ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2018)
67, *Part* 1, *pp.* 231–253

Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution

Gavin Shaddick,
University of Exeter and University of Bath, UK

Matthew L. Thomas and Amelia Green,
University of Bath, UK

Michael Brauer,
University of British Columbia, Vancouver, Canada

Aaron van Donkelaar,
Dalhousie University, Halifax, Canada

Rick Burnett,
Health Canada, Ottawa, Canada

Howard H. Chang,
Emory University, Atlanta, Canada

Aaron Cohen,
Health Effects Institute, Boston, USA

Rita Van Dingenen,
European Commission, Ispra, Italy

Carlos Dora and Sophie Gumy,
World Health Organization, Geneva, Switzerland

Yang Liu,
Emory University, Atlanta, USA

Randall Martin,
Dalhousie University, Halifax, Canada

Address for correspondence: Gavin Shaddick, Department of Mathematics, University of Exeter, Streatham Campus, Exeter, EX4 4QT, UK.

0035–9254/18/67231

Lance A. Waller,
Emory University, Atlanta, USA

Jason West,
University of North Carolina, Chapel Hill, USA

James V. Zidek
University of British Columbia, Vancouver, Canada

and Annette Prüss-Ustün
World Health Organization, Geneva, Switzerland

[Received September 2016. Revised April 2017]

Summary. 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, regions remain in which monitoring is limited. The data integration model for air quality 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.

Keywords: Air pollution; Bayesian hierarchical modelling; Data fusion; Environmental health effects; Global burden of disease; Integrated nested Laplace approximations; Spatial modelling

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*., 2015; Sava and Carlsten, 2012; World Health Organization, 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 $PM_{2.5}$ (World Health Organization, 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 (an annual mean of 10 μg m^{-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 (World Health Organization, 2016b) are shown in Fig. 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.

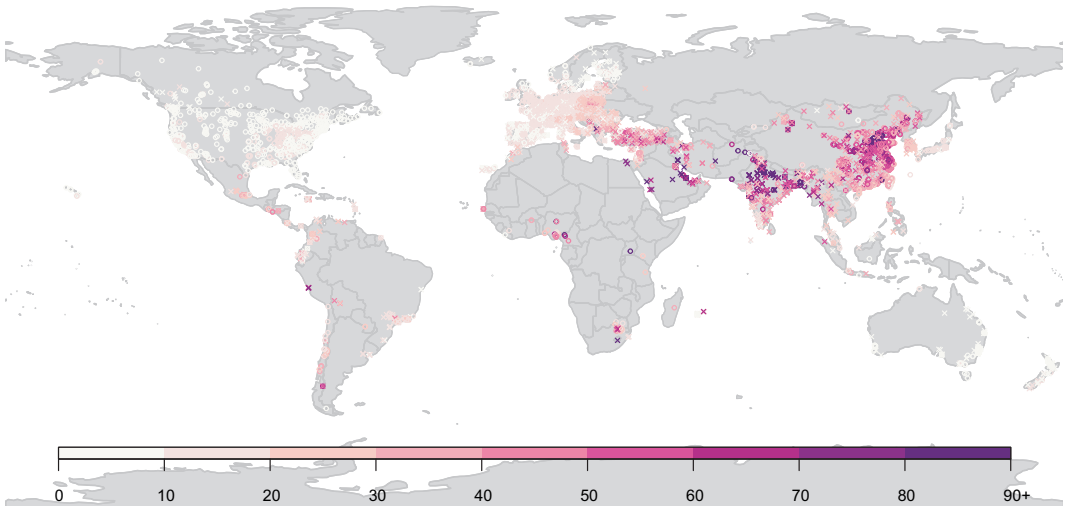


Fig. 1. Locations of ground monitors measuring $PM_{2.5}$ (○) and PM_{10} (×): colours denote the annual average concentrations ($\mu g\ m^{-3}$) of $PM_{2.5}$ (or $PM_{2.5}$ converted from PM_{10}); the data are from 2014 (46%), 2013 (36%), 2012 (9%) and 2006–2011 and 2015 (9%)

For this reason, there is a need to use information from other sources to obtain estimates of exposures for all areas of the world. In 2013, the 'Global burden of disease' (GBD) study, as described in Forouzanfar *et al*. (2015), used a regression calibration approach to utilize 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 11 km × 11 km at the equator) that were then matched to population estimates to estimate disease burden. In the 2013 GBD, a fused estimate of $PM_{2.5}$, calculated as the average of estimates from satellites and chemical transport models, was calibrated against ground measurements by 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, enabling a comprehensive set of global estimates of $PM_{2.5}$ to be produced.

