table 1). Table 1, Fig. 1 (for selected years), and Tables S3–S6 (for all years) document that GWR improved predictions compared to SLR and RF consistently across years. The largest improvement in R^2 was found for O_3 and PM_{10} (about 10% on average) and the smallest for $PM_{2.5}$ (about 2% on average) compared to SLR (see column 'GWR vs SLR' in Table 1). The difference between the performance for the two particle metrics may be because the ratios between $PM_{2.5}$ and PM_{10} concentrations are different across Europe (Eeftens et al., 2012). RF improved model performance compared to SLR modestly for NO_2, O_3, and PM_{10}, but not $PM_{2.5}$ for which it performed modestly worse (Table 1). The small improvement of model improvement by RF compared to SLR is consistent with previous work for European models of the year 2010 (Chen et al., 2019) and

Table 1
Descriptive statistics of 5-fold CV R^2 values and differences in 5-fold CV R^2 between different models from 2000 to 2019.

		Absolute CV R^2 values		Difference in CV R^2 between single-year models			Difference in CV R^2 between multi-year models			Difference in CV R2 between single-year and multi-year models		
		GWR	GTWR00-19	GWR vs SLR	GWR vs RF	RF vs SLR	GTWR00-19 vs SLR00-19	GTWR00-19 vs RF00-19	RF00-19 vs SLR00-19	SLR vs SLR00-19	GWR vs GTWR00-19	RF vs RF00-19
NO2	mean	0.67	0.66	0.05	0.02	0.03	0.07	0.03	0.04	0.02	0.00	0.01
	min	0.62	0.61	0.04	0.01	0.01	0.06	0.00	−0.07	0.01	−0.01	−0.02
	max	0.70	0.69	0.07	0.04	0.05	0.09	0.15	0.07	0.04	0.01	0.13
	sd	0.03	0.03	0.01	0.01	0.01	0.01	0.03	0.03	0.01	0.01	0.03
O3	mean	0.57	0.58	0.11	0.04	0.06	0.13	0.09	0.04	0.01	−0.01	0.04
	min	0.43	0.45	0.07	0.00	0.02	0.08	0.01	−0.04	−0.03	−0.04	−0.02
	max	0.66	0.67	0.15	0.10	0.13	0.19	0.19	0.14	0.05	0.03	0.12
	sd	0.05	0.05	0.03	0.03	0.04	0.03	0.05	0.05	0.02	0.02	0.03
PM10	mean	0.61	0.62	0.10	0.05	0.05	0.12	0.08	0.04	0.02	−0.01	0.02
	min	0.48	0.50	0.07	0.00	−0.01	0.08	0.04	0.00	0.00	−0.03	−0.02
	max	0.71	0.73	0.14	0.11	0.12	0.16	0.14	0.11	0.07	0.04	0.12
	sd	0.06	0.07	0.02	0.03	0.04	0.03	0.03	0.04	0.02	0.02	0.03
PM2.5	mean	0.77	0.77	0.02	0.09	−0.08	0.03	0.07	−0.04	0.00	−0.01	−0.03
	min	0.69	0.71	−0.01	0.07	−0.20	0.00	0.04	−0.13	−0.03	−0.06	−0.12
	max	0.82	0.82	0.06	0.19	−0.02	0.10	0.13	0.04	0.06	0.01	0.04
	sd	0.04	0.04	0.02	0.03	0.04	0.02	0.03	0.04	0.02	0.02	0.04

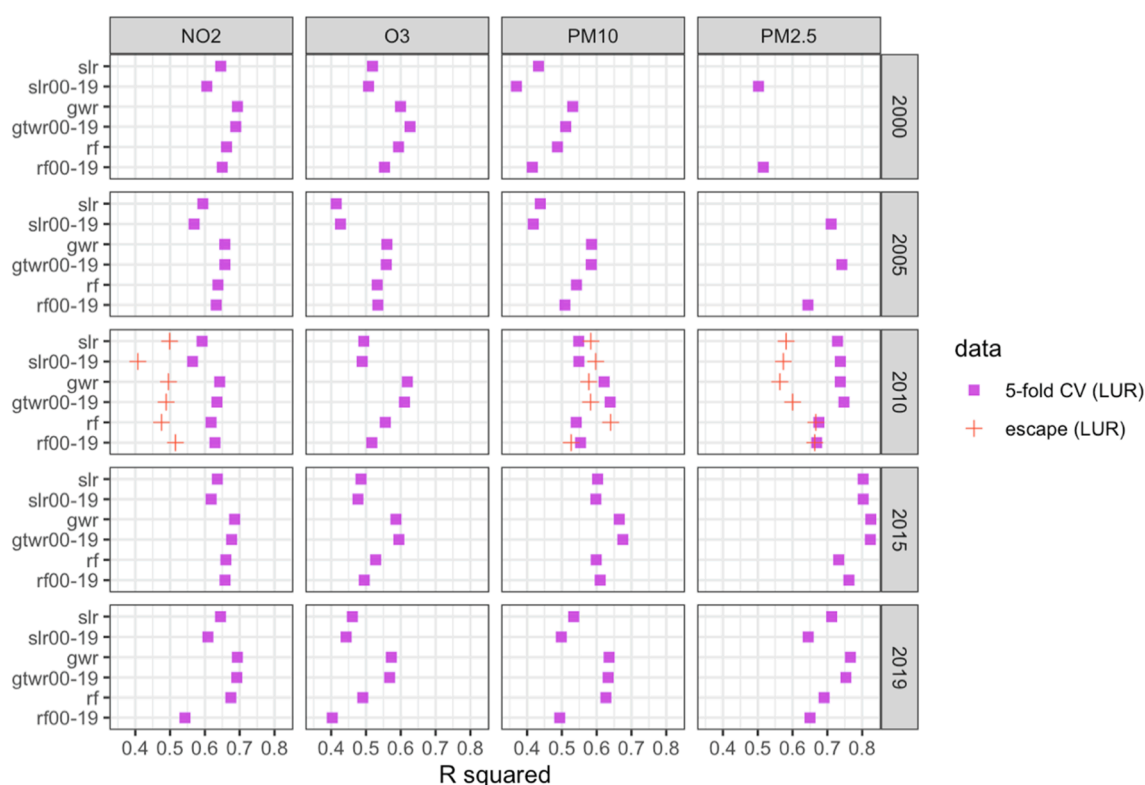mean: average; min: minimum; max: maximum; sd: standard deviation.

Fig. 1. R^2 values of 5-fold CV R^2 values for SLR, GWR, GTWR, and RF models for selected years. Internal validation using 5-fold CV was done for each year (shown in solid-square icons), whereas external ESCAPE validation was only available for 2010 (shown in cross icons). 'slr' and 'gwr' are the single-year SLR and GWR models built for each year. 'slr00-19' and 'gtwr00-19' are the multi-year SLR and GTWR models built for period from 2000 to 2019. For PM2.5, because the number of observations was too few to train GTWR (between 2000 and 2003) for some folds and GWR (between 2000 and 2005) for all folds, the values of GWR and GTWR were missing. Values for other years are shown in Fig. S6 and Tables S3–S6. RMSE values are shown in Fig. S7.

models based upon mobile monitoring data (Kerckhoffs et al., 2019).

The model performance of the annual GWR LUR models was quite stable over time for NO$_2$ (Table S3, Figs. S6 & S7). For O$_3$, the spatial variations explained by the GWR LUR models fluctuated between years (Table S4). For PM$_{10}$ (Table S5) and PM$_{2.5}$ (Table S6), model performance improved in recent years compared to earlier years, probably related to the significant increase in the number of sites and possibly harmonization of monitoring methods across Europe. The trends and

improvement in R^2 (Fig. S6) and RMSE values (Fig. S7) were similar.

3.1.2. GTWR: spatially- and temporally-varying coefficients in multi-year models

GTWR00-19 explained more variation in observations than the other multi-year models (SLR00-19, RF00-19) (Table 1, Fig. 1 for selected years, Tables S3–S6 for all years). GTWR00-19 improved the R^2 values compared to SLR00-19 by 7%, 13%, 12%, 3% on average and compared

to RF00-19 by 3%, 9%, 8%, 7% on average for NO_2, O_3, PM_{10}, and $PM_{2.5}$ respectively (see columns 'Difference in CV R^2 between multi-year models' in Table 1). Table 1 and Tables S3–S6 document that GTWR improved model performance for all years compared to other multi-year models. The multi-year RF model (RF00-19) overall slightly improves model performance compared to the multi-year SLR model (SLR00-19) for NO_2, O_3, and PM_{10} but not $PM_{2.5}$ for some years (Tables S3–S6). The scatterplots of GTWR predictions against ground-based monitoring observations are shown in Fig. S8.

We evaluated whether using observations from shorter multi-year periods would lead to different model performance and model predictions for each year. We used observations from shorter multi-year periods (ranging from 3 to 8 years) to train GTWR. These GTWR models with shorter multi-year periods gave similar 5-fold CV R^2 values (Table S7) in 2010, 2015, and 2019. The correlations were high between predictions from GTWR00-19 (our default GTWR model using observations from 20 years) and other GTWR models built by observations from multi-year periods shorter than the 20-year period (as shown in row 'cor' in Table S7).

3.1.3. Comparison between single-year and multi-year models

When we compared single-year models and multi-year models, we noted that GTWR00-19 estimated spatial variation in annual average observations as good as the annual GWR models (column 'GWR vs GTWR00-19' in Table 1, Fig. 1 for selected years, and Tables S3–S6 for all years). For $PM_{2.5}$, because the number of observations was too small to train GTWR (between 2000 and 2003) for some folds and GWR (between 2000 and 2005) for all folds, the values of 5-fold CV R^2 were unavailable. The performance of multi-year models with the SLR and RF algorithm was slightly less than the performance of the single-year models with these algorithms (column 'SLR vs SLR00-19' and 'RF vs RF00-19' in Table 1).

Using the ESCAPE data as external validation data available for the year 2010 for only NO_2, $PM_{2.5}$, and PM_{10}, we observed that GWR and GTWR00-19 performed similarly compared to other algorithms for NO_2. For $PM_{2.5}$ and PM_{10}, GWR and GTWR00-19 performed similarly compared to SLR and SLR00-19 but slightly worse than RF (indicated by the plus icon in Fig. 1). As the ESCAPE monitoring data is much more spatially clustered and covers fewer countries than the routine monitoring data (Fig. S5), we gave more importance to the 5-fold cross-validation when comparing modeling approaches.

GWR and GTWR both improved the performance of regression-based methods by incorporating spatial-varying coefficients in the LUR models. The GWR models used slightly different predictors with spatially-varying coefficient values each year, whereas the GTWR model had a fixed model structure in terms of predictors and buffer sizes but with spatially- and temporally-varying coefficients. Although the annual GWR models have different predictors across years, these mostly reflect the same determinant (e.g., nearby road traffic represented by different buffer sizes) (Fig. S9). Although GTWR and GWR showed similar 5-fold CV accuracies, GTWR had the extra advantage of estimating PM2.5 concentrations when the number of observations was too small to train an annual LUR model for years before 2006. GTWR includes observations in other years to inform the model for a specific year. Thus, we will use the GTWR models as our default model.

3.1.4. Model structure of single-year and multi-year models

Fig. S9 shows the variables selected by and used SLR and used in GWR in each year. The variables selected for NO_2 and PM_{10} were quite similar across years with some different variables selected in some years, whereas the variables for O_3 and $PM_{2.5}$ were quite different across years. Road variables were selected with slightly different road types and buffer sizes in all years for all pollutants. The chemical transport model estimates were also selected for all pollutants: estimates from the MACC-II ensemble model for NO_2 (no2_10MACC), estimates from the DEHM model for O_3 (O3_dehm), and estimates from the satellite retrievals

converted by the GEOS-Chem model for $PM_{2.5}$ and PM_{10} (gwr_sat). Impervious densities with different buffer sizes were selected in all years for NO_2 and O_3. Population and altitude were selected almost every year for O_3. Different meteorological variables were selected almost every year for all pollutants, which are wind speed for NO_2 and temperature for O_3, PM_{10}, and $PM_{2.5}$.

Fig. S11 shows the optimized GWR adaptive bandwidth (nngb) for each pollutant and each year. Figs. S14–S17 illustrate spatially-varying coefficients estimated by GWR for all pollutants for the year 2010, indicating noticeable spatial variability in the coefficient values across Europe. We do not have a clear interpretation of the spatial pattern of coefficients related to known sources or dispersion characteristics.

Fig. S10 gives the top-12 variables in reducing MSE the most for each RF model. Most variables selected in SLR were also the top-12 variables in RF. The optimized RF hyperparameters were similar in some RF models (Fig. S12, S13).

Fig. S9 (in the columns of '00-19' for four pollutants) shows the variables selected by and used in the multi-year SLR model (SLR00-19) and used in the multi-year GTWR model (GTWR00-19). Figs. S18–S25 illustrates the spatially- and temporally-varying coefficient values of the GTWR00-19 multi-year model in the year 2010 and 2015. In Figs. S18–S25, the legend scales remain fixed for the same predictors and pollutants in the two years to observe changes in coefficient values over time. We observed fair consistent spatial pattern in the coefficient values over time. In other words, as observed from the spatially- and temporally-varying coefficients in the GTWR00-19 (Figs. S18–S25), the spatial variability in the coefficient values was larger than the temporal variability. Table S12 shows the optimized GTWR parameters for all pollutants.

3.1.5. Discussion of the importance of allowing spatially-varying coefficients

Overall, the modest improvement of model performance by allowing spatially-varying coefficients over a large spatial area, such as Europe, is consistent with the previous modeling at the North-American and global scale (Hammer et al., 2020; Lu et al., 2020; Van Donkelaar et al., 2015). Incorporating spatial variations in the relationships between predictors and air pollution is important for estimating air pollution over a large spatial extent. The GTWR00-19 models were used to estimate Europe-wide annual average concentrations for NO_2, O_3, PM_{10}, and $PM_{2.5}$ at a 25 m × 25 m spatial resolution, a considerable improvement compared to our earlier work on a 100 m × 100 m spatial resolution (de Hoogh et al., 2018a). The high spatial resolution (25 m) provided by the road predictors could improve capturing traffic-related emission sources, especially for NO_2.

Our previous Europe-wide model (de Hoogh et al., 2018a) resulted in 5-fold CV R^2 values of 0.66 and 0.64 in 2010 and 2013 using SLR followed by kriging compared to our GTWR R^2 of 0.74 and 0.81 for $PM_{2.5}$. For NO_2, our SLR (0.65 in 2000, 0.59 in 2005, 0.59 in 2010) had slightly higher 5-fold CV R^2 values than the previous SLR model (0.54 in 2000, 0.5 in 2005, 0.58 in 2010). For O_3, our SLR model (0.52 in 2000, 0.47 in 2005, 0.63 in 2010) had slightly lower R^2 values than the previous SLR model (0.58 in 2000, 0.47 in 2005 and 0.63 in 2010). But these lower R^2 values for O_3 were because we did not include the X and Y coordinates as predictors in our SLR model, whereas the previous SLR included the X and Y coordinates. Our GWR model, however, had higher R^2 values (0.6 in 2000, 0.56 in 2005, 0.6 in 2010) than the previous SLR model. The improvement compared to the previous Europe-wide model could be because in this study we included more variables (e.g., meteorological variables) and variables with improved spatial resolution (e.g., road variables improved from 100-m to 25-m).

Our GTWR model also outperformed another European model created in a global study in 2011 (Larkin et al., 2017) for NO_2 (0.64 vs 0.57). Although in these previous studies the study areas and validation methods were slightly different from ours, our higher R^2 values could be because of two factors. Firstly, we used spatially-varying regression

coefficients instead of spatially-fixed linear regression. Secondly, the predictors we used here were more spatially resolved and we also included some additional CTM, SAT, and meteorological data as predictors to capture spatial variations in air pollution (Fig. S9). For $PM_{2.5}$, the second factor was especially important; for O_3 and NO_2 both factors were important.

The improvement in model performance by including spatial variations in linear regression was larger than by including spatial variations in non-linear models as achieved by using climate zone in RF to capture the spatial variations in air pollution (shown in columns 'GWR vs SLR' and 'GTWR00-19 vs RF00-19′ in Table 1). This may be specific to modeling at the annual time scale, where the relationships between predictors and concentrations do not materially deviate from linear

(Chen et al., 2019). Moreover, the similar model performance between GWR and GTWR and the relatively stable temporal variability in GTWR coefficient values indicate that capturing spatial variability in linear regression is more important than capturing temporal variability at an annual scale in Europe.

We note that the improvement of model performance by GTWR and GWR compared to SLR and RF was generally fairly modest, suggesting that models derived with these methods provide reasonable model predictions as well.

3.2. Modeled spatial patterns over 20 years

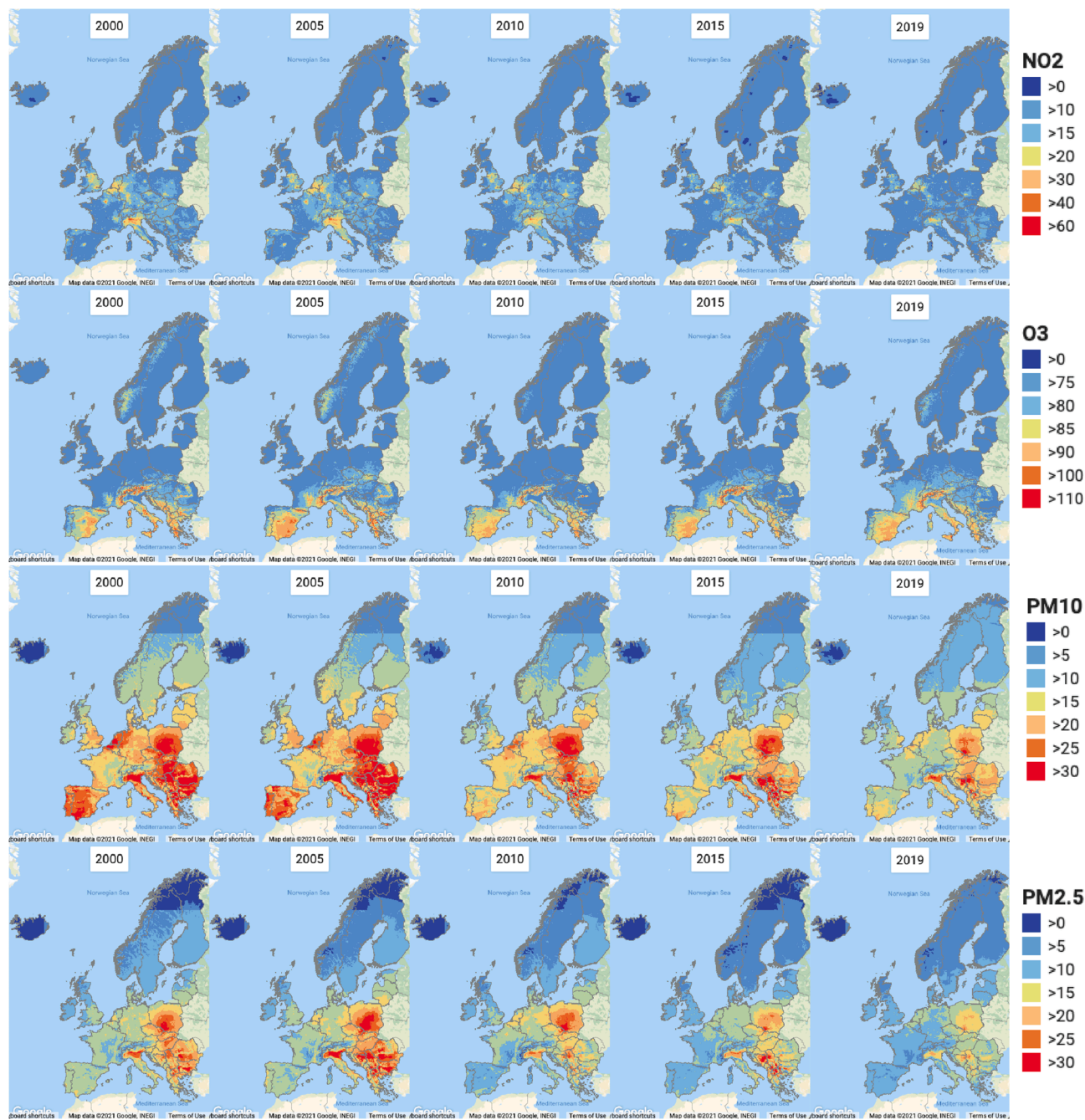Fig. 2 shows Europe-wide annual average concentration maps of

Fig. 2. Europe-wide annual average ground-level NO_2, O_3, PM_{10}, and $PM_{2.5}$ concentrations (μg/m³) estimated by GTWR00-19 in 2000, 2005, 2010,2015, and 2019 (Base map source: Google Maps). For maps for all years from 2000 to 2019 for all pollutants, readers are referred to https://youchenshenuu.users.earthengine.app /view/expanse-air-pollution-20-yr-maps.

NO_2, O_3, PM_{10}, and $PM_{2.5}$ estimated by GTWR00-19 and Fig. 3 shows the zoom-in maps in Paris. All annual maps from 2000 to 2019 are available on the website (https://youchenshenuu.users.earthengine. app/view/expanse-air-pollution-20-yr-maps).

For NO_2, the large-scale spatial patterns were similar over time with a decreasing trend. High concentrations were in cities and in specific areas such as the Italian Po-Valley, the surroundings of the Netherlands, and eastern Europe, related to population density (de Hoogh et al., 2018a). For O_3, the large-scale spatial patterns were similar over time with a small increasing trend. Overall, the Alps, southern Europe, and the Balkans had higher annual average O_3 concentrations than the rest of the study area.

For $PM_{2.5}$, the spatial patterns were also similar over time with a decreasing trend, but the patterns varied more in time compared to NO_2 and O_3. Overall, the northern part of Italy, eastern Europe and the Balkans had higher annual average $PM_{2.5}$ concentrations than the rest of Europe. For PM_{10}, the concentrations were decreasing over time, but the spatial patterns were more dynamic compared to NO_2 and O_3. In the early 2000 s, the PM_{10} concentrations were above 20 μg/m^3 in all regions except in the Alps, Ireland, and northern Europe. In the recent 5 years, only the northern part of Italy, eastern Europe, and the Balkans had higher PM_{10} concentrations (>25 μg/m^3) than the rest of Europe. For both PM_{10} and $PM_{2.5}$, some distinct dividing lines in northern Europe were because areas above these lines had no values in the

satellite-derived product (gwr_sat in Table S1) (Hammer et al., 2020; Van Donkelaar et al., 2019), and we replaced these missing values with zero values for this product.

The overall spatial patterns for NO_2 remained quite stable from year to year with a decreasing trend, but for PM_{10}, $PM_{2.5}$, and O_3 some regional changes over time were visible. The ambient air pollution concentrations are mostly driven by anthropogenic emission sources and are also influenced by meteorological factors and long-range transport of precursor gases for some pollutants. For NO_2, the ambient concentrations are mainly driven by the anthropogenic emission sources (i.e., mostly from road transport and energy combustion) reduced by policy regulation and improved efficiencies in energy combustion (Colette and Rouïl, 2020; EEA, 2021a, 2021b, 2021c). Thus, the declining estimated NO_2 concentrations were spatially stable. For O_3, PM_{10}, and $PM_{2.5}$, the ambient concentrations are influenced by not only emission sources but also meteorological factors and long-range transport of precursor gases (such as nitrogen oxides) (Colette and Rouïl, 2020; EEA, 2021b). The annual average of the daily maximum 8-hour mean for O_3 was the only pollutant with an increasing trend over time, as found in both our study and ground-based observations (Colette and Rouïl, 2020).

3.3. Back-extrapolation

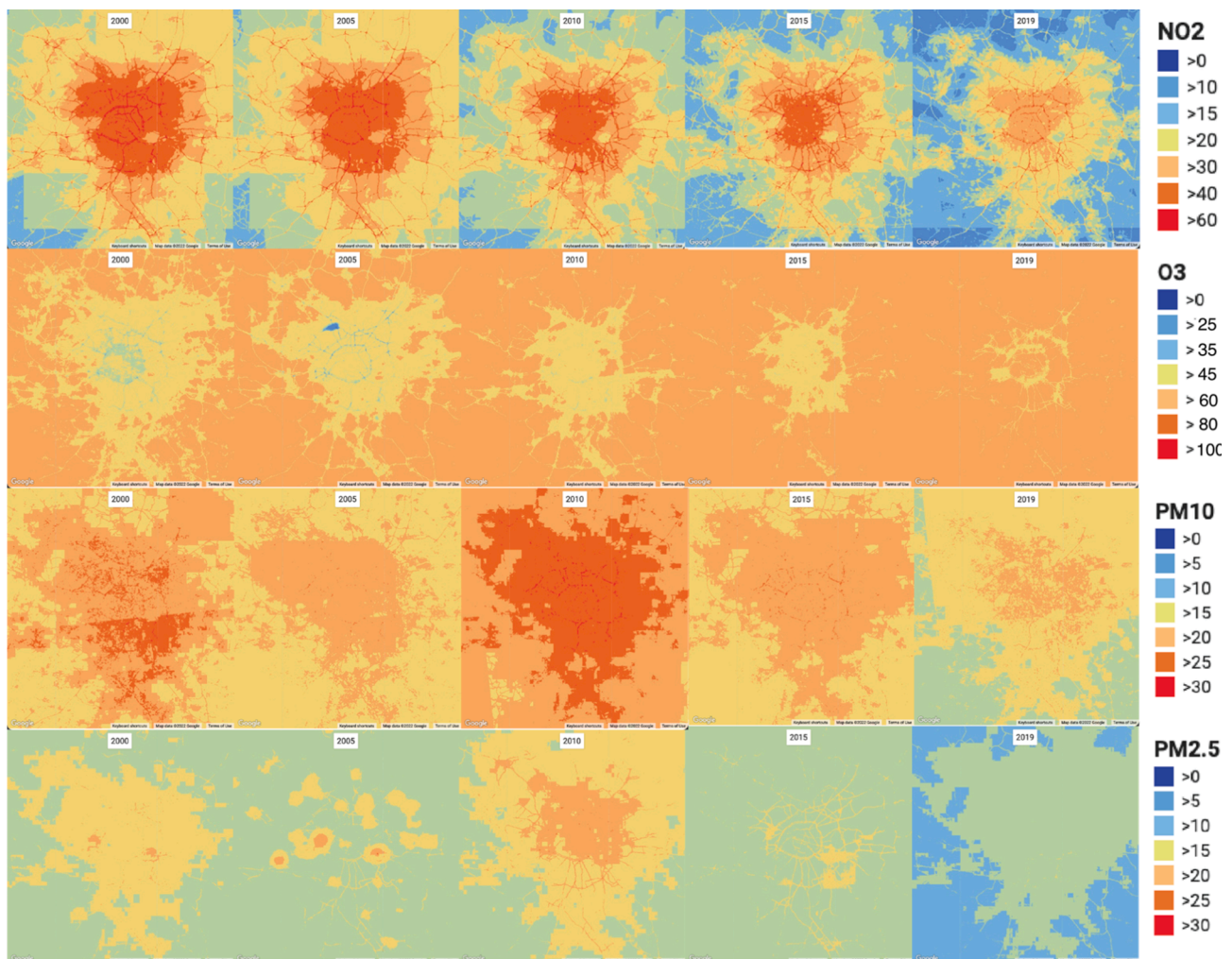In section 3.3.1 we first discuss the performance of back-

Fig. 3. Annual average ground-level NO_2, O_3, PM_{10}, and $PM_{2.5}$ concentrations (μg/m^3) estimated by GTWR00-19 in 2000, 2005, 2010, 2015, and 2019 in Paris. For maps for all years from 2000 to 2019 for all pollutants, readers are referred to https://youchenshenuu.users.earthengine.app/view/expanse-air-pollution-2 0-yr-maps.

extrapolation based on 5-fold CV. In section 3.3.2 we compare the performance of back-extrapolation with the new GTWR model described in section 3.1.

3.3.1. Performance of back-extrapolated models

The 5-fold CV performance of the back-extrapolation methods generally decreased when the extrapolation year (t_2 in Eq (1) and (2)) became more distant from the reference year ($t_1 = 2010$ in Eq (1) and (2)), as shown in Fig. S6 and Tables S8–S11 (in the 2nd-4th columns). For NO_2 (Table S8), the differencing method had higher CV R^2 values than the ratio method, with only a small decreasing trend for backward extrapolation (before 2010) and with a large decreasing trend for forward extrapolation (after 2010). For O_3 (Table S9), both back-extrapolation methods had similar performance that decreased with time especially for backward extrapolation and that decreased slightly for forward extrapolation (except 2018 & 2019). For PM_{10} (Table S10) and $PM_{2.5}$ (Table S11), both back-extrapolation methods had similar performance that decreased with time especially for backward extrapolation and that decreased slightly for forward extrapolation (except 2019). The large decreasing in CV R^2 values for backward extrapolation

for PM_{10} and $PM_{2.5}$ could be because of the significant decrease in the available observations and the quality of the observations in the early 2000 s. The number of observations for the earliest was small, especially for $PM_{2.5}$ (with less than 200 observations available before 2006).

Overall, the average annual predictions from the back-extrapolation were less similar to the annual average observations in years more distant to the year 2010 (for years both before and after 2010) for all pollutants.

3.3.2. Comparison between GTWR LUR model and back-extrapolation

The 5-fold CV performance of the back-extrapolation methods was lower compared to the GTWR00-19 models for all pollutants when going further away in time from 2010. For NO_2 (Table S8), the CV R^2 was 8% lower for back-extrapolated values in 2000 using the differencing method compared to GTWR00-19. For O_3 and especially PM_{10} and $PM_{2.5}$, the difference in CV R^2 values of GTWR00-19 and back-extrapolation increased away from the year 2010 for both differencing and ratio methods in this study.

The overall correlations were above 0.86 between the GTWR00-19 annual average predictions and the differencing back-extrapolated
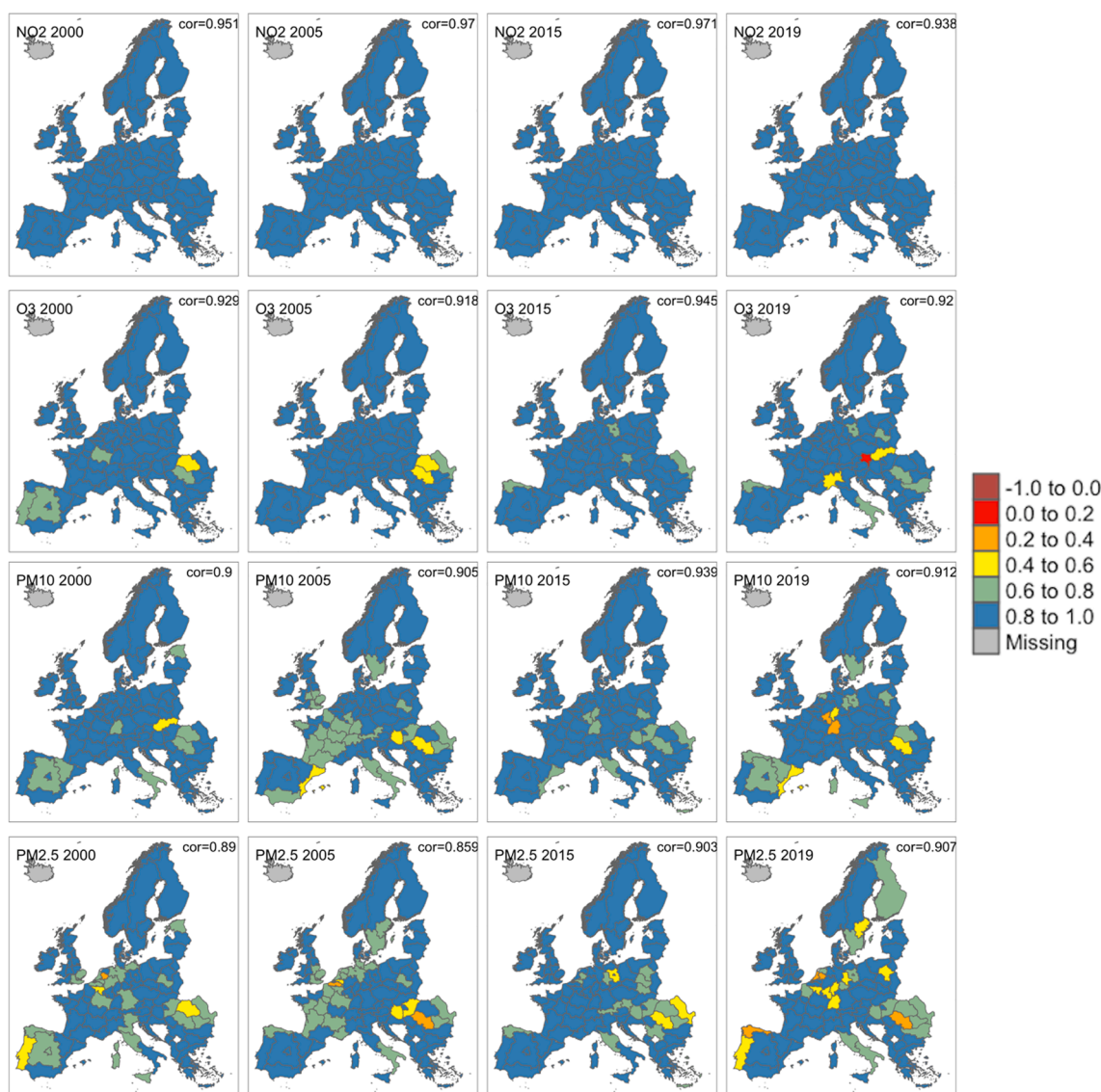

Fig. 4. Correlation between predictions from GTWR00-19 (GTWR built for period from 2000 to 2019) and predictions extrapolated from the 2010 GTWR00-19 predictions to other years by the differencing method using DEHM data described in Table S1 ($t_1 = 2010$ in Eq (2)) for NO_2, O_3, PM_{10}, and $PM_{2.5}$. Correlation values shown in the upper corner of each graph are overall correlations.

predictions at random points for all years (Fig. 4). For NO_2, the correlations were high also within all NUTS1-regions (>0.8). For O_3, PM_{10}, and $PM_{2.5}$, the correlations were above 0.8 in most regions but were below 0.6 in some regions and years.

Between the GTWR00-19 LUR models and back-extrapolated models for all years, the decreasing difference in CV R^2 (calculated as MSE-based R^2) and the high overall correlation could be explained by the slightly increasing differences between absolute average levels estimated by the large-scale DEHM model and the average levels estimated by the GTWR00-19 models (as shown in the 5th-8th columns of Tables S8–S11). Boxplots of the predicted values showed that the outliers and the average of the back-extrapolated values were slightly higher than the outliers and the average of the GTWR00-19 predictions (Fig. 5).

Overall, our result is similar to previous studies for NO_2. In Great Britain (Gulliver et al., 2016), a decrease of 8% in R^2 was found between a 1991 LUR model and a back-extrapolated LUR (extrapolated from 2009 LUR). In Vancouver in Canada, a decrease of 3% in the R^2 was found between a 2003 LUR model and a 2010 LUR model recalibrated to 2003 (Wang et al., 2013). The moderate to high correlations indicate that the GTWR predictions and the back-extrapolated predictions may give similar relative exposure ranking/classification across the population in Europe-wide cohort studies, but the difference in absolute values, as indicated by the low MSE-based R^2 values, could be high. Thus, GTWR would be preferred especially to study the shape of the exposure–response relationships between health effects and air pollution (i. e., at what level of air pollution health effects occur) because accurate absolute values are required. The back-extrapolated values with systematic bias would be a reasonable approximation in health studies in which the occurrence of specific health outcomes is compared in linear models or compared between low and high exposed subjects.

3.4. Limitations and strengths

An important limitation of this study is the lack of ground-level observations for $PM_{2.5}$ before 2006. This limitation was mitigated by the multi-year modeling method, but we were unable to perform 5-fold CV

for GTWR00-19 before 2004 because of the limited observations, with the number of observations in each of those years less than 100. This limited number of ground-level observations however could make our cross-validation less reliable in the early 2000s than in the later years for $PM_{2.5}$. With GTWR00-19 we were able to estimate $PM_{2.5}$ concentrations with limited observations and to increase 5-fold CV R^2 values by 7% in NO_2, 13% in O_3, 12% in PM_{10}, and 3% in $PM_{2.5}$ compared to SLR.

We did not have traffic intensity data available across Europe and instead used buffered road length of different road types from OpenStreetMap (OSM). We used the road type data from the year 2020 only because we judged that using historical data was associated with too large methodological issues (improved quality of OSM over time).

Despite these limitations, our GTWR00-19 and GWR have captured the overall spatial variations in air pollution every year with satisfactory 5-fold CV R^2 values against the ground-level observations. Our output gives harmonized exposure estimates in European cohort studies at a high spatial resolution (25 m × 25 m).

Our annual average air pollution exposure maps will be applied in health effect studies of long-term exposure to air pollution. For studies of short-term exposure, a more refined temporal resolution is needed, following the methodology of some recent studies that have developed daily pollution maps for multiple years at a resolution of 1 km or coarser with good performance (de Hoogh et al., 2019, 2018b; Di et al., 2019; Liu et al., 2020; Requia et al., 2020; Shtein et al., 2018). For our current objective, the spatial resolution of 1 km × 1 km is not sufficient.

All code to build LUR models in R programming language is available on Github for single-year (Shen et al., 2021a) and multi-year modeling (Shen et al., 2021b).

4. Conclusions

We showed the importance of including spatially-varying relationships in LUR models to improve long-term air pollution estimates in Europe. The spatially-varying linear regression models (GWR, GTWR) explained a modestly larger amount of spatial variation in air pollution concentrations across Europe than the spatially-fixed linear regression (SLR) and the machine learning method (RF). Our harmonized annual
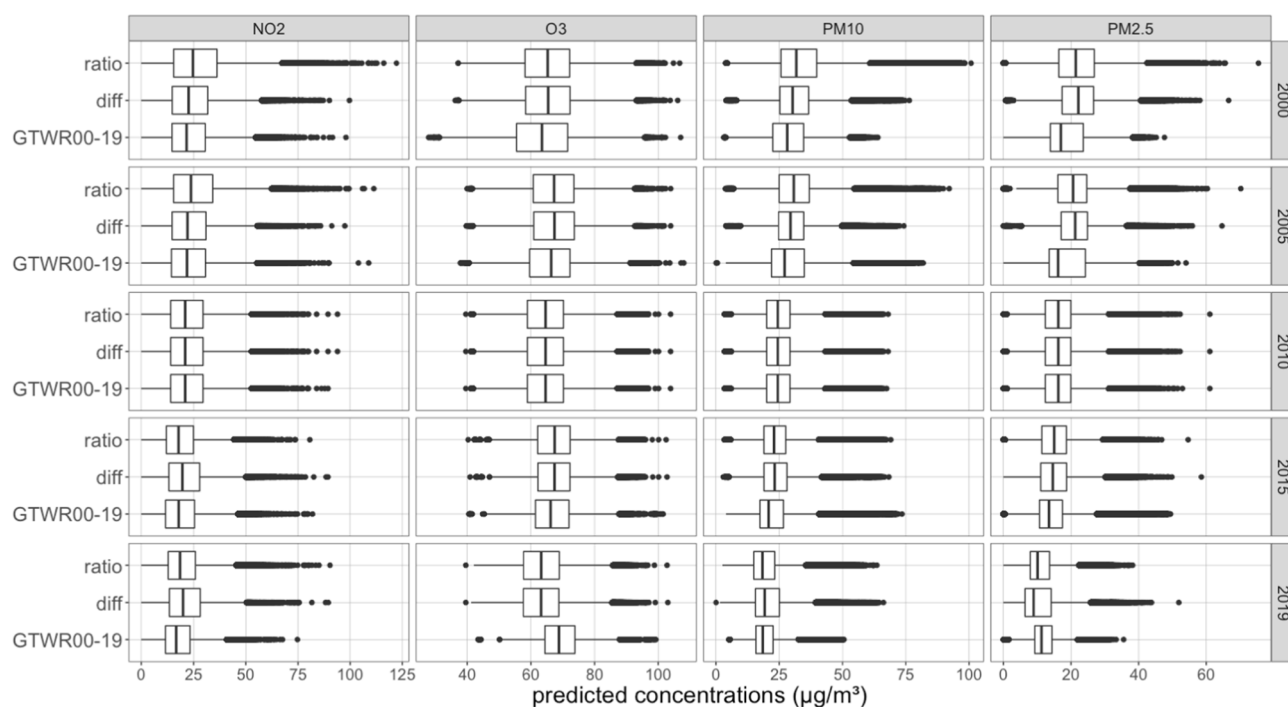

Fig. 5. Boxplots of predictions at random points from GTWR00-19, differencing back-extrapolation (diff) and ratio back-extrapolation (ratio) using GTWR00-19 for year 2010 extrapolated to other years using DEHM data described in Table S1 ($t_1 = 2010$ in Eq (1) and (2)).

estimates available for NO2, O3, PM10, and PM2.5 from 2000 to 2019 will allow time-varying exposure-health risk models for Europe-wide health analysis studies.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2022.107485.

References

Bai, Y., Wu, L., Qin, K., Zhang, Y., Shen, Y., Zhou, Y., 2016. A geographically and temporally weighted regression model for ground-level PM2.5 estimation from satellite-derived 500 m resolution AOD. Remote Sens. 8, 262. https://doi.org/10.3390/rs8030262.

Bechle, M.J., Millet, D.B., Marshall, J.D., 2015. National spatiotemporal exposure surface for NO2: Monthly scaling of a satellite-derived land-use regression, 2000–2010. Environ. Sci. Technol. 49, 12297–12305. https://doi.org/10.1021/acs.est.5b02882.

Brandt, J., Silver, J.D., Frohn, L.M., Geels, C., Gross, A., Hansen, A.B., Hansen, K.M., Hedegaard, G.B., Skjøth, C.A., Villadsen, H., Zare, A., Christensen, J.H., 2012. An integrated model study for Europe and North America using the Danish Eulerian Hemispheric Model with focus on intercontinental transport of air pollution. Atmos. Environ. 53, 156–176. https://doi.org/10.1016/j.atmosenv.2012.01.011.

Brauer, M., Brook, J.R., Christidis, T., Chu, Y., Crouse, D.L., Erickson, A., Hystad, P., Li, C., Martin, R.V., Meng, J., Pappin, A.J., Pinault, L.L., Tjepkema, M., van Donkelaar, A., Weichenthal, S., Burnett, R.T., 2019. Mortality-air pollution associations in low-exposure environments (MAPLE): Phase 1. Res. Rep. Health. Eff. Inst. 1–87.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Brunsdon, C., Fotheringham, A.S., Charlton, M.E., 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. Geogr. Anal. 28, 281–298.

Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., van Donkelaar, A., Hvidtfeldt, U.A., Katsouyanni, K., Janssen, N.A.H.H., Martin, R. V., Samoli, E., Schwartz, P.E., Stafoggia, M., Bellander, T., Strak, M., Wolf, K., Vienneau, D., Vermeulen, R., Brunekreef, B., Hoek, G., 2019. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. Environ. Int. 130, 104934 https://doi.org/10.1016/j.envint.2019.104934.

Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Weinmayr, G., Bauwelinck, M., van Donkelaar, A., Hvidtfeldt, U.A., Atkinson, R., Janssen, N.A.H., Martin, R.V., Samoli, E., Andersen, Z.J., Oftedal, B.M., Stafoggia, M., Bellander, T., Strak, M., Wolf, K., Vienneau, D., Brunekreef, B., Hoek, G., 2020. Development of Europe-wide models for particle elemental composition using supervised linear regression and random forest. Environ. Sci. Technol. 54 (24), 15698–15709.

Chen, J., Rodopoulou, S., de Hoogh, K., Strak, M., Andersen, Z.J., Atkinson, R., Bauwelinck, M., Bellander, T., Brandt, J., Cesaroni, G., Concin, H., Fecht, D., Forastiere, F., Gulliver, J., Hertel, O., Hoffmann, B., Hvidtfeldt, U.A., Janssen, N.A. H., Jöckel, K.H., Jørgensen, J., Katsouyanni, K., Ketzel, M., Klompmaker, J.O., Lager, A., Leander, K., Liu, S., Ljungman, P., Macdonald, C.J., Magnusson, P.K.E., Mehta, A., Nagel, G., Oftedal, B., Pershagen, G., Peters, A., Raaschou-Nielsen, O., Renzi, M., Rizzuto, D., Samoli, E., van der Schouw, Y.T., Schramm, S., Schwarze, P., Sigsgaard, T., Sørensen, M., Stafoggia, M., Tjønneland, A., Vienneau, D., Weinmayr, G., Wolf, K., Brunekreef, B., Hoek, G., 2021. Long-term exposure to fine

particle elemental components and natural and cause-specific mortality—a pooled analysis of eight european cohorts within the ELAPSE project. Environ. Health Perspect. 129 https://doi.org/10.1289/EHP8368.

Christensen, J.H., 1997. The Danish eulerian hemispheric model - a three-dimensional air pollution model used for the arctic. Atmos. Environ. 31, 4169–4191. https://doi.org/10.1016/S1352-2310(97)00264-1.

Colette, A., Rouïl, L., 2020. Air Quality Trends in Europe: 2000-2017. Assessment for surface SO2, NO2, Ozone, PM10 and PM2.5. Eionet Report - ETC/ATNI 2019/16.

Cyrys, J., Eeftens, M., Heinrich, J., Ampe, C., Armengaud, A., Beelen, R., Bellander, T., Beregszaszi, T., Birk, M., Cesaroni, G., Cirach, M., de Hoogh, K., De Nazelle, A., de Vocht, F., Declercq, C., Dedele, A., Dimakopoulou, K., Eriksen, K., Galassi, C., Graulevičiene, R., Grivas, G., Gruzieva, O., Gustafsson, A.H., Hoffmann, B., Iakovides, M., Ineichen, A., Krämer, U., Lanki, T., Lozano, P., Madsen, C., Meliefste, K., Modig, L., Mölter, A., Mosler, G., Nieuwenhuijsen, M., Nonnemacher, M., Oldenwening, M., Peters, A., Pontet, S., Probst-Hensch, N., Quass, U., Raaschou-Nielsen, O., Ranzi, A., Sugiri, D., Stephanou, E.G., Taimisto, P., Tsai, M.Y., Vaskövi, É., Villani, S., Wang, M., Brunekreef, B., Hoek, G., 2012. Variation of NO2 and NOx concentrations between and within 36 European study areas: Results from the ESCAPE study. Atmos. Environ. 62, 374–390. https://doi.org/10.1016/j.atmosenv.2012.07.080.

de Hoogh, K., Chen, J., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., van Donkelaar, A., Hvidtfeldt, U.A., Katsouyanni, K., Klompmaker, J., Martin, R.V., Samoli, E., Schwartz, P.E., Stafoggia, M., Bellander, T., Strak, M., Wolf, K., Vienneau, D., Brunekreef, B., Hoek, G., 2018a. Spatial PM 2.5, NO 2, O 3 and BC models for Western Europe – Evaluation of spatiotemporal stability. Environ. Int. 120, 81–92. https://doi.org/10.1016/j.envint.2018.07.036.

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

de Hoogh, K., Saucy, A., Shtein, A., Schwartz, J., West, E.A., Strassmann, A., Puhan, M., Röösli, M., Stafoggia, M., Kloog, I., 2019. Predicting fine-scale daily NO 2 for 2005–2016 incorporating OMI satellite data across Switzerland. Environ. Sci. Technol. 53, 10279–10287. https://doi.org/10.1021/acs.est.9b03107.

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. Environ. Int. 130, 104909 https://doi.org/10.1016/j.envint.2019.104909.

Di, Q., Wang, Y., Zanobetti, A., Wang, Y., Koutrakis, P., Choirat, C., Dominici, F., Schwartz, J.D., 2017. Air Pollution and Mortality in the Medicare Population. N. Engl. J. Med. 376, 2513–2522. https://doi.org/10.1056/nejmoa1702747.

EEA, 2021a. Europe's air quality status 2021- update — European Environment Agency [WWW Document]. URL https://www.eea.europa.eu/publications/air-quality-in-europe-2021/air-quality-status-briefing-2021 (accessed 7.13.22).

EEA, 2021b. Sources and emissions of air pollutants in Europe [WWW Document]. Eur. Environ. agency. URL https://www.eea.europa.eu/publications/air-quality-in-europe-2021/sources-and-emissions-of-air (accessed 7.13.22).

EEA, 2021c. Air quality in Europe 2021 Key messages [WWW Document]. Air Qual. Eur. 2021. URL https://www.eea.europa.eu/publications/air-quality-in-europe-2021/air-quality-in-europe-2021 (accessed 7.13.22).

Eeftens, M., Beelen, R., De Hoogh, K., Bellander, T., Cesaroni, G., Cirach, M., Declercq, C., Dedele, A., Dons, E., De Nazelle, A., Dimakopoulou, K., Eriksen, K., Falq, G., Fischer, P., Galassi, C., Gražulevičiene, R., Heinrich, J., Hoffmann, B., Jerrett, M., Keidel, D., Korek, M., Lanki, T., Lindley, S., Madsen, C., Mölter, A., Nádor, G., Nieuwenhuijsen, M., Nonnemacher, M., Pedeli, X., Raaschou-Nielsen, O., Patelarou, E., Quass, U., Ranzi, A., Schindler, C., Stempfelet, M., Stephanou, E., Sugiri, D., Tsai, M.Y., Yli-Tuomi, T., Varró, M.J., Vienneau, D., Klot, S.V., Wolf, K., Brunekreef, B., Hoek, G., 2012a. Development of land use regression models for PM2.5, PM 2.5 absorbance, PM10 and PMcoarse in 20 European study areas; Results of the ESCAPE project. Environ. Sci. Technol. 46, 11195–11205. https://doi.org/10.1021/es301948k.

Eeftens, M., Tsai, M.Y., Ampe, C., Anwander, B., Beelen, R., Bellander, T., Cesaroni, G., Cirach, M., Cyrys, J., de Hoogh, K., De Nazelle, A., de Vocht, F., Declercq, C., Dedele, A., Eriksen, K., Galassi, C., Gražulevičiene, R., Grivas, G., Heinrich, J., Hoffmann, B., Iakovides, M., Ineichen, A., Katsouyanni, K., Korek, M., Krämer, U., Kuhlbusch, T., Lanki, T., Madsen, C., Meliefste, K., Mölter, A., Mosler, G., Nieuwenhuijsen, M., Oldenwening, M., Pennanen, A., Probst-Hensch, N., Quass, U., Raaschou-Nielsen, O., Ranzi, A., Stephanou, E., Sugiri, D., Udvardy, O., Vaskövi, É., Weinmayr, G., Brunekreef, B., Hoek, G., 2012. Spatial variation of PM2.5, PM10, PM2.5 absorbance and PMcoarse concentrations between and within 20 European study areas and the relationship with NO2 - Results of the ESCAPE project. Atmos. Environ. 62, 303–317. https://doi.org/10.1016/j.atmosenv.2012.08.038.

European Environmental Agency, 2020a. AirBase - The European air quality database — European Environment Agency (EEA) [WWW Document]. URL https://www.eea.europa.eu/data-and-maps/data/airbase-the-european-air-quality-database-8 (accessed 11.18.20).

European Environmental Agency, 2020b. Air Quality e-Reporting (AQ e-Reporting) — European Environment Agency [WWW Document]. URL https://www.eea.europa.eu/data-and-maps/data/aqereporting-8 (accessed 11.18.20).

Fotheringham, A.S., Crespo, R., Yao, J., 2015. Geographical and Temporal Weighted Regression (GTWR). Geogr. Anal. 47, 431–452. https://doi.org/10.1111/gean.12071.

Girres, J.F., Touya, G., 2010. Quality Assessment of the French OpenStreetMap Dataset. Trans. GIS 14, 435–459. https://doi.org/10.1111/J.1467-9671.2010.01203.X.

Gollini, I., Lu, B., Charlton, M., Brunsdon, C., Harris, P., 2015. GWmodel}: An {R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models. J. Stat. Softw. 63, 1–50.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031.

Gulliver, J., De Hoogh, K., Hansell, A., Vienneau, D., 2013. Development and back-extrapolation of NO2 land use regression models for historic exposure assessment in Great Britain. Environ. Sci. Technol. 47, 7804–7811. https://doi.org/10.1021/es4008849.

Gulliver, J., de Hoogh, K., Hoek, G., Vienneau, D., Fecht, D., Hansell, A., 2016. Back-extrapolated and year-specific NO2 land use regression models for Great Britain - Do they yield different exposure assessment? Environ. Int. 92–93, 202–209. https://doi.org/10.1016/j.envint.2016.03.037.

Hammer, M.S., Van Donkelaar, A., Li, C., Lyapustin, A., Sayer, A.M., Hsu, N.C., Levy, R. C., Garay, M.J., Kalashnikova, O.V., Kahn, R.A., Brauer, M., Apte, J.S., Henze, D.K., Zhang, L., Zhang, Q., Ford, B., Pierce, J.R., Martin, R.V., 2020. Global Estimates and Long-Term Trends of Fine Particulate Matter Concentrations (1998–2018). Environ. Sci. Technol. 54, 7879–7890. https://doi.org/10.1021/acs.est.0c01764.

He, Q., Huang, B., 2018. Satellite-based high-resolution PM2.5 estimation over the Beijing-Tianjin-Hebei region of China using an improved geographically and temporally weighted regression model. Environ. Pollut. 236, 1027–1037. https://doi.org/10.1016/j.envpol.2018.01.053.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 2018, e5518.

Huang, B., Wu, B., Barry, M., 2010. Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. Int. J. Geogr. Inf. Sci. 24, 383–401. https://doi.org/10.1080/13658810802672469.

Kerckhoffs, J., Hoek, G., Gehring, U., Vermeulen, R., 2021. Modelling nationwide spatial variation of ultrafine particles based on mobile monitoring. Environ. Int. 154, 106569 https://doi.org/10.1016/j.envint.2021.106569.

Kerckhoffs, J., Hoek, G., Portengen, L., Brunekreef, B., Vermeulen, R.C.H., 2019. Performance of prediction algorithms for modeling outdoor air pollution spatial surfaces. Environ. Sci. Technol. 53, 1413–1421. https://doi.org/10.1021/acs.est.8b06038.

Knibbs, L.D., Hewson, M.G., Bechle, M.J., Marshall, J.D., Barnett, A.G., 2014. A national satellite-based land-use regression model for air pollution exposure assessment in Australia. Environ. Res. 135, 204–211. https://doi.org/10.1016/j.envres.2014.09.011.

Larkin, A., Geddes, J.A., Martin, R.V., Xiao, Q.Y., Liu, Y., Marshall, J.D., Brauer, M., Hystad, P., 2017. Global land use regression model for nitrogen dioxide air pollution. Environ. Sci. Technol. 51, 6957–6964. https://doi.org/10.1021/acs.est.7b01148.

Li, J., Heap, A.D., Potter, A., Daniell, J.J., 2011. Application of machine learning methods to spatial interpolation of environmental variables. Environ. Model. Softw. 26, 1647–1659. https://doi.org/10.1016/j.envsoft.2011.07.004.

Liu, R., Ma, Z., Liu, Y., Shao, Y., Zhao, W., Bi, J., 2020. Spatiotemporal distributions of surface ozone levels in China from 2005 to 2017: A machine learning approach. Environ. Int. 142, 105823 https://doi.org/10.1016/j.envint.2020.105823.

Lu, B., Harris, P., Charlton, M., Brunsdon, C., 2014. The GWmodel {R} package: further topics for exploring spatial heterogeneity using geographically weighted models. Geo-spatial Inf. Sci. 17 (2), 85–101.

Lu, M., Schmitz, O., de Hoogh, K., Kai, Q., Karssenberg, D., 2020. Evaluation of different methods and data sources to optimise modelling of NO2 at a global scale. Environ. Int. 142, 105856 https://doi.org/10.1016/j.envint.2020.105856.

Marécal, V., Peuch, V.-H., Andersson, C., Andersson, S., Arteta, J., Beekmann, M., Benedictow, A., Bergström, R., Bessagnet, B., Cansado, A., Chéroux, F., Colette, A., Coman, A., Curier, R.L., Denier Van Der Gon, H.A.C., Drouin, A., Elbern, H., Emili, E., Engelen, R.J., Eskes, H.J., Foret, G., Friese, E., Gauss, M., Giannaros, C., Guth, J., Joly, M., Jaumouillé, E., Josse, B., Kadygrov, N., Kaiser, J.W., Krajsek, K., Kuenen, J., Kumar, U., Liora, N., Lopez, E., Malherbe, L., Martinez, I., Melas, D., Meleux, F., Menut, L., Moinat, P., Morales, T., Parmentier, J., Piacentini, A., Plu, M., Poupkou, A., Queguiner, S., Robertson, L., Rouïl, L., Schaap, M., Segers, A., Sofiev, M., Tarasson, L., Thomas, M., Timmermans, R., Valdebenito, Á., Van Velthoven, P., Van Versendaal, R., Vira, J., Ung, A., 2015. A regional air quality

forecasting system over Europe: the MACC-II daily ensemble production. Geosci. Model Dev 8, 2777–2813. https://doi.org/10.5194/gmd-8-2777-2015.

Neis, P., Zielstra, D., Zipf, A., 2011. The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. Futur. Internet 2012, Vol. 4, Pages 1-21 4, 1–21. https://doi.org/10.3390/FI4010001.

OpenStreetMap contributors, 2020. Planet dump retrieved from https://planet.osm.org [WWW Document]. URL https://www.openstreetmap.org.

Ortiz, A., Guerreiro, C., 2020. Air Quality in Europe - 2020 report. https://doi.org/10.2800/786656.

R Core Team, 2020. R: A language and environment for statistical computing.

Requia, W.J., Di, Q., Silvern, R., Kelly, J.T., Koutrakis, P., Mickley, L.J., Sulprizio, M.P., Amini, H., Shi, L., Schwartz, J., 2020. An ensemble learning approach for estimating high spatiotemporal resolution of ground-level ozone in the contiguous United States. Environ. Sci. Technol. 54, 11037–11047. https://doi.org/10.1021/acs.est.0c01791.

Shen, Y., de Hoogh, K., Schmitz, O., Clinton, N., Tuxen-Bettman, K., Brandt, J., Christensen, J.H., Frohn, L.M., Geels, C., Karssenberg, D., Vermeulen, R., Hoek, G., 2021a. EXPANSE_algorithm - version 0.1.0 [WWW Document]. URL https://github.com/co822ee/EXPANSE_algorithm.

Shen, Y., de Hoogh, K., Schmitz, O., Clinton, N., Tuxen-Bettman, K., Brandt, J., Christensen, J.H., Frohn, L.M., Geels, C., Karssenberg, D., Vermeulen, R., Hoek, G., 2021b. expanse_multiyear - version 0.1.0 [WWW Document]. URL https://github.com/co822ee/expanse_multiyear.

Shtein, A., Karnieli, A., Katra, I., Raz, R., Levy, I., Lyapustin, A., Dorman, M., Broday, D. M., Kloog, I., 2018. Estimating daily and intra-daily PM10 and PM2.5 in Israel using a spatio-temporal hybrid modeling approach. Atmos. Environ. 191, 142–152. https://doi.org/10.1016/j.atmosenv.2018.08.002.

Stafoggia, M., Oftedal, B., Chen, J., Rodopoulou, S., Renzi, M., Atkinson, R.W., Bauwelinck, M., Klompmaker, J.O., Mehta, A., Vienneau, D., Andersen, Z.J., Bellander, T., Brandt, J., Cesaroni, G., de Hoogh, K., Fecht, D., Gulliver, J., Hertel, O., Hoffmann, B., Hvidtfeldt, U.A., Jöckel, K.H., Jørgensen, J.T., Katsouyanni, K., Ketzel, M., Kristoffersen, D.T., Lager, A., Leander, K., Liu, S., Ljungman, P.L.S., Nagel, G., Pershagen, G., Peters, A., Raaschou-Nielsen, O., Rizzuto, D., Schramm, S., Schwarze, P.E., Severi, G., Sigsgaard, T., Strak, M., van der Schouw, Y.T., Verschuren, M., Weinmayr, G., Wolf, K., Zitt, E., Samoli, E., Forastiere, F., Brunekreef, B., Hoek, G., Janssen, N.A.H., 2022. Long-term exposure to low ambient air pollution concentrations and mortality among 28 million people: results from seven large European cohorts within the ELAPSE project. Lancet. Planet. Heal. 6, e9–e18. https://doi.org/10.1016/S2542-5196(21)00277-1.

Van Donkelaar, A., Martin, R.V., Li, C., Burnett, R.T., 2019. Regional estimates of chemical composition of fine particulate matter using a combined geoscience-statistical method with information from satellites, models, and monitors. Environ. Sci. Technol. 53, 2595–2611. https://doi.org/10.1021/acs.est.8b06392.

Van Donkelaar, A., Martin, R.V., Spurr, R.J.D., Burnett, R.T., 2015. High-resolution satellite-derived PM2.5 from optimal estimation and geographically weighted regression over North America. Environ. Sci. Technol. 49, 10482–10491. https://doi.org/10.1021/acs.est.5b02076.

Wang, R., Henderson, S.B., Sbihi, H., Allen, R.W., Brauer, M., 2013. Temporal stability of land use regression models for traffic-related air pollution. Atmos. Environ. 64, 312–319. https://doi.org/10.1016/j.atmosenv.2012.09.056.

WHO, 2021. Ambient (outdoor) air pollution [WWW Document]. URL https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health (accessed 10.19.21).

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.

Wright, M.N., Ziegler, A., 2017. ranger}: A Fast Implementation of Random Forests for High Dimensional Data in {C++ and {R}. J. Stat. Softw. 77, 1–17. https://doi.org/10.18637/jss.v077.i01.

Wu, B., Li, R., Huang, B., 2014. A geographically and temporally weighted autoregressive model with application to housing prices. Int. J. Geogr. Inf. Sci. 28, 1186–1204. https://doi.org/10.1080/13658816.2013.878463.

Zhan, Y., Luo, Y., Deng, X., Zhang, K., Zhang, M., Grieneisen, M.L., Di, B., 2018. Satellite-Based Estimates of Daily NO2 Exposure in China Using Hybrid Random Forest and Spatiotemporal Kriging Model. Environ. Sci. Technol. 52, 4180–4189. https://doi.org/10.1021/acs.est.7b05669.