Data Integration Model for Air Quality: A Hierarchical Approach to the Global Estimation of Exposures to Ambient Air Pollution

Gavin Shaddick[i, ii], Matthew L. Thomas[ii], Amelia Green[ii], Michael Brauer[iii], Aaron van Donkelaar[iv], Rick Burnett[v], Howard H. Chang[vi], Aaron Cohen[vii], Rita Van Dingenen[viii], Carlos Dora[ix], Sophie Gumy[ix], Yang Liu[x], Randall Martin[iv], Lance A. Waller[vi], Jason West[xi], James V. Zidek[xii] and Annette Prüss-Ustün[ix]

[i] *College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK*
[ii] *Department of Mathematical Sciences, University of Bath, U.K.*
[iii] *School of Population and Public Health, The University of British Columbia, Canada*
[iv] *Department of Physics and Atmospheric Science, Dalhousie University, Canada*
[v] *Health Canada, Ottawa, Canada*
[vi] *Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, U.S.A.*
[vii] *Health Effects Institute, Boston, U.S.A.*
[viii] *Institute for Environment and Sustainability, Joint Research Centre, European Commission, Italy*
[ix] *World Health Organization, Geneva, Switzerland*
[x] *Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, U.S.A.*
[xi] *Department of Environmental Sciences and Engineering, University of North Carolina, Chapel Hill, U.S.A.*
[xii] *Department of Statistics, University of British Columbia, Canada*

April 13, 2017

Abstract

Air pollution is a major risk factor for global health, with 3 million deaths annually being attributed to fine particulate matter ambient pollution ($PM_{2.5}$). The primary source of information for estimating population exposures to air pollution has been measurements from ground monitoring networks but, although coverage is increasing, there remain regions in which monitoring is limited. The Data Integration Model for Air Quality (DIMAQ) supplements ground monitoring data with information from other sources, such as satellite retrievals of aerosol optical depth and chemical transport models. Set within a Bayesian hierarchical modelling framework, the model allows spatially-varying relationships between ground measurements and other factors that estimate air quality. The model is used to estimate exposures, together with associated measures of uncertainty, on a high resolution grid covering the entire world from which it is estimated that 92% of the world's population reside in areas exceeding the World Health Organization's Air Quality Guidelines.

1 Introduction

Ambient air pollution poses a significant threat to global health and has been associated with a range of adverse health effects, including cardiovascular and respiratory diseases in addition to some cancers (Brook et al., 2010; Hoek et al., 2013; Loomis et al., 2013; Newby et al., 2014; Sava and Carlsten, 2012; WHO, 2013). Fine particulate matter ($PM_{2.5}$) in particular has been established as a key driver of global health with an estimated 3 million deaths in 2014 being attributable to $PM_{2.5}$ (WHO, 2016a). It has been estimated that the majority of the world's population (87%) reside in areas in which the World Health Organization (WHO) air quality guideline (annual mean of 10 μgm^{-3}) for $PM_{2.5}$ is exceeded (Brauer et al., 2015).

It is vital that the subsequent risks, trends and consequences of air pollution are monitored and modelled to develop effective environmental and public health policy to lessen the burden of air pollution. Accurate measurements of exposure in any given area are required but this is a demanding task: the processes involved are extremely complex and ground monitoring is scarce in many regions. The locations of ground monitoring sites within the WHO Air Pollution in Cities database (WHO, 2016b) are shown in Figure 1 where it can be seen that the density of monitoring sites varies considerably, with extensive measurements available in North America, Europe, China and India but with little or no measurement data available for large areas of Africa, South America and the Middle East.

For this reason, there is a need to use information from other sources in order to obtain estimates of exposures for all areas of the world. In 2013, the Global Burden of Disease study (henceforth referred to as GBD2013), as described in Forouzanfar et al. (2015), used a regression calibration approach to utilise information from satellite remote sensing and chemical transport models to create a set of estimates of exposures on a high–resolution grid ($0.1^{\circ} \times 0.1^{\circ}$, approximately 11km $\times$ 11km at the equator) that were then matched to population estimates to estimate disease burden. In GBD2013, a fused estimate of $PM_{2.5}$, calculated as the average of estimates from satellites and chemical transport models, was calibrated against ground measurements using linear regression. For cells that contained a ground monitor, measurements were regressed against this fused estimate in conjunction with information related to local monitoring networks (Brauer et al., 2015). The resulting calibration function was applied to all grid cells, allowing a comprehensive set of global estimates of $PM_{2.5}$ to be produced.

This allowed data from the three sources to be utilised, but the use of a single, global, calibration function resulted in underestimation in a number of areas (Brauer et al., 2015). In reality, the relationships between ground measurements and estimates from other sources will vary spatially due to regional differences in biases and errors that will be present in the different methods of estimation. Recently, Van Donkelaar et al. (2016) extended this approach using geographically weighted regression (GWR) to allow calibration (between the measurements and estimates) equations to vary spatially and to utilise additional information related to land use and the chemical composition of particulate matter. However, both the original linear regression and GWR approaches only provide an informal analysis of the uncertainty associated with the resulting estimates of exposure.

In addition to regional differences in calibration functions, additional challenges arise when combining data that are generated in fundamentally different ways. Satellite pixels and chemical transport model cells are not the same with each potentially not capturing different micro-scale features that may be reflected in the ground measurements and all three sources of data will have different error structures that may not align. The difference in resolution between ground monitors (point locations) and estimates from satellite and chemical transport models (grid cells) has led to the use of spatially varying coefficient models, often referred to as *downscaling models* (Chang, 2016). In the purely spatial model presented in Berrocal et al. (2010) for example, the intercepts and coefficients are assumed to arise from a continuous bivariate spatial process. Downscaling/upscaling models, set within a Bayesian hierarchical framework, have been used for both spatial and spatio-temporal modelling of air pollution

with examples including Guillas et al. (2006), who used the UIUC 2-D chemical-transport model of the global atmosphere; Van de Kassteele et al. (2006), who modelled PM_{10} concentrations over Western Europe using information from both satellite observations and a chemical transport model; McMillan et al. (2010) who modelled $PM_{2.5}$ in the North Eastern U.S. using estimates from the Community Multi-scale Air Quality (CMAQ) numerical model; Kloog et al. (2014) who modelled $PM_{2.5}$ in the Northeastern U.S. using satellite-based aerosol optical depth (AOD) and Berrocal et al. (2010) and Zidek et al. (2012) who modelled ozone in the Eastern U.S. (Eastern and Central in the case of Zidek *et al.*) using estimates from CMAQ and a variant of the MAQSIP (Multiscale Air Quality Simulation Platform) model respectively.

An alternative approach to the calibration used in downscaling is *Bayesian melding* (Poole and Raftery, 2000) in which both the measurements and the estimates are assumed to arise from an underlying latent process that represents the true level of the pollutant. This latent process itself is unobservable but measurements can be taken, possibly with error, at locations in space and time. For example, the underlying latent process represents the true level of $PM_{2.5}$ and this gives rise to the measurements from ground monitors and the estimates from satellite remote sensing and atmospheric models, all of which will inform the posterior distribution of the underlying latent process. Bayesian melding has been used to model SO_2 in the Eastern U.S. combining ground measurements with information from the Models-3 air quality model (Fuentes and Raftery, 2005).

In this paper, a model is presented for integrating data from multiple sources, that enables accurate estimation of global exposures to fine particulate matter. Set within a Bayesian hierarchical framework, this Data Integration Model for Air Quality (DIMAQ) estimates exposures, together with associated measures of uncertainty, at high geographical resolution by utilising information from multiple sources and addresses many of the issues encountered with previous approaches. The structure of the paper is as follows: after this introduction, Section 2 provides details of the data that are available, including measurements from ground monitoring and estimates from satellites and chemical transport models. Section 3 provides details of the statistical model (DIMAQ) that is used to integrate data from these different sources and methods for inference when performing Bayesian analysis with large datasets. In Section 4 the results of applying DIMAQ are presented, including examples of global and country specific estimates of exposure to $PM_{2.5}$ together with details of the methods used for model evaluation and comparison. Finally, Section 5 provides a concluding summary and a discussion of potential areas for future research.

2 Data

The sources of data used here can be allocated to one of three groups: (i) ground monitoring data; (ii) estimates of $PM_{2.5}$ from remote sensing satellites and chemical transport models; (iii) other sources including population, land–use and topography. Ground monitoring is available at a distinct number of locations, whereas the latter two groups provide near complete global coverage (and have previously been shown to have strong associations with global concentrations of $PM_{2.5}$, see below for details). Utilising such data will allow estimates of exposures to be made for all areas, including those for which ground monitoring is sparse or non-existent.

2.1 Ground measurements

Ground measurements were available for locations reported within the WHO Air Pollution in Cities database (WHO, 2016b), but rather than using the city averages reported in that database, monitor-specific measurements are used. The result was measurements of concentrations of PM_{10} and $PM_{2.5}$ from 6,003 ground monitors. The locations and annual average concentrations for these monitors can be seen in Figure 1. The database was compiled to represent measurements in 2014 with the majority

of measurements coming from that year (2760 monitors). Where data were not available for 2014, data were used from 2015 (18 monitors), 2013 (2155), 2012 (564), 2011 (60), 2010 (375), 2009 (49), 2008 (21) and 2006 (1). In addition to annual average concentrations, additional information related to the ground measurements was also included where available, including monitor geo coordinates and monitor site type.

For locations measuring only PM_{10}, $PM_{2.5}$ measurements were estimated from PM_{10}. This was performed using a locally derived conversion factor ($PM_{2.5}/PM_{10}$ ratio, for stations where measurements are available for the same year) that was estimated using population-weighted averages of location-specific conversion factors for the country as detailed in Brauer et al. (2012). If country-level conversion factors were not available, the average of country-level conversion factors within a region were used.

2.2 Satellite-based estimates

Satellite remote sensing is a method that estimates pollution from satellite retrievals of aerosol optical depth (AOD), a measurement of light extinction by aerosols in the atmosphere. AOD indicates how aerosols modify the radiation leaving the top of the atmosphere after being scattered by the Earth's atmosphere and surface. Estimates of $PM_{2.5}$ are obtained by correcting AOD using a spatially varying term, η,

$$PM_{2.5} = \eta \times \text{AOD} \; .$$

Here η is the coincident ratio of $PM_{2.5}$ to AOD and accounts for local variation in vertical structure, meteorology, and aerosol type. This ratio is simulated from the GEOS-Chem global chemical transport model (Bey et al., 2001).

The estimates used here combine AOD retrievals from multiple satellites with simulations from the GEOS-Chem chemical transport model and land use information, produced at a spatial resolution of $0.1^{\circ} \times 0.1^{\circ}$, which is approximately 11km $\times$ 11km at the equator. This is described in detail in Van Donkelaar et al. (2016). A map of the estimates of $PM_{2.5}$ from this model can be seen in Figure 2.

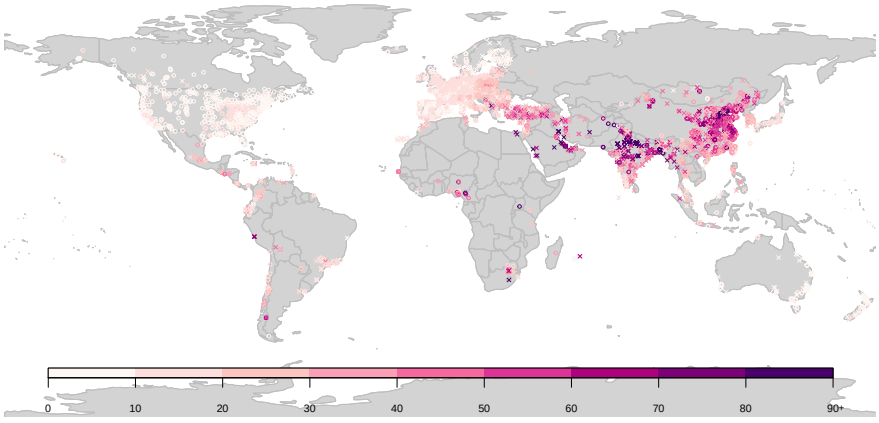


Figure 1: Locations of ground monitors measuring $PM_{2.5}$ (circles) and PM_{10} (crosses). Colours denote the annual average concentrations (μgm^{-3}) of $PM_{2.5}$ (or $PM_{2.5}$ converted from PM_{10}). Data are from 2014 (46%), 2013 (36%), 2012 (9%) and 2006-2011, 2015 (9%).

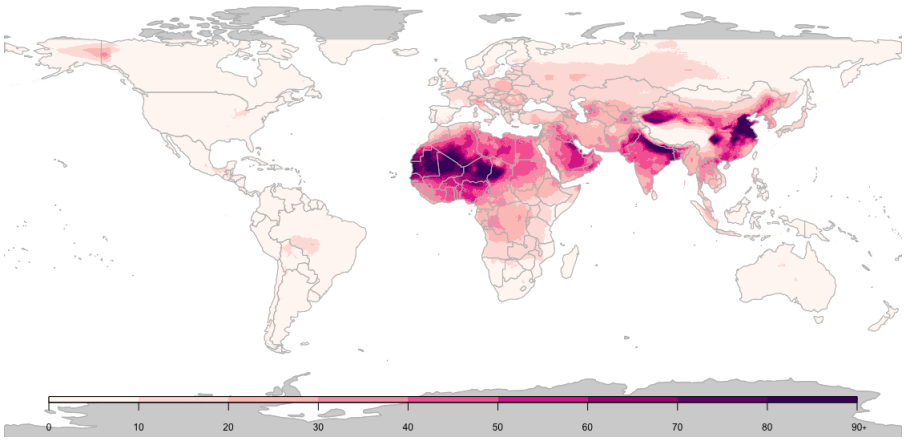


Figure 2: Satellite-based estimates of $PM_{2.5}$ (μgm^{-3}) for 2014, by grid cell ($0.1^\circ \times 0.1^\circ$ resolution).

2.3 Chemical transport model simulations

Numerically simulated estimates of $PM_{2.5}$ were obtained from atmospheric chemical transport models. There are a variety of such models available including GEOS-Chem (Bey et al., 2001), TM5 (Huijnen et al., 2010) and TM5-FASST (Van Dingenen et al., 2014). The first two of these are nested 3-dimensional global atmospheric transport models which can be used to simulate levels of $PM_{2.5}$ with TM5-FASST being a reduced form of the full TM5 model, developed to allow faster computation for impact assessment (Van Dingenen et al., 2014). Estimates at a spatial resolution of $1^\circ \times 1^\circ$ were allocated to a higher resolution grid, of $0.1^\circ \times 0.1^\circ$, based on population density (Brauer et al., 2012, 2015). Estimates for $PM_{2.5}$ from the TM5-FASST model were available for 2010, as described in Brauer et al. (2015). A map of these estimates can be seen in Figure 3a.

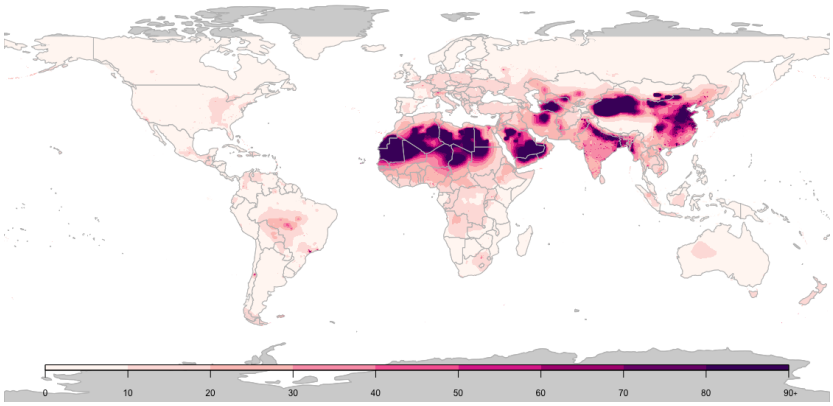
In addition to the estimates of $PM_{2.5}$, estimates of the sum of sulphate, nitrate, ammonium and organic carbon (SNAOC) and the compositional concentrations of mineral dust (DUST) based on simulations from the GEOS-Chem chemical transport model (Van Donkelaar et al., 2016) were available for 2014. Maps of the estimates of SNAOC and DUST can be seen in Figures 3b and 3c respectively.

2.4 Population data

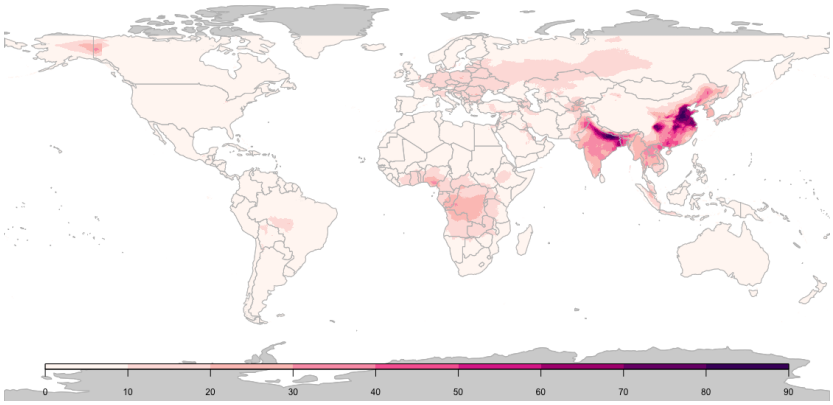
A comprehensive set of population data on a high-resolution grid was obtained from the Gridded Population of the World (GPW) database (GPW, 2016). These data are provided on a $0.0417^\circ \times 0.0417^\circ$ resolution. Aggregation to each $0.1^\circ \times 0.1^\circ$ grid cell was performed as detailed in Brauer et al. (2015). GPW version 4 provides population estimates for 2000, 2005, 2010, 2015 and 2020. Following the methodology used in Brauer et al. (2015), populations for 2014 were obtained by interpolation using cubic splines (performed for each grid cell) with knots placed at 2000, 2005, 2010, 2015 and 2020. A map of the resulting estimates of populations for 2014 can be seen in Figure 4.

2.5 Land use

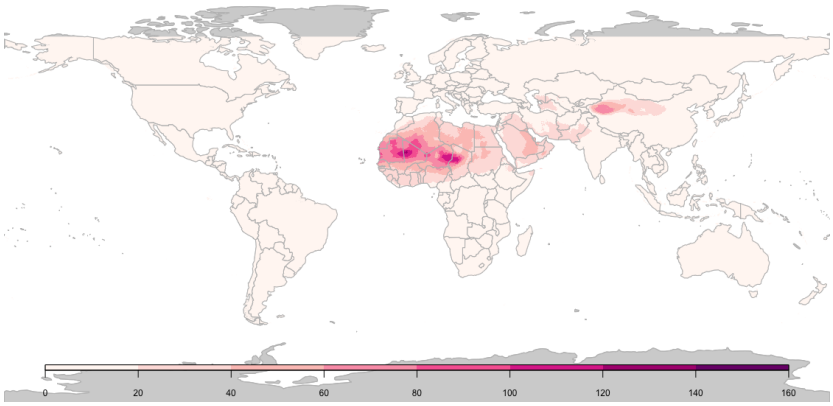
Van Donkelaar et al. (2016) developed a measure combining information on elevation and land use that was shown to be a significant predictor of $PM_{2.5}$. For each ground monitor, the following are calculated: (i) the difference between the elevation (of the ground monitor) and that of the surrounding grid cell, as defined by the GEOS-Chem chemical transport model (ED); (ii) the distance to the nearest urban



(a) Estimates of $PM_{2.5}$ (μgm^{-3}) for 2010 from the TM5 chemical transport model used in GBD2013.



(b) Estimates of the sum of sulphate, nitrate, ammonium and organic carbon (μgm^{-3}) for 2014 from the GEOS-Chem chemical transport model.



(c) Estimates of the compositional concentrations of mineral dust (μgm^{-3}) for 2014 from the GEOS-Chem chemical transport model.

Figure 3: Estimates from chemical transport models, by grid cell ($0.1^\circ \times 0.1^\circ$ resolution).

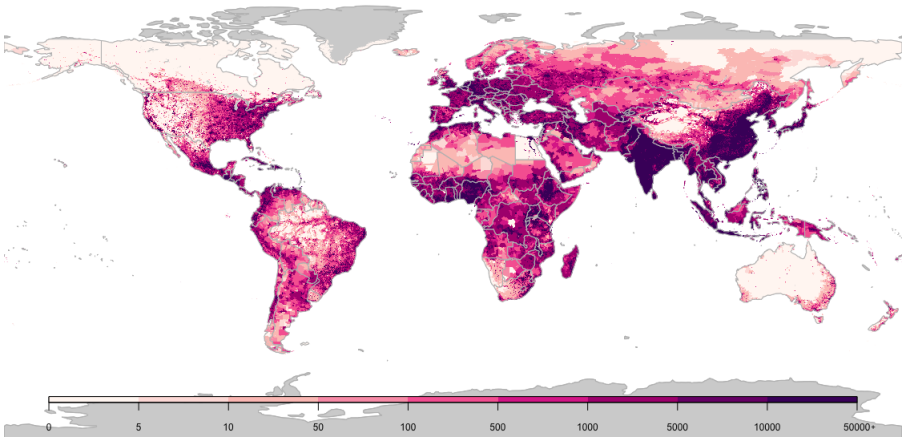


Figure 4: Population estimates for 2014 from the Gridded Population of the World version 4 (GPW v4) database, by grid cell ($0.1^\circ \times 0.1^\circ$ resolution).

land surface (DU), based upon MODIS land cover descriptions (Friedl et al., 2010). The resulting measure (ED × DU) was available for 2014 for each $0.1^\circ \times 0.1^\circ$ grid cell.

3 Statistical Modelling

The aim is to obtain estimates of concentrations of $PM_{2.5}$ for each of 1.4 million grid cells, together with associated measures of uncertainty. This will be achieved by finding the posterior distributions for each cell, from which summary measures will be calculated.

The overall approach is statistical calibration as described in Chang (2016): a regression model is used to express ground measurements, Y_s, available at a discrete set of N_S locations $S \in \mathcal{S}$ with labels $S = \{s_0, s_1, \ldots, s_{N_S}\}$, that are a function of covariates, X_{sr}: $r = 1, \ldots, R$, that reflect information from other sources, as described in Section 2. Covariate information may be available for point locations (as with the ground measurements) or on a grid of N_L cells, $l \in L$ where $L = l_1, ..., l_{N_L}$.

Considering a single covariate, X_{lr}, for ease of explanation,

$$Y_s = \tilde{\beta}_{0s} + \tilde{\beta}_{1s} X_{lr} + \epsilon_s \tag{1}$$

where X_{lr} is measured on a grid. Here, $\epsilon_s \sim N(0, \sigma^2_\epsilon)$ is a random error term. The terms $\tilde{\beta}_{0s}$ and $\tilde{\beta}_{1s}$ denote random effects that allow the intercept and coefficient to vary over space

$$\begin{aligned} \tilde{\beta}_{0s} &= \beta_0 + \beta_{0s} \\ \tilde{\beta}_{1s} &= \beta_1 + \beta_{1s} \, . \end{aligned}$$

Here, β_0 and β_1 are fixed effects representing the mean value of the intercept and coefficients respectively, with β_{0s} and β_{1s} zero mean spatial random effects providing (spatially driven) adjustments to these means, allowing the calibration functions to vary over space. In downscaling models, it is assumed that the parameters β_{0s} and β_{1s} arise from a continuous spatial process which allows within grid cell variation (see Berrocal et al. (2010) for an example using a continuous bivariate spatial process).

Despite monitoring data being increasingly available, there are issues that may mean using a spatial continuous process may be problematic in this setting. Monitoring protocols, measurement techniques,

quality control procedures and mechanisms for obtaining annual averages may vary from country-to-country (Brauer et al., 2012) leading to natural discontinuities in ground measurements, and their precision, between countries. In addition, the geographic distribution of measurements, as seen in Figure 1, is heavily biased toward North America, Europe, China and India with some areas of the world, e.g. Africa, having very little monitoring information to inform such a model. Therefore, the spatial random effects used here are based on country level geography rather than continuous spatial processes.

The structure of the random effects used here exploits a geographical nested hierarchy: each of the 187 countries considered are allocated to one of 21 regions and, further, to one of 7 super-regions. Each region must contain at least two countries and is broadly based on geographic regions/sub-continents and groupings based on country level development status and causes of death (Brauer et al., 2012). The geographical structure of regions within super-regions can be seen in Figure 5. Where there are limited monitoring data within a country, information can be borrowed from higher up the hierarchy, i.e. from other countries within the region and further, from the wider super-region. It is noted that the 'high income' super-region is non-contiguous and for North Africa/Middle East the region is the same as the super-region and therefore will be a single set of random effects, i.e. no distinction between region and super-region, for this area.

3.1 A Data Integration Model for Air Quality

Annual averages of ground measurements (of $PM_{2.5}$) at point locations, s, within grid cell, l, country, i, region, j, and super–region, k are denoted by Y_{slijk}. As described in Section 3, there is a nested hierarchical structure with $s = 1, \ldots, N_{lijk}$ sites within grid cell, l; $l = 1, \ldots, N_{ijk}$, grid cells within country i; $i = 1, \ldots, N_{jk}$, countries within region j; $j = 1, \ldots, N_k$, regions within super–region k: $k = 1, \ldots, N$. In order to allow for the skew in the measurements and the constraint of non–negativity, the (natural) logarithm of the measurements are used.

The model consists of sets of fixed and random effects, for both intercepts and covariates, and is given as follows,

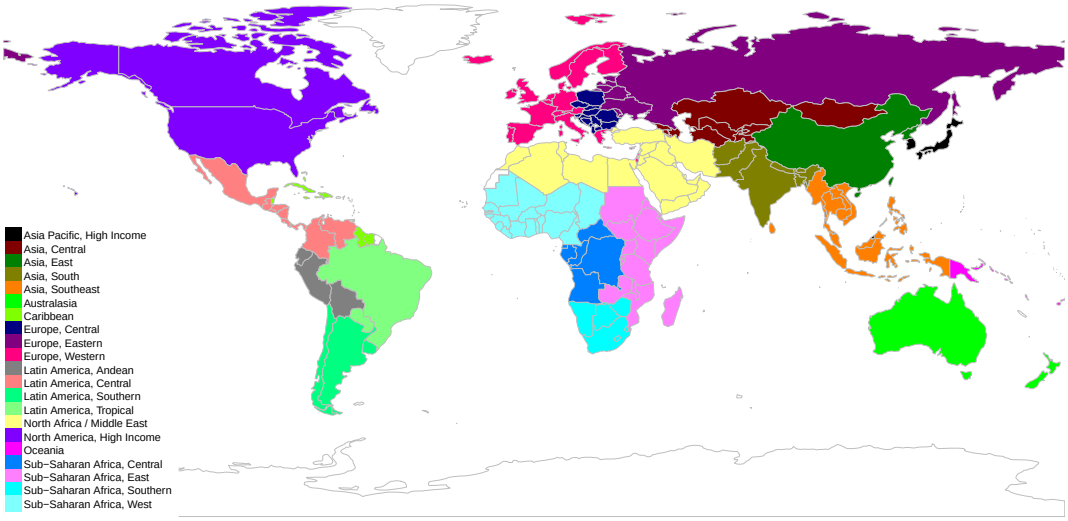
$$\begin{aligned} \log(Y_{slijk}) = \tilde{\beta}_{0,lijk} &+ \sum_{q \in Q} \tilde{\beta}_{q,ijk} X_{q,lijk} \\ &+ \sum_{p_1 \in P_1} \beta_{p_1} X_{p_1,lijk} + \sum_{p_2 \in P_2} \beta_{p_2} X_{p_2,slijk} \\ &+ \epsilon_{slijk} \ , \end{aligned} \tag{2}$$

where $\epsilon_{slijk} \sim N(0, \sigma^2_\epsilon)$ is a random error term. A set of R covariates contains two groups, $R = (P, Q)$, where P are those which have fixed effects (across space) and Q those assigned random effects. The main estimates of air quality, e.g. those from satellites and chemical transport models, will be assigned random effects and are in Q, with other variables being assigned fixed effects. Within the group, P, of covariates that have fixed effects; P_1 are available at the grid cell level, l, with others, P_2, being available for the point locations, s, of the monitors, $P = (P_1, P_2)$.

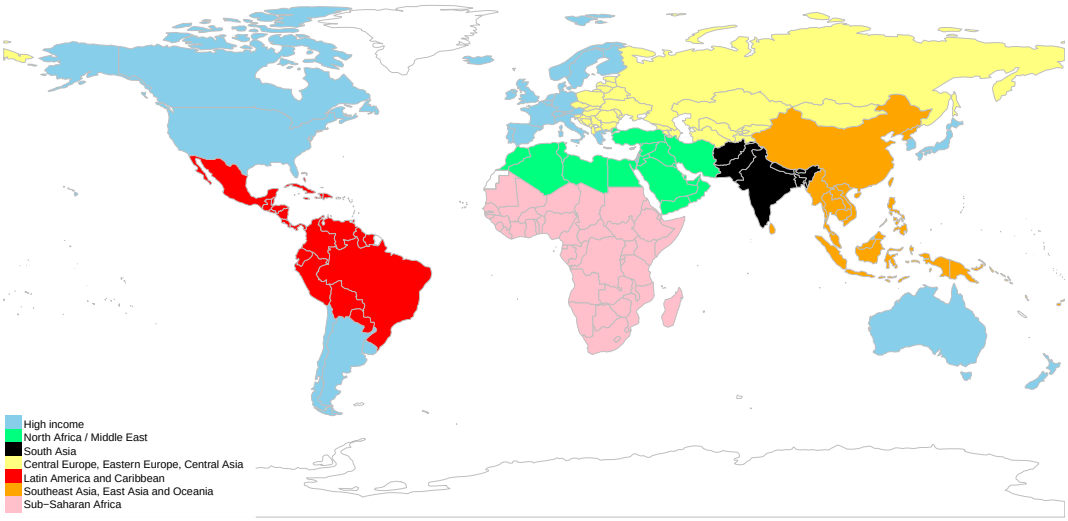
3.1.1 Structure of the random effects

Here, the random effect terms, $\tilde{\beta}_{0,lijk}$ and $\tilde{\beta}_{q,ijk}$, have contributions from the country, the region and the super–region, with the intercept also having a random effect for the cell representing within-cell variation in ground measurements,

$$\begin{aligned} \tilde{\beta}_{0,lijk} &= \beta_0 + \beta^G_{0,lijk} + \beta^C_{0,ijk} + \beta^R_{0,jk} + \beta^{SR}_{0,k} \\ \tilde{\beta}_{q,ijk} &= \beta_q + \beta^C_{q,ijk} + \beta^R_{q,jk} + \beta^{SR}_{q,k} \ . \end{aligned}$$



(a) Map of regions.



(b) Map of super-regions.

Figure 5: Schematic showing the nested geographical structure of countries within regions within super-regions.

For clarity of exposition, the following description is restricted to a generic parameter, β. Let β_k^{SR} denote the coefficient for super-region k. The coefficients for super-regions are distributed with mean equal to the overall mean (β_0, the fixed effect) and variance, σ^2_{SR}, representing between super-region variability,

$$\beta_k^{SR} \sim N(\beta_0, \sigma^2_{SR})$$

where $k = 1, \ldots, N = 7$. Similarly, each super-region contains a number of regions. Let β_{jk}^{R} denote the coefficient for region j (in super–region k) that will be distributed with mean equal to to the coefficient for the super-region and variance representing the between region (within super–region) variability,

$$\beta_{jk}^{R} \sim N(\beta_k^{SR}, \sigma^2_{R,k}),$$

where $j = 1, \ldots, N_k$, the number of regions in super-region j. Each region will contain a number of countries. Let β_{ijk}^{C} denote the coefficient for country i in region j and super-region k. The country level effect will be distributed with mean equal to the coefficient for region j within super-region k with variance representing the between country (within region) variability,

$$\beta_{ijk}^{C} \sim N(\beta_{jk}^{R}, \sigma^2_{C,jk}), \tag{3}$$

where $i = 1, \ldots, N_{jk}$ is the number of countries in region j (in super–region k). Note that in the case of the intercepts, there is an additional term, β_{lijk}^{G}, representing within grid cell (between monitoring locations) variability.

Country effects within regions and regional effects within super–regions are assumed to be independent within their respective geographies. However, the geographical hierarchy is broadly based on geographic regions, sub-continents, mortality and economic factors (Brauer et al., 2012) and, as such, there are countries for which the allocation may not be optimal when considering environmental factors, such as air pollution. For example, Mongolia is included within the Asia Central region and Central Eastern Europe and Central Asia super–region (see Figure 5) but its pollution profile might be expected to be more similar to those of its direct neighbours, including China (which is in a different region (Asia East)) and super–region (South East Asia, East Asia and Oceania), than the profiles of more western countries. For this reason, it might be advantageous to allow the borrowing of information in Equation (3) to include countries that are immediate neighbours rather than all of the countries in the surrounding administrative region. This could be achieved using an intrinsic conditionally autoregressive (ICAR) model (Besag, 1974) in place of Equation (3),

$$\beta_i^C | \beta_{i'}^C,\ i' \in \partial_i \sim N\left(\overline{\beta}_i^C, \frac{\psi^2}{N_{\partial_i}}\right),$$

where ∂_i is the set of neighbours of country i, $N_{\partial i}$ is the number of neighbours, and $\overline{\beta}_i^C$ is the mean of the spatial random effects of these neighbours.

3.1.2 Hyperpriors

Gaussian priors, $N(0, \sigma^2)$, are assigned to each of the fixed effects β_0 and β_q where $\sigma^{-2} = 0.0001$. Gamma priors, $Ga(a, b)$ are assigned to the log of the precisions, i.e. $\log(\sigma^{-2}_{G,i}), \log(\sigma^{-2}_{C,j}), \log(\sigma^{-2}_{R,j}), \log(\sigma^{-2}_{SR})$ and $\log(\psi^{-2})$, with $a = 1, b = 0.00005$.

3.2 Inference

The model presented in Section 3.1 is a Latent Gaussian Model (LGM) and therefore advantage can be taken of methods offering efficient computation when performing Bayesian inference. LGMs can be implemented using approximate Bayesian inference using integrated nested Laplace approximations (INLA) as proposed in Rue et al. (2009) using the R-INLA software (Rue et al., 2012). The following sections provide a brief summary of LGMs (Section 3.2.1) and INLA (Section 3.2.2) with additional details linking to the model described in Section 3.1.

3.2.1 Latent Gaussian models

The model presented in Equation (2) can be expressed in general form as follows: Given $\eta_s = g(E(Y_s))$, where $g(\cdot)$ is a link function,

$$\eta_s = \beta_0 + \sum_{p=1}^{P} \beta_p X_{qs} + \sum_{q=1}^{Q} f_q(Z_{qs})$$

where β_0 is an overall intercept term, the set of β_p $(p = 1, \ldots, P)$ are the coefficients associated with covariates, X; the fixed effects. The set of functions, $f_1(\cdot), \ldots, f_Q(\cdot)$ represent the random effects with the form of the function being determined by the model. For example, a hierarchical model may have $f_1(\cdot) \sim N(0, \sigma_f^2)$, with a distribution defined for σ_f^2, whereas for standard regression, $f(\cdot) \equiv 0$, leaving just fixed effects.

The set of unknown parameters, $\boldsymbol{\theta}$, will include both the coefficients of the model shown above and the parameters required for the functions, i.e. $\boldsymbol{\theta} = (\beta_p, f_q)$. Here $\boldsymbol{\theta}$ will contain the parameters of the model as described in Section 3.1 and will include $\beta_0, \beta^G_{0,lijk}, \beta^C_{0,ijk}, \beta^R_{0,jk}, \beta^{SR}_{0,k}, \beta_q, \beta^C_{q,ijk}, \beta^R_{q,jk}$ and $\beta^{SR}_{q,k}$, with the set of hyperparameters associated with $\boldsymbol{\theta}$ being $\boldsymbol{\psi}_2 = (\sigma^2_{G,i}, \sigma^2_{C,j}, \sigma^2_{R,j}, \sigma^2_{SR})$. The overall set of parameters, $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)$, also contains $\boldsymbol{\psi}_1 = (\sigma^2_\epsilon)$, which relates to the variance of the measurement error in the data.

Assigning a Gaussian distribution to the parameters in $\boldsymbol{\theta}$, $\boldsymbol{\theta}|\boldsymbol{\psi} \sim MVN(\mathbf{0}|\Sigma(\boldsymbol{\psi}_2))$ will result in a LGM. The computation required to perform inference will be largely determined by the characteristics of the covariance matrix, $\Sigma(\boldsymbol{\psi}_2)$, which will often be dense, i.e. it will have many entries that are non-zero, leading to a high computational burden when performing the matrix inversions that will be required to perform inference. If $\boldsymbol{\theta}|\boldsymbol{\psi}_2$ can be expressed in terms of a Gaussian Markov random field (GMRF), then it may be possible to take advantage of methods that reduce computation when performing Bayesian analysis on models of this type (Rue and Held, 2005). Using a GMRF means that typically the inverse of the covariance matrix, $Q = \Sigma^{-1}$ will be sparse (i.e. more zero entries) due to the conditional independence between sets of parameters in which $\theta_l \perp\!\!\!\perp \theta_m|\boldsymbol{\theta}_{-lm} \iff Q_{lm} = 0$ (where $-lm$ denotes the vector of $\boldsymbol{\theta}$ with the l and m elements removed) (Rue and Held, 2005). Expressing $\boldsymbol{\theta}|\boldsymbol{\psi}_2$ in terms of the precision, rather than the covariance, gives $\boldsymbol{\theta}|\boldsymbol{\psi} \sim MVN(\mathbf{0}|Q(\boldsymbol{\psi}_2)^{-1})$, where $\boldsymbol{\psi}_2$ denotes the parameters associated with Q rather than Σ.

3.2.2 Integrated Laplace approximations

Estimation of the (marginal) distributions of the model parameters and hyperparameters of a LGM will require evaluation of the following integrals:

$$\begin{aligned} p(\theta_j|\boldsymbol{Y}) &= \int p(\theta_j|\boldsymbol{Y}, \boldsymbol{\psi}) p(\boldsymbol{\psi}|\boldsymbol{Y}) d\boldsymbol{\psi} \\ p(\psi_k|\boldsymbol{Y}) &= \int p(\boldsymbol{\psi}|\boldsymbol{Y}) d\boldsymbol{\psi}_{-k} \end{aligned} \tag{4}$$

In all but the most stylised cases, these will not be analytically tractable. Samples from these distributions could be obtained using Markov chain Monte Carlo (MCMC) methods but there may be issues when fitting LGMs using MCMC, as described in Rue et al. (2009), and the computational burden may be excessive, especially large numbers of predictions are required. Here, approximate Bayesian inference is performed using INLA. It is noted that the dimension of $\boldsymbol{\theta}$ is much larger than the dimension of $\boldsymbol{\psi}$ and this will help in the implementation of the model as the computational burden increases linearly with the dimension of $\boldsymbol{\theta}$ but exponentially with the dimension of $\boldsymbol{\psi}$.

The aim is to find approximations for the distributions shown in Equation (4). For the hyperparameters, the posterior of $\boldsymbol{\psi}$ given $\boldsymbol{Y}$ can be written as

$$\begin{aligned} p(\boldsymbol{\psi}|\boldsymbol{Y}) &= \frac{p(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{Y})}{p(\boldsymbol{\theta}|\boldsymbol{\psi},\boldsymbol{Y})} \\ &\propto \frac{p(\boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\psi})p(\boldsymbol{\psi})}{p(\boldsymbol{\theta}|\boldsymbol{\psi},\boldsymbol{Y})} \\ &\approx \left.\frac{p(\boldsymbol{Y}|\boldsymbol{\theta})(\boldsymbol{\theta}|\boldsymbol{\psi})p(\boldsymbol{\psi})}{\tilde{p}(\boldsymbol{\theta}|\boldsymbol{\psi},\boldsymbol{Y})}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\psi})} \\ &= \tilde{p}(\boldsymbol{\psi}|\boldsymbol{Y})\ . \end{aligned}$$

Here a Laplace approximation (LA) is used in the denominator for $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{\psi},\boldsymbol{Y})$. For univariate θ with an integral of the form $\int e^{g(\theta)}$, the LA takes the form $g(\theta) \sim N(\hat{\theta}(\boldsymbol{\psi}),\hat{\sigma}^2)$, where $\hat{\theta}(\boldsymbol{\psi})$ is the modal value of θ for specific values of the hyperparameters, $\boldsymbol{\psi}$ and $\hat{\sigma}^2 = \left\{\frac{d^2 \log g(\theta)}{d\theta^2}\right\}^{-1}$.

The mode of $\tilde{p}(\boldsymbol{\psi}|\boldsymbol{Y})$ can be found numerically by Newton-type algorithms. Around the mode, the distribution, $\log \tilde{p}(\boldsymbol{\psi}|\boldsymbol{Y})$, is evaluated over a grid of H points, $\boldsymbol{\psi}_h^*$, each with associated integration weight Δ_h. For each point on the grid, the marginal posterior, $\tilde{p}(\boldsymbol{\psi}_h^*|\boldsymbol{Y})$ is obtained from which approximations to the marginal distributions, $\tilde{p}(\boldsymbol{\psi}|\boldsymbol{Y})$, can be found using numerical integration.

For the individual model parameters, θ_j,

$$\begin{aligned} p(\theta_j|\boldsymbol{Y}) &= \frac{p((\theta_j,\boldsymbol{\theta}_{-j}),\boldsymbol{\psi}|\boldsymbol{Y})}{p(\boldsymbol{\theta}_{-j}|\theta_j,\boldsymbol{\psi},\boldsymbol{Y})} \\ &\propto \frac{p(\boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\psi})p(\boldsymbol{\psi})}{p(\boldsymbol{\theta}_{-j}|\theta_j,\boldsymbol{\psi},\boldsymbol{Y})} \\ &\approx \left.\frac{p(\boldsymbol{Y}|\boldsymbol{\theta})(\boldsymbol{\theta}|\boldsymbol{\psi})p(\boldsymbol{\psi})}{\tilde{p}(\boldsymbol{\theta}_{-j}|\theta_j,\boldsymbol{\psi},\boldsymbol{Y})}\right|_{\boldsymbol{\theta}_{-j}=\hat{\boldsymbol{\theta}}_{-j}(\theta_j,\boldsymbol{\psi})} \\ &= \tilde{p}(\theta_j|\boldsymbol{\psi},\boldsymbol{Y})\ . \end{aligned}$$

For an LGM, $(\boldsymbol{\theta}_{-j}|\theta_j,\boldsymbol{\psi},\boldsymbol{Y})$ will be approximately Gaussian. There are a number of ways of constructing the approximation in the denominator including a simple Gaussian approximation which will be computationally attractive but may be inaccurate. Alternatively, a LA would be highly accurate but computationally expensive. R-INLA uses a computationally efficient method, a simplified LA, that consists of performing a Taylor expansion around the LA of $\tilde{p}(\theta_j|\boldsymbol{\psi},\boldsymbol{Y})$, aiming to 'correct' the Gaussian approximation for location and skewness (Rue et al., 2009).

The marginal posteriors, $\tilde{p}(\boldsymbol{\psi}_h^*|\boldsymbol{Y})$, evaluated at each of the points $\boldsymbol{\psi}_h^*$, are used to obtain the conditional posteriors, $\tilde{p}(\theta_j|\boldsymbol{\psi}_h^*,\boldsymbol{Y})$, on a grid of values for θ_j. The marginal posteriors, $\tilde{p}(\theta_j|\boldsymbol{Y})$, are then found by numerical integration: $\tilde{p}(\theta_j|\boldsymbol{Y}) = \sum_h^H \tilde{p}(\theta_j|\boldsymbol{\psi}_h^*,\boldsymbol{Y})\tilde{p}(\boldsymbol{\psi}_h^*|\boldsymbol{Y})\Delta_h$, with the integration weights, Δ_h, being equal when the grid takes the form of a regular lattice.

The model presented in Section 3.1 was implemented using R-INLA (Rue et al., 2012) installed on the Balena high performance computing system (HPC) at the University of Bath (`www.bath.ac.uk/bucs/services/hpc/facilities/`). Fitting the model described in Section 3.1 to data from the 6003 monitors (and associated covariates) does not itself require the use of an HPC but the prediction on the entire grid (of 1.4 million cells) did present some computational challenges. When fitting the model, the prediction locations are treated as missing data and their posterior distributions

are approximated simultaneously with model fitting. INLA requires a copy of the model to be stored on a single node which, even with the high-memory compute nodes (32GBs per core) available with the Balena HPC, resulted in memory issues when attempting to perform estimation and prediction on the entire grid in a single step. Therefore, prediction was performed using subsets of the prediction grid, each containing groups of regions. Each subset, including the satellite estimates and other variables included in the model, was appended to the modelling dataset with both estimation and prediction performed for each combination. The resulting sets of predictions were combined to give a complete set of global predictions.

4 Results

A series of models based on the structure described in Section 3.1 were applied with the aim of assessing the predictive ability of potential explanatory factors. The choice of which variables were included in the final model was made based on their contribution to within-sample model fit and out-of-sample predictive ability.

Details of the variables included in five candidate models can be seen in Table 1. They include information on local network characteristics; indicator variables for whether the type of monitor was unspecified, X_1; whether the exact location is known, X_2, and whether $PM_{2.5}$ was estimated from PM_{10} (X_3); satellite-based estimates of $PM_{2.5}$ concentrations (X_4), estimates of $PM_{2.5}$ (X_5) from the TM5-FASST chemical transport model, dust (DUST; X_6) and the sum of sulphate, nitrate, ammonium and organic carbon (SNAOC; X_7) from atmospheric models; estimates of population (X_8) and a function of land–use and elevation (ED $\times$ DU; X_9). Except for the measurements themselves, all of these variables are spatially aligned to the resolution of the grid. Further details can be found in Section 2.

In the comparisons that follow, model (i) is the model used in GBD2013 (Brauer et al., 2015) and is a linear regression model with response equal to the average concentration from monitors within a grid cell and covariates X_1, X_2 and X_3 together with the average of the satellite-based estimates and those from the TM5-FASST chemical transport model for each cell, $(X_4 + X_5)/2$. Models (ii) to (v) are variants of the model presented in Section 3.1.

For evaluation, cross validation was performed using 25 combinations of training (80%) and validation (20%) datasets. Validation sets were obtained by taking a stratified random sample, using sampling probabilities based on the cross-tabulation of $PM_{2.5}$ categories (0-24.9, 25-49.9, 50-74.9, 75-99.9, 100+ μgm^{-3}) and super-regions, resulting in concentrations in each validation sets having the same distribution of $PM_{2.5}$ concentrations and super-regions as the overall set of sites. The following metrics were calculated for each training/evaluation set combination: for model fit, R^2 and the deviance information criteria (DIC, a measure of model fit for Bayesian models); and for predictive accuracy, root mean squared error (RMSE) and population weighted root mean squared error (PwRMSE). For the measures of predictive accuracy, each measurement (arising at a point location) is compared to the prediction for the grid cell that contains the ground monitor in question.

The results of fitting the five candidate models can be seen in Table 2, which shows R^2 and DIC for within sample model fit and RMSE and PwRMSE for out-of-sample predictive ability, and in Figure 6 which shows the PwRMSE for each model by super-region. It can be seen that using any of the hierarchical models based on the structure described in Section 3.1 provides an immediate improvement in all metrics when compared to the linear model, with a single global calibration function, used in GBD2013. For example, using model (ii) which contains satellite-based estimates, and population and local network characteristics, results in the overall R^2 improving from 0.54 to 0.90, DIC from 7828 to 1105 and reductions of 5.9 and 10.1 μgm^{-3} for RMSE and population weighted RMSE respectively.

Model	**(i)**		**(ii)**		**(iii)**		**(iv)**		**(v)**	
Variable	F	R	F	R	F	R	F	R	F	R
Intercept	✓		✓	✓	✓	✓	✓	✓	✓	✓
$X_1^{\dagger}$	✓		✓		✓		✓		✓	
$X_2^{\dagger}$	✓		✓		✓		✓		✓	
$X_3^{\dagger}$	✓		✓		✓		✓		✓	
X_4	✓		✓	✓	✓	✓	✓	✓	✓	✓
X_5	✓				✓	✓			✓	✓
X_6							✓		✓	
X_7							✓		✓	
$X_8^{\dagger\dagger}$			✓	✓	✓	✓	✓	✓	✓	✓
X_9							✓		✓	

† Together with interaction with X_4, X_5 where they are included within the model.
†† Country level random effects are assigned a conditional autoregressive prior.

Table 1: Variables includes in each of five candidate models: X_1, whether the type of monitor was unspecified; X_2, whether the exact location is known; X_3, whether $PM_{2.5}$ was estimated from PM_{10}; X_4, satellite-based and X_5 chemical transport model estimates of $PM_{2.5}$; X_6 and X_7, estimates of compositional concentrations of mineral dust and the sum of sulphate, nitrate, ammonium and organic carbon from atmospheric models; X_9, a function of elevation difference and land–use. Here, F and R denote the inclusion of fixed and random effects respectively for each variable.

This improvement can be seen in each of the super-regions (Figure 6), with the most marked improvements in areas where there is limited ground monitoring. Sensitivity analyses were performed to assess the effects of the given allocation of countries to regions. Repeating the analyses after switching a selection of countries that lay on regional borders to their adjacent region did not produce any discernible differences in the results.

Adding either estimates of $PM_{2.5}$ from the TM5-FASST chemical transport model; model (iii), or estimates of specific chemical components (SNAOC) and dust (DUST) from the GEOS-Chem chemical transport model together with information on differences in elevation between a ground monitor and its surrounding grid cell (ED $\times$ DU); model (iv), to this resulted in further improvements with model (iv) showing the most improvement. Although it resulted in a reduction in the DIC, adding the estimates of $PM_{2.5}$ from the TM5-FASST chemical transport model to model (iv) did not result in any substantial improvement in predictive ability. This may be in part due to the fact that the variables used in model (iv) are for 2014 whereas the estimates from the TM5-FASST model are from 2010. Considering the lack in improvement of predictive ability and the increased complexity and computational burden involved when incorporating an additional set of random effects, these estimates are not not included in the final model (model (iv)).

Predictions from the final model (model (iv)) can be seen in Figures 7a and 7b. The point estimates shown in Figure 7a give a summary of air quality for each grid cell and show clearly the spatial variation in global $PM_{2.5}$. For each grid cell, there is an underlying (posterior) probability distribution which incorporates information about the uncertainty of these estimates. There are a number of ways of presenting this uncertainty and Figure 7b shows one of these; half of the length of the 95% credible intervals (Denby et al., 2007). Here, higher uncertainty is associated with a combination of sparsity of monitoring data and higher concentrations, examples of which can seen in areas of North Africa and the Middle East.

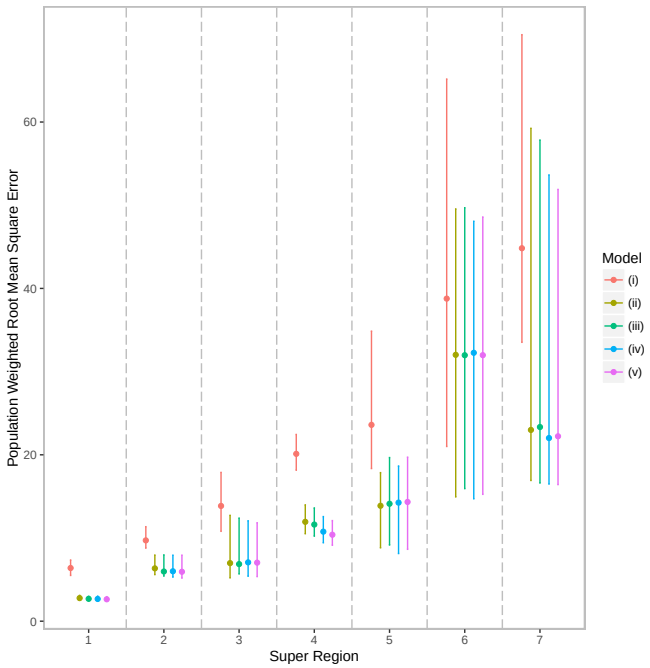


Figure 6: Summaries of predictive ability of the GBD2013 model (i) and four candidate models (ii-iv), for each of seven super–regions: 1, high income; 2, Central Europe, Eastern Europe, Central Asia; 3, Latin America and Caribbean; 4, Southeast Asia, East Asia and Oceania; 5, North Africa / Middle East; 6, Sub-Saharan Africa; 7, South Asia. For each model, population weighted root mean squared errors (μgm^{-3}) are given with dots denoting the median of the distribution from 25 training/evaluation sets and the vertical lines the range of values.

Model	R^2	DIC	RMSE†	PwRMSE†
(i)	0.54 (0.53, 0.54)	7828 (7685, 8657)	17.1 (16.5, 18.1)	23.1 (20.5, 29.3)
(ii)	0.90 (0.90, 0.91)	1105 (849, 1239)	11.2 (10.1, 12.9)	13.0 (11.5, 23.5)
(iii)	0.90 (0.90, 0.91)	986 (704, 1115)	11.1 (10.0, 13.3)	12.8 (11.2, 23.0)
(iv)	0.91 (0.90, 0.91)	877 (640, 1015)	10.7 (9.5, 12.3)	12.1 (10.7, 21.4)
(v)	0.91 (0.90, 0.92)	777 (508, 919)	10.7 (9.5, 12.5)	12.0 (10.7, 20.7)

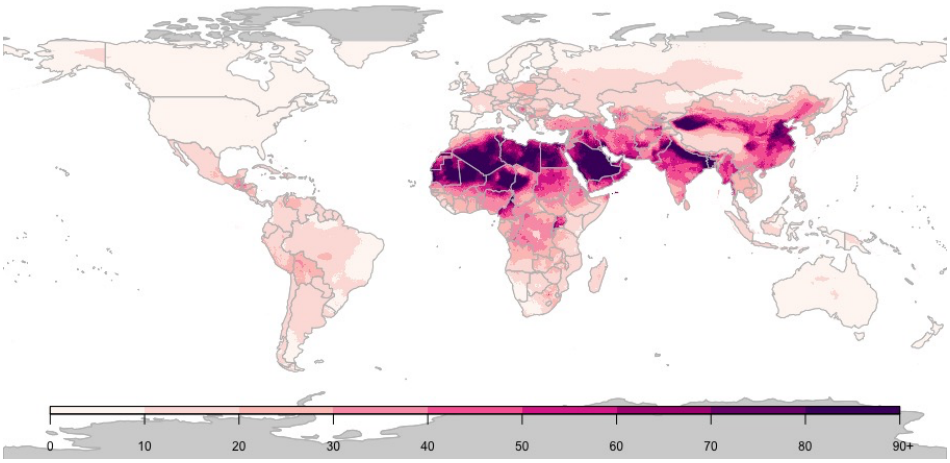
† μgm^{-3}

Table 2: Summary of results from fitting five candidate models described in Table 1. Results are presented for both in-sample model fit and out-of-sample predictive ability and are the median (minimum, maximum) values from 25 training-validation set combinations. For within sample model fit, R^2 and Deviance Information Criteria (DIC) are given and for out-of-sample predictive ability, root mean squared error (RMSE) and population weighted root mean squared error (PwRMSE).

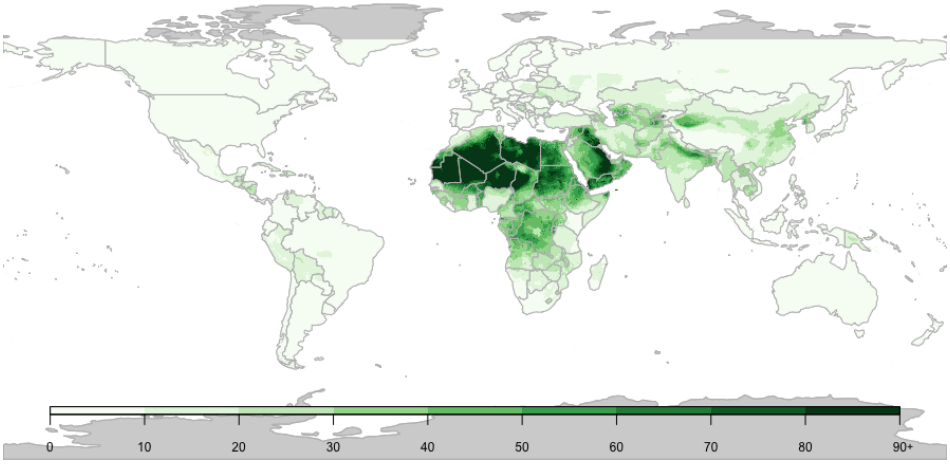
The distributions for each cell can also be used to examine the probabilities of exceeding particular thresholds. Figure 8 shows an example of this and contains predicted concentrations for China (Figure 8a) together with the probability for each cell that the value exceeds 35 μgm^{-3} (Figure 8b) and 75 μgm^{-3} (Figure 8c). High probabilities of exceeding the greater of the two thresholds are observed in the area around Beijing and in the Xinjiang province in the far west of the country. For the latter, a substantial component of the high (estimated) concentrations will be due to mineral dust from the large deserts in the region, as can be seen in Figure 3c. The distribution of estimated exposures shown in the map of median values (of the marginal posterior distributions) shown in Figure 8a can also be seen in Figure 9a which the profile of air pollution ($PM_{2.5}$) in this country contains three distinct components: (i) a land mass with low levels of air pollution; (ii) a much larger proportion of the total land mass with (comparatively) high levels; and (iii) a substantial area with very high levels. In terms of potential risks to health, it is high levels in areas of high population that will drive the disease burden. Figure 9b shows the distribution of estimated population level exposures, calculated by multiplying the estimate in each grid cell by its population. It can be seen that only a small proportion of the population reside in areas with the lowest concentrations with the vast majority of the population experiencing much higher levels of $PM_{2.5}$.

5 Discussion

In this paper we have developed a model to produce a comprehensive set of high-resolution estimates of exposures to fine particulate matter. The approach builds on that used for the GBD2013 project that calibrated ground measurements against estimates obtained from satellites and a chemical transport model using linear regression. This allowed data from the three sources to be utilised, but only provided an informal analysis of the uncertainty associated with the resulting estimates of exposure. There was also limited scope for considering changes in the calibration functions between geographical regions. As discussed in Brauer et al. (2015), the increase in the availability of ground measurements has increased the feasibility of allowing spatially varying calibration functions. This was performed using geographically weighted regression in Van Donkelaar et al. (2016), but here a hierarchical modelling approach is used in which country-specific calibration functions are used and information 'borrowed' from the surrounding region and super–region where local monitoring data is inadequate for stable estimation of the coefficients in the calibration models. This is achieved using sets of random effects, for countries within regions within super–regions, reflecting a nested geographical hierarchy. The models are fitted within a Bayesian hierarchical framework which produces full posterior distributions for estimated levels of $PM_{2.5}$ for each grid cell rather than just point estimates. Summaries of these posterior distributions can be used to give point estimates, e.g, medians and medians, together with measures of uncertainty, e.g. 95% credible intervals. They can also be used to estimate exceedance probabilities,

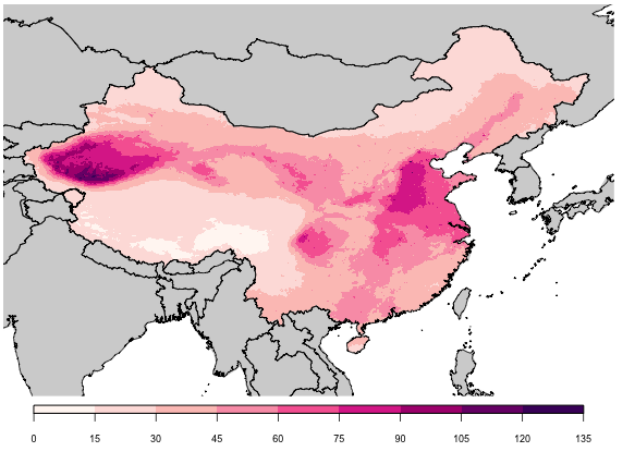


(a) Medians of posterior distributions.

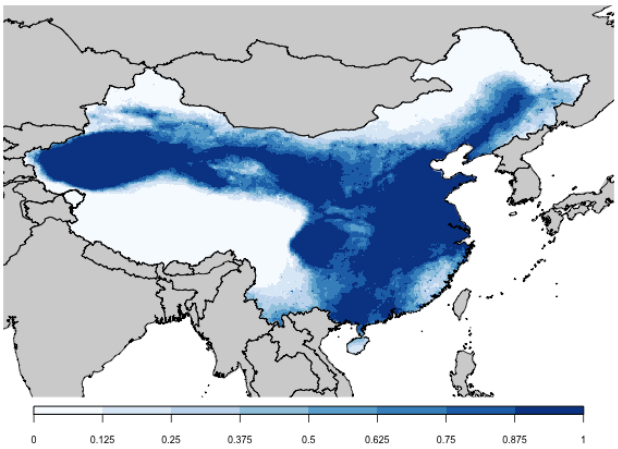


(b) Half the width of 95% posterior credible intervals.

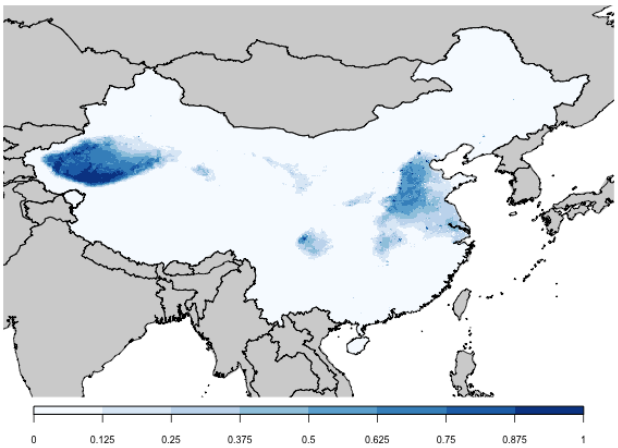
Figure 7: Estimates of annual averages of $PM_{2.5}$ (μgm^{-3}) for 2014 together with associated uncertainty for each grid cell ($0.1^\circ \times 0.1^\circ$ resolution) using a Bayesian hierarchical model (see text for details).



(a) Medians of posterior distributions.

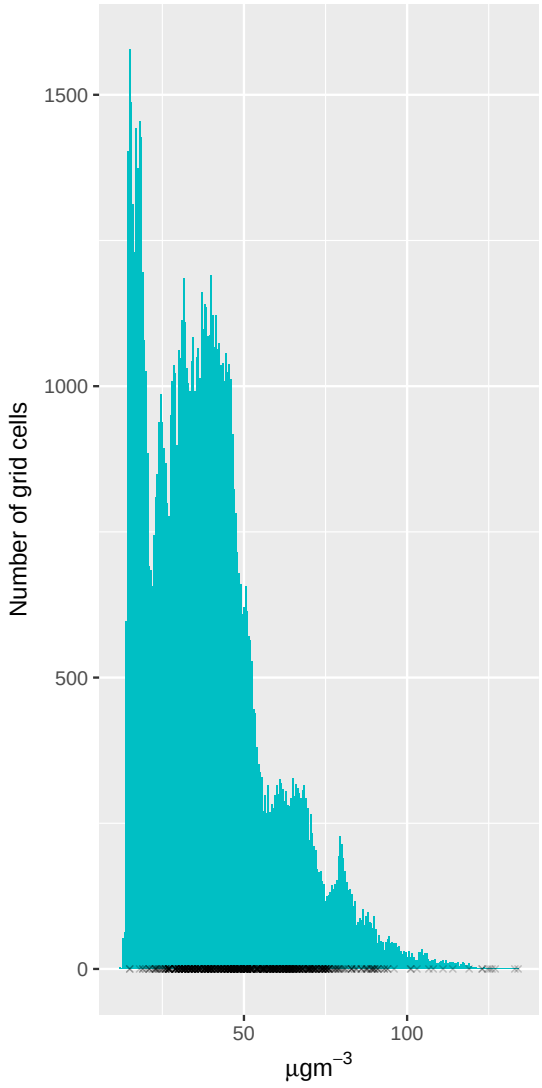


(b) Probability of exceeding 35 μgm^{-3}.

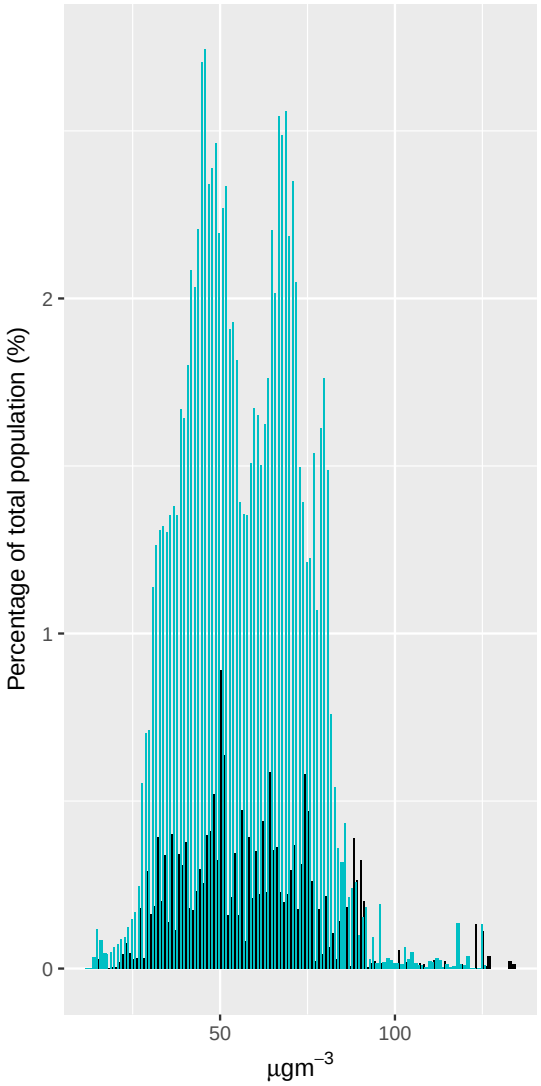


(c) Probability of exceeding 75 μgm^{-3}.

Figure 8: Estimates of annual mean $PM_{2.5}$ concentrations (μgm^{-3}) for 2014 together with exceedance probabilities using a Bayesian hierarchical model (see text for details) for each grid cell ($0.1^\circ \times 0.1^\circ$ resolution) in China.



(a) Estimated annual average concentrations of $PM_{2.5}$ by grid cell ($0.1^o \times 0.1^o$ resolution). Black crosses denote the annual averages recorded at ground monitors with the level of transparency denoting the density of monitors at each concentration.



(b) Estimated population level exposures (blue bars) and, for cells containing at least one monitoring station, population weighted measurements from ground monitors (black bars).

Figure 9: Distributions of annual mean concentrations and population level exposures for $PM_{2.5}$ (μgm^{-3}) in China.

e.g. the probability of exceeding air quality guidelines. Based on posterior estimates (medians for each grid cell), it is estimated that 92% of the world's population reside in grid cells for which the annual average is greater than the WHO guideline of 10 μgm^{-3}, which is greater than the 87% reported in Brauer et al. (2015) for 2013.

In addition to the hierarchical approach to modelling used here, the increased availability of ground monitoring data has been utilised in the analysis. Ground measurements were available from 6003 locations (compared with 4073 for GBD2013) and, in addition, estimates of specific components of air pollution, including mineral dust and the sum of sulphate, nitrate, ammonium and organic carbon, were available from atmospheric models. A series of candidate models, containing different sets of variables and structures for the random effects, were considered with the final choice of model being made on predictive ability. This was assessed by cross-validation in which models were fitted to 25 training datasets (each containing 80% the overall data with stratified sampling to ensure samples were representative in terms of the distribution of concentrations within each super–region) and predictions compared to measurements within the corresponding validation set. The final model contained information on local network characteristics, including whether $PM_{2.5}$ was measured or values converted from PM_{10}, and whether the exact site type and location were known, together with satellite-based estimates, estimates of specific components from the GEOS-Chem chemical transport model, land use and elevation, and population. The final model includes country-level (within region, within super–region) random effects for satellite-based estimates and neighbouring country level random effects for population, with interactions between the fixed effects for variables and those reflecting local network characteristics. Notably, the estimates of $PM_{2.5}$ from the TM5-FASST model used in GBD2013 were not found to improve the predictive ability and they were not included in the final model. In preference, estimates of specific components of pollution and the interaction between altitude and land–use from Van Donkelaar et al. (2016) were found to provide marked improvements in predictively ability and are included in the model.

The model presented here has been shown to offer improved estimates of $PM_{2.5}$ but there is certainly room for improvement, especially in areas such as Sub-Saharan Africa and South Asia. One of the potential uses of the outputs from the model, i.e. the information on areas with high predicted exposures and high uncertainty shown in Figures 7 and 8, would be to guide where future monitoring efforts might be focused. It may also be possible to utilise other sources of information related to air quality in addition to those considered here, such as road networks and other land–use variables.

In the current implementation, a single annual average of ground measurements is used for each monitoring location. For 2014, 46% of the measurements from the WHO cities database come from that year with the remainder coming from the closest year for which data were available. This results in 82% of the measurements coming from 2014 or 2013 with the majority of the remainder coming from the period 2010-2012. As monitoring networks develop, in some areas there will be the possibility of multiple measurements at specific locations over time and future developments of the model might include a temporal component that would acknowledge the temporal aspect of the data, possibly with lower weight given to less recent measurements. At present, one approach to reducing the issues that might arise when comparing measurements from locations in close proximity where there are differences over time, would be to only use data from the most recent years. However, such data is often not available in precisely the regions where ground measurements are most needed to produce accurate calibration functions.

In the calibration approach used here there is an implicit assumption the covariates are error free, an assumption that may be untenable in practice. When integrating data from many different sources, each source will have its own error structures and spatially varying biases. For example, the estimates of $PM_{2.5}$ from satellite retrievals and the estimates of specific components from the chemical transport are all the result of modelling and, as such, will be subject to uncertainties and biases arising from

errors in inputs and possible model misspecification. Therefore, a Bayesian melding approach may be more suitable in this setting, in which each source of information is assumed to be related to an underlying 'true' level of pollution (at any location) with additive and multiplicative bias terms. In addition, Bayesian melding allows provides a coherent framework in which data from different sources at different levels of aggregation can be integrated, and allows for prediction at any required level of aggregation with associated estimates of uncertainty.

However, Bayesian melding is complex to implement and can be very computationally demanding, particularly using Markov chain Monte Carlo (MCMC), due to the requirement to perform a stochastic integral of the underlying continuous process to the resolution of the grid cells, for each grid cell. In contrast, one of the major advantages of downscaling is the computational saving that is made by only considering grid cells containing measurement locations within the estimation, after which prediction at unknown locations is relatively straightforward (Chang, 2016). In its current incarnation, using MCMC, Bayesian melding is computationally infeasible for large-scale problems of this type. Future research will involve developing computationally efficient methods for performing Bayesian melding, using approximate Bayesian inference.

In summary, this work presents an important step forward in large-scale data integration in this setting, allowing information on air quality to be drawn from a wide variety of sources, each potentially measured at different resolutions, with different error structures and with different levels of uncertainty. Ultimately, this will lead to more accurate estimates of air quality together with measures of uncertainty that acknowledge the uncertainty associated with the individual data sources. This information can also be incorporated within a health effects model leading to improved characterisation of uncertainty when estimating disease burden. This in turn will lead to increased understanding of the effects of air pollution on health and the potential effects of mitigation strategies.

6 Acknowledgments

The model was developed by the a multi-disciplinary group of experts established as part of the recommendations from the first meeting of the WHO Global Platform for Air Quality, Geneva, January 2014. The resulting *Data Integration Task Force* consists of authors 1, 4-9 and 12-16 of this paper together with members of the WHO (authors 10,11 and 17). The views expressed in this article are those of the authors and they do not necessarily represent the views, decisions or policies of the institutions with which they are affiliated. The model presented and reviewed at the second meeting of the Global Platform for Air Quality, Geneva, August 2015. Matthew Lloyd Thomas is supported by a scholarship from the EPSRC Centre for Doctoral Training in Statistical Applied Mathematics at Bath (SAMBa), under the project EP/L015684/1. Amelia Green was supported for this work by WHO contracts APW 201255146 and 201255393.

References

Berrocal, V. J., Gelfand, A. E. and Holland, D. M. (2010) A Spatio-Temporal Downscaler for Output from Numerical Models. *Journal of Agricultural, Biological, and Environmental Statistics*, **15**, 176–197.

Besag, J. (1974) Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Statistical Methodological)*, 192–236.

Bey, I., Jacob, D. J., Yantosca, R. M., Logan, J. A., Field, B. D., Fiore, A. M., Li, Q., Liu, H. Y., Mickley, L. J. and Schultz, M. G. (2001) Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation. *Journal of Geophysical Research: Atmospheres*, **106**, 23073–23095.

Brauer, M., Amann, M., Burnett, R. T., Cohen, A., Dentener, F., Ezzati, M., Henderson, S. B., Krzyzanowski, M., Martin, R. V., Van Dingenen, R., van Donkelaar, A. and Thurston, G. D. (2012) Exposure Assessment for Estimation of the Global Burden of Disease Attributable to Outdoor Air Pollution. *Environmental Science and Technology*, **46**, 652–660.

Brauer, M., Freedman, G., Frostad, J., Van Donkelaar, A., Martin, R. V., Dentener, F., van Dingenen, R., Estep, K., Amini, H., Apte, J. S., Balakrishnan, K., Barregard, L., Broday, D., Feigin, V., Ghosh, S., Hopke, P. K., Knibbs, L. D., Kokubo, Y., Liu, Y., Ma, S., Morawska, L., Texcalac Sangrador, J. L., Shaddick, G., Anderson, H. R., Vos, T., Forouzanfar, M. H., Burnett, R. T. and Cohen, A. (2015) Ambient Air Pollution Exposure Estimation for the Global Burden of Disease 2013. *Environmental Science and Technology*, **50**, 79–88.

Brook, R., Rajagopalan, S., Pope, C. r., Brook, J., Bhatnagar, A., Diez-Roux, A., F, H., Hong, Y., Luepker, R., Mittleman, M., Peters, A., Siscovick, D., Smith, S. J., L, W. and Kaufman, J. (2010) Particulate matter air pollution and cardiovascular disease an update to the scientific statement from the American Heart Association. *Circulation*, **121**, 2331–2378.

Chang, H. (2016) Data Assimilation for Environmental Pollution Fields. In *Handbook of Spatial Epidemiology* (eds. A. B. Lawson, S. Banerjee, R. P. Haining and M. D. Ugarte), chap. 16, 289–302. Boca Raton: CRC Press.

Denby, B., Costa, A., Monteiro, A., Dudek, A. and Erik, S. (2007) Uncertainty Mapping for Air Quality Modelling and Data Assimilation. In *Proceedings of the 11th International Conference on Harmonisation within Atmospheric Dispersion Purposes, Cambridge, UK*.

Forouzanfar, M. H., Alexander, L., Anderson, H. R., Bachman, V. F., Biryukov, S., Brauer, M., Burnett, R., Casey, D., Coates, M. M., Cohen, A. and *et al.* (2015) Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*, **386**, 2287–2323.

Friedl, M. A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A. and Huang, X. (2010) MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sensing of Environment*, **114**, 168–182.

Fuentes, M. and Raftery, A. E. (2005) Model Evaluation and Spatial Interpolation by Bayesian Combination of Observations with Outputs from Numerical Models. *Biometrics*, **61**, 36–45.

GPW (2016) *Gridded Population of the World, Version 4 (GPWv4): Administrative Unit Center Points with Population Estimates.* Center for International Earth Science Information Network - CIESIN - Columbia University.

Guillas, S., Tiao, G., Wuebbles, D. and Zubrow, A. (2006) Statistical diagnostic and correction of a chemistry-transport model for the prediction of total column ozone. *Atmospheric Chemistry and Physics*, **6**, 525–537.

Hoek, G., Krishnan, R. M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B. and Kaufman, J. D. (2013) Long-term air pollution exposure and cardio-respiratory mortality: a review. *Environmental Health*, **12**, 1.

Huijnen, V., Williams, J., van Weele, M., van Noije, T., Krol, M., Dentener, F., Segers, A., Houweling, S., Peters, W., Laat, J. d., Boersma, F., Bergamaschi, P., van Velthoven, P., Le Sager, P., Eskes, H., Alkemade, F., Scheele, R., Nédélec, P. and Pätz, H. W. (2010) The global chemistry transport model TM5: Description and evaluation of the tropospheric chemistry version 3.0. *Geoscientific Model Development*, **3**, 445–473.

Van de Kassteele, J., Koelemeijer, R., Dekkers, A., Schaap, M., Homan, C. and Stein, A. (2006) Statistical mapping of PM_{10} concentrations over Western Europe using secondary information from dispersion modeling and MODIS satellite observations. *Stochastic Environmental Research and Risk Assessment*, **21**, 183–194.

Kloog, I., Chudnovsky, A. A., Just, A. C., Nordio, F., Koutrakis, P., Coull, B. A., Lyapustin, A., Wang, Y. and Schwartz, J. (2014) A new hybrid spatio-temporal model for estimating daily multi-year $PM_{2.5}$ concentrations across northeastern USA using high resolution aerosol optical depth data. *Atmospheric Environment*, **95**, 581–590.

Loomis, D., Grosse, Y., Lauby-Secretan, B., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Baan, R., Mattock, H. and Straif, K. (2013) International Agency for Research on Cancer Monograph Working Group, IARC. The carcinogenicity of outdoor air pollution. *Lancet Oncology*, **14**, 1262–1263.

McMillan, N. J., Holland, D. M., Morara, M. and Feng, J. (2010) Combining numerical model output and particulate data using Bayesian space–time modeling. *Environmetrics*, **21**, 48–65.

Newby, D., Mannucci, P., Tell, G., Baccarelli, A., Brook, R., Donaldson, K., Forastiere, F., Franchini, M., Franco, O., Graham, I., Hoek, G., Hoffmann, B., Hoylaerts, M., Künzli, N., Mills, N., Pekkanen, J., Peters, A., Piepoli, M., Rajagopalan, S. and Storey, R. (2014) ESC Working Group on Thrombosis, European Association for Cardiovascular Prevention and Rehabilitation; ESC Heart Failure Association. Expert position paper on air pollution and cardiovascular disease. *European Heart Journal*, ehu458.

Poole, D. and Raftery, A. E. (2000) Inference for Deterministic Simulation Models: the Bayesian Melding Approach. *Journal of the American Statistical Association*, **95**, 1244–1255.

Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*. CRC Press.

Rue, H., Martino, S. and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**, 319–392.

Rue, H., Martino, S. and Lindgren, F. (2012) The R-INLA Project. *R-INLA http://www. r-inla. org*.

Sava, F. and Carlsten, C. (2012) Respiratory Health Effects of Ambient Air Pollution: An Update. *Clinics in Chest Medicine*, **33**, 759–769.

Van Dingenen, R., Leitao, J. and Dentener, F. (2014) A multi-metric global source-receptor model for integrated impact assessment of climate and air quality policy scenarios. In *EGU General Assembly Conference Abstracts*, vol. 16, 13949.

Van Donkelaar, A., Martin, R. V., Brauer, M., Hsu, N. C., Kahn, R. A., Levy, R. C., Lyapustin, A., Sayer, A. M. and Winker, D. M. (2016) Global Estimates of Fine Particulate Matter using a Combined Geophysical-Statistical Method with Information from Satellites, Models, and Monitors. *Environmental Science and Technology*, **50**, 3762–3772.

WHO (2013) *Review of evidence on health aspects of air pollution – REVIHAAP Project*. World Health Organization.

— (2016a) *Ambient Air Pollution: A global assessment of exposure and burden of disease.* World Health Organization, Geneva.

— (2016b) *WHO Global Urban Ambient Air Pollution Database (update 2016).* World Health Organization, Geneva.

Zidek, J. V., Le, N. D. and Liu, Z. (2012) Combining data and simulated data for space–time fields: application to ozone. *Environmental and Ecological Statistics*, **19**, 37–56.