This allowed data from the three sources to be utilized, but the use of a single, global, calibration function resulted in underestimation in various areas (Brauer *et al*., 2015). In reality, the relationships between ground measurements and estimates from other sources will vary spatially because of 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 by using geographically weighted regression to allow calibration (between the measurements and estimates) equations to vary spatially and to utilize additional information related to land use and the chemical composition of particulate matter. However, both the original linear regression and geographically weighted regression approaches provide only an informal analysis of the uncertainty that is 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 microscale 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, which are often referred to as *downscaling models* (Chang, 2016). In the purely spatial model that was 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 spatiotemporal modelling of air pollution with examples including Guillas *et al*. (2006), who used the University of Illinois at Urbana—Champaign two-dimensional chemical transport model of the global atmosphere, Van de Kassteele *et al*. (2006), who modelled PM_{10}-concentrations over Western Europe by using information from both satellite observations and a chemical transport model, McMillan *et al*. (2010) who modelled $PM_{2.5}$ in the North Eastern USA by using estimates from the community multiscale air quality numerical model, Kloog *et al*. (2014) who modelled $PM_{2.5}$ in the North Eastern USA by using satellite-based aerosol optical depth AOD and Berrocal *et al*. (2010) and Zidek *et al*. (2012) who modelled ozone in the eastern USA (eastern and central in the case of Zidek *et al*. (2012)) by using estimates from the community multiscale air quality model and a variant of the 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 sulphur dioxide in the Eastern USA, 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 estimates exposures, together with associated measures of uncertainty, at high geographical resolution by utilizing information from multiple sources and addresses many of the issues that were 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 data integration model for air quality statistical model that is used to integrate data from these different sources and methods for inference when performing Bayesian analysis with large data sets. In Section 4 the results of applying the data integration model for air quality model are presented, including examples of global and country-specific estimates of exposure to $PM_{2.5}$ together with details of the methods that are 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 that are used here can be allocated to one of three groups:

(a) ground monitoring data;
(b) estimates of $PM_{2.5}$ from remote sensing satellites and chemical transport models;
(c) other sources including population, land use and topography.

Ground monitoring is available at a distinct number of locations, whereas the last two groups provide near complete global coverage (and have previously been shown to have strong associations with global $PM_{2.5}$-concentrations; see below for details). Utilizing 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 (World Health Organization, 2016b) but, rather than using the city averages that are reported in that database, monitor-specific measurements are used. The result was measurements of PM_{10}- and $PM_{2.5}$-concentrations from 6003 ground monitors. The locations and annual average concentrations for these monitors can be seen in Fig. 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-co-ordinates 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 by 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 was 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:

$$PM_{2.5} = \eta \, AOD.$$

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 global chemical transport model GEOS-Chem (Bey *et al.*, 2001).

The estimates that are 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 11 km $\times$ 11 km 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 Fig. 2.

2.3. Chemical transport model simulations

Numerically simulated estimates of $PM_{2.5}$ were obtained from atmospheric chemical transport models. A variety of such models are 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 three-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

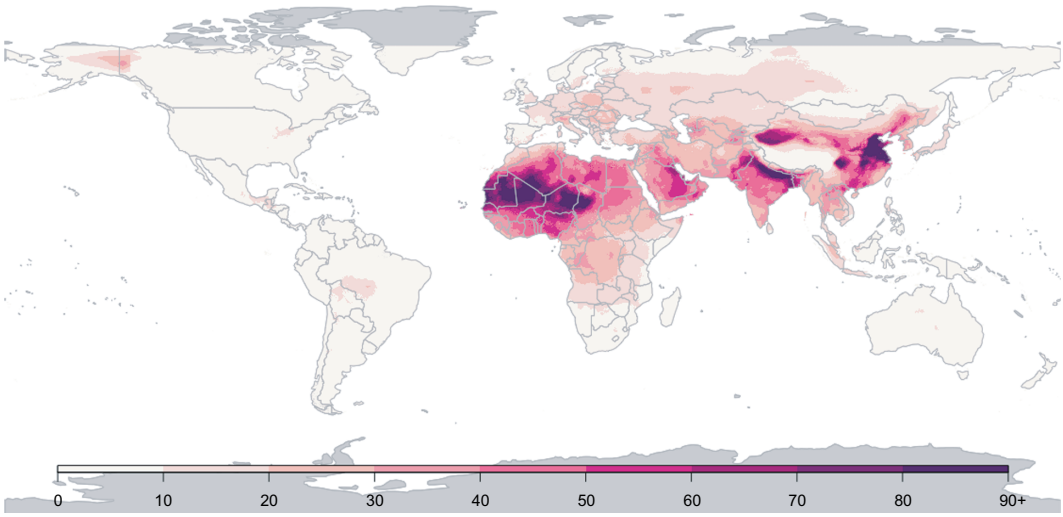


Fig. 2. Satellite-based estimates of $PM_{2.5}$ (μg m^{-3}) for 2014, by grid cell (0.1° × 0.1° resolution)

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 TM5-FASST were available for 2010, as described in Brauer *et al*. (2015). A map of these estimates can be seen in Fig. 3(a).

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 Figs 3(b) and 3(c) respectively.

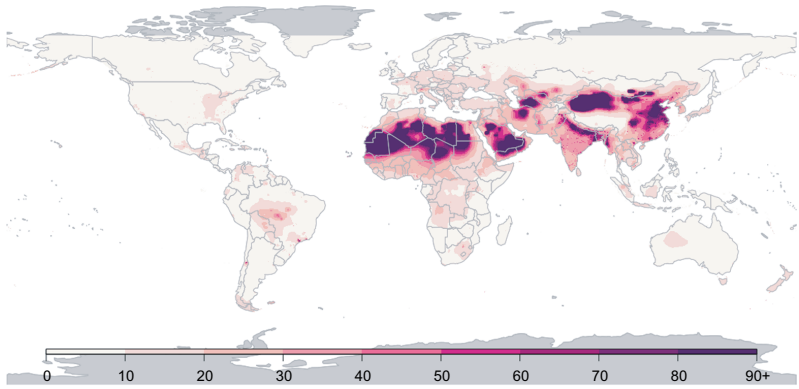
2.4. Population data

A comprehensive set of population data on a high resolution grid was obtained from the 'Gridded population of the world' (GPW) database (Center for International Earth Science Information Network, 2016). These data are provided at $0.0417^\circ \times 0.0417^\circ$ resolution. Aggregation to each $0.1^\circ \times 0.1^\circ$ grid cell was performed as detailed in Brauer *et al*. (2015). Version 4 of the database provides population estimates for 2000, 2005, 2010, 2015 and 2020. Following the methodology that was 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 Fig. 4.

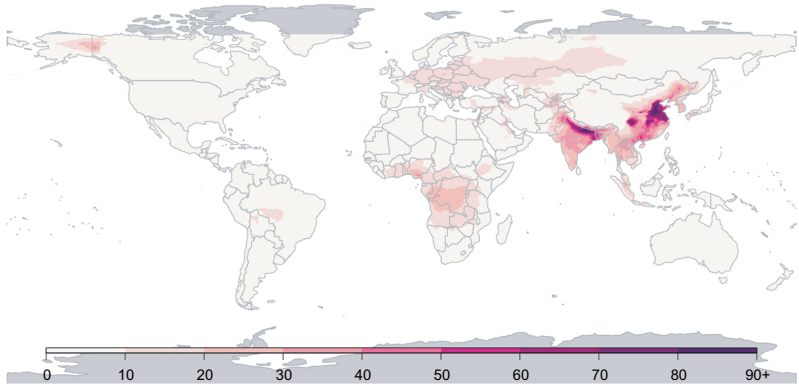
2.5. Land use

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:

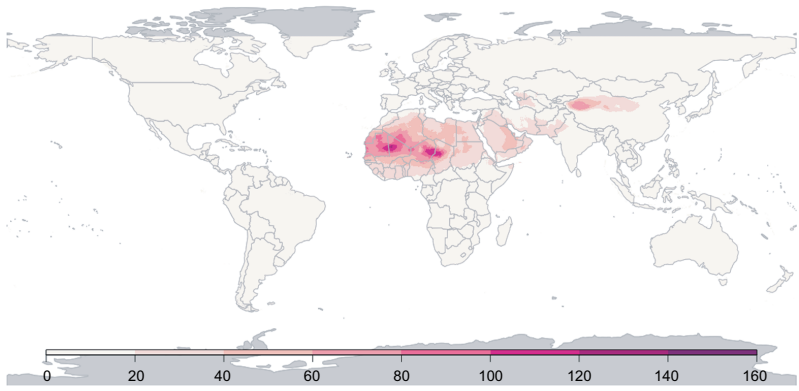
(a) the difference between the elevation (of the ground monitor) and that of the surrounding grid cell (ED) as defined by the GEOS-Chem chemical transport model;



(a)



(b)



(c)

Fig. 3. Estimates from chemical transport models, by grid cell (0.1° × 0.1° resolution): (a) estimates of $PM_{2.5}$ (μg m^{-3}) for 2010 from the TM5 chemical transport model used in the 2013 GBD; (b) estimates of the sum of sulphate, nitrate, ammonium and organic carbon (μg m^{-3}) for 2014 from the GEOS-Chem chemical transport model; (c) estimates of the compositional concentrations of mineral dust (μg m^{-3}) for 2014 from the GEOS-Chem chemical transport model

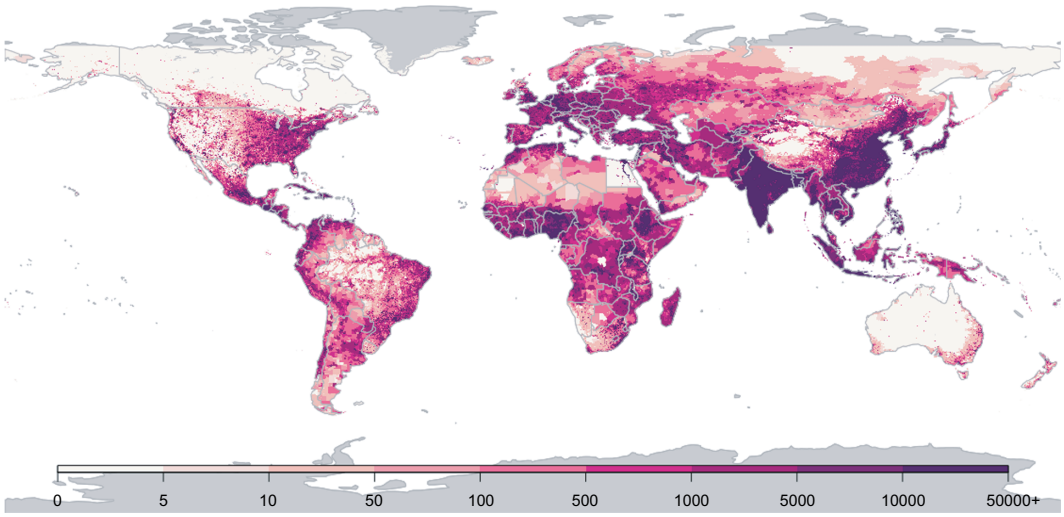


Fig. 4. Population estimates for 2014 from the GPW version 4 database, by grid cell (0.1° × 0.1° resolution)

(b) the distance to the nearest urban land surface (DU) based on the 'Moderate resolution imaging spectroradiometer' MODIS land cover descriptions (Friedl *et al.*, 2010). The combination of these measures, 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 $PM_{2.5}$-concentrations 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 $L = l_1, \ldots, l_{N_L}$.

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, $\epsilon_s \sim N(0, \sigma_\epsilon^2)$ 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:

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

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).

Although monitoring data are increasingly available, there are issues that may mean that 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 Fig. 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 that are used here are based on country level geography rather than continuous spatial processes.

The structure of the random effects that are used here exploits a geographical nested hierarchy: each of the 187 countries that were considered are allocated to one of 21 regions and, further, to one of seven super-regions. Each region must contain at least two countries and is broadly based on geographic regions or subcontinents 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 Fig. 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 $Y_{sli\,jk}$. As described in Section 3, there is a nested hierarchical structure with $s = 1, \ldots, N_{li\,jk}$ sites within grid cell l, $l = 1, \ldots, N_{i\,jk}$, grid cells within country i, $i = 1, \ldots, N_{jk}$, countries within region j, $j = 1, \ldots, N_k$, and regions within super-region k, $k = 1, \ldots, N$. 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

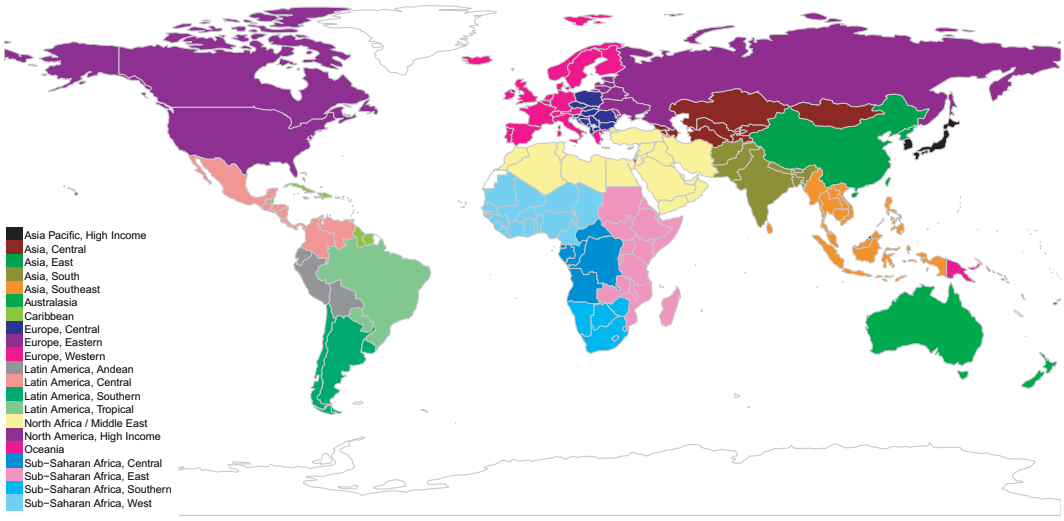
$$\log(Y_{sli\,jk}) = \tilde{\beta}_{0,li\,jk} + \sum_{q \in Q} \tilde{\beta}_{q,i\,jk} X_{q,li\,jk} + \sum_{p_1 \in P_1} \beta_{p_1} X_{p_1,li\,jk} + \sum_{p_2 \in P_2} \beta_{p_2} X_{p_2,sli\,jk} + \epsilon_{sli\,jk}, \qquad (2)$$

where $\epsilon_{sli\,jk} \sim N(0, \sigma_\epsilon^2)$ 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, $P = (P_1, P_2)$.

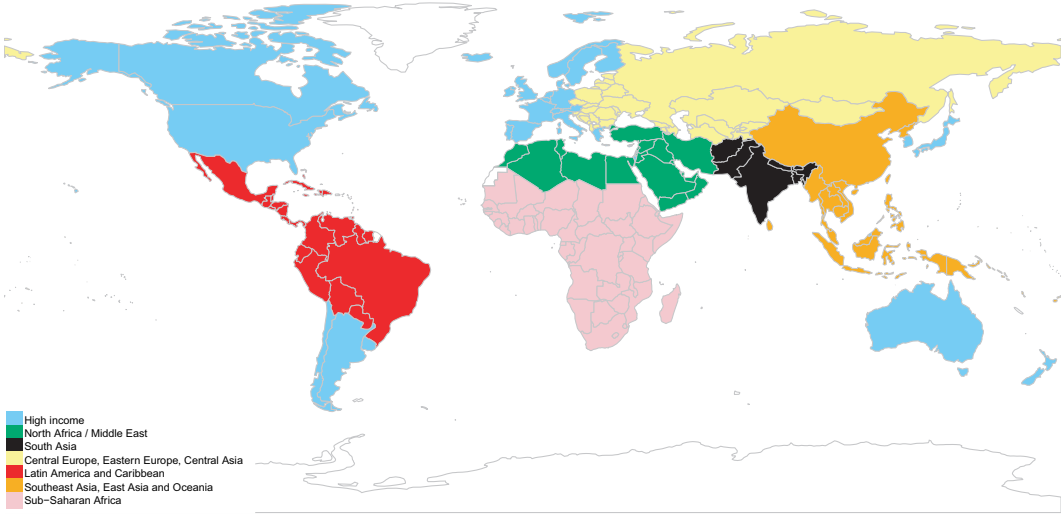
3.1.1. Structure of the random effects

Here, the random-effect terms $\tilde{\beta}_{0,li\,jk}$ and $\tilde{\beta}_{q,i\,jk}$ 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:

$$\tilde{\beta}_{0,li\,jk} = \beta_0 + \beta^G_{0,li\,jk} + \beta^C_{0,i\,jk} + \beta^R_{0,\,jk} + \beta^{\mathrm{SR}}_{0,k}$$
$$\tilde{\beta}_{q,i\,jk} = \beta_q + \beta^C_{q,i\,jk} + \beta^R_{q,\,jk} + \beta^{\mathrm{SR}}_{q,k}.$$



(a)



(b)

Fig. 5. Schematic diagrams showing the nested geographical structure of countries (a) within regions and (b) 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:

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

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 the

coefficient for the super-region and variance representing the between-region (within-super-region) variability:

$$\beta^R_{jk} \sim N(\beta^{\mathrm{SR}}_k, \sigma^2_{R,k}),$$

where $j = 1, \ldots, N_k$, the number of regions in super-region j. Each region will contain a number of countries. Let β^C_{ijk} 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:

$$\beta^C_{ijk} \sim N(\beta^R_{jk}, \sigma^2_{C,jk}), \tag{3}$$

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^G_{lijk} 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, subcontinents, 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 Fig. 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 the countries in the surrounding administrative region. This could be achieved by using an intrinsic conditionally auto-regressive model (Besag, 1974) in place of equation (3):

$$\beta^C_i | \beta^C_{i'}, \qquad i' \in \partial_i \sim N\left(\bar{\beta}^C_i, \frac{\psi^2}{N_{\partial_i}}\right),$$

where ∂_i is the set of neighbours of country i, N_{∂_i} is the number of neighbours and $\bar{\beta}^C_i$ 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 $\mathrm{Ga}(a, b)$ are assigned to the logarithm of the precisions, i.e. $\log(\sigma^{-2}_{G,i})$, $\log(\sigma^{-2}_{C,j})$, $\log(\sigma^{-2}_{R,j})$, $\log(\sigma^{-2}_{\mathrm{SR}})$ and $\log(\psi^{-2})$, with $a = 1$ and $b = 0.000\,05$.

3.2. Inference

The model that was presented in Section 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 by using approximate Bayesian inference using integrated nested Laplace approximations (INLAs) 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 INLAs (Section 3.2.2) with additional details linking to the model that was described in Section 3.1.

3.2.1. Latent Gaussian models

The model that is 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 that are associated with covariates X; the fixed effects. The set of functions $f_1(\cdot), \ldots, f_Q(\cdot)$ represents 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 that are 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 β_0, $\beta^G_{0,lijk}$, $\beta^C_{0,ijk}$, $\beta^R_{0,jk}$, $\beta^{\mathrm{SR}}_{0,k}$, β_q, $\beta^C_{q,ijk}$, $\beta^R_{q,jk}$ and $\beta^{\mathrm{SR}}_{q,k}$, with the set of hyperparameters associated with $\boldsymbol{\theta}$ being $\boldsymbol{\psi}_2 = (\sigma^2_{G,i}, \sigma^2_{C,j}, \sigma^2_{R,j}, \sigma^2_{\mathrm{SR}})$. 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 $\boldsymbol{\theta}$, $\boldsymbol{\theta}|\boldsymbol{\psi} \sim \mathrm{MVN}\{\mathbf{0}|\Sigma(\boldsymbol{\psi}_2)\}$, will result in an LGM. The computation that is 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, 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 Gaussian Markov random field means that typically the inverse of the covariance matrix, $Q = \Sigma^{-1}$, will be sparse (i.e. more 0-entries) because of the conditional independence between sets of parameters in which $\theta_l \perp\!\!\!\perp \theta_m | \boldsymbol{\theta}_{-lm} \Leftrightarrow Q_{lm} = 0$ (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 $\boldsymbol{\theta}|\boldsymbol{\psi} \sim \mathrm{MVN}\{0|Q(\boldsymbol{\psi}_2)^{-1}\}$, where $\boldsymbol{\psi}_2$ denotes the parameters that are associated with Q rather than Σ.

3.2.2. Integrated Laplace approximations

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

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

In all except the most stylized cases, these will not be analytically tractable. Samples from these distributions could be obtained by using Markov chain Monte Carlo (MCMC) methods but there may be issues when fitting LGMs by using MCMC sampling, 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 INLAs. 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 that are shown in equation (4). For the hyperparameters, the posterior of ψ given $\mathbf{Y}$ can be written as

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

Here a Laplace approximation (LA) is used in the denominator for $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{Y})$. For univariate θ with an integral of the form $\int \exp\{g(\theta)\}$, 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

$$\hat{\sigma}^2 = \left[\frac{\mathrm{d}^2 \log\{g(\theta)\}}{\mathrm{d}\theta^2} \right]^{-1}.$$

The mode of $\tilde{p}(\boldsymbol{\psi}|\mathbf{Y})$ can be found numerically by Newton-type algorithms. Around the mode, the distribution $\log\{\tilde{p}(\boldsymbol{\psi}|\mathbf{Y})\}$ 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 $\tilde{p}(\boldsymbol{\psi}_h^*|\mathbf{Y})$ is obtained from which approximations to the marginal distributions, $\tilde{p}(\boldsymbol{\psi}|\mathbf{Y})$, can be found by using numerical integration.

For the individual model parameters, θ_j,

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

For an LGM, $(\boldsymbol{\theta}_{-j}|\theta_j, \boldsymbol{\psi}, \mathbf{Y})$ will be approximately Gaussian. There are various ways to construct the approximation in the denominator, including a simple Gaussian approximation which will be computationally attractive but may be inaccurate. Alternatively, an 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 $\tilde{p}(\theta_j|\boldsymbol{\psi}, \mathbf{Y})$, aiming to 'correct' the Gaussian approximation for location and skewness (Rue *et al.*, 2009).

The marginal posteriors $\tilde{p}(\boldsymbol{\psi}_h^*|\mathbf{Y})$, evaluated at each of the points $\boldsymbol{\psi}_h^*$, are used to obtain the conditional posteriors, $\tilde{p}(\theta_j|\boldsymbol{\psi}_h^*, \mathbf{Y})$, on a grid of values for θ_j. The marginal posteriors $\tilde{p}(\theta_j|\mathbf{Y})$ are then found by numerical integration:

$$\tilde{p}(\theta_j|\mathbf{Y}) = \sum_h^H \tilde{p}(\theta_j|\boldsymbol{\psi}_h^*, \mathbf{Y})\, \tilde{p}(\boldsymbol{\psi}_h^*|\mathbf{Y})\, \Delta_h,$$

with the integration weights Δ_h being equal when the grid takes the form of a regular lattice.

The model that was presented in Section 3.1 was implemented using R-INLA (Rue *et al.*, 2012) installed on the Balena high performance computing system at the University of Bath

(www.bath.ac.uk/bucs/services/hpc/facilities/). Fitting the model that was described in Section 3.1 to data from the 6003 monitors (and associated covariates) does not itself require the use of a high performance computer 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. The INLA algorithm requires a copy of the model to be stored on a single node which, even with the high memory compute nodes (32 Gbytes per core) available with the Balena high performance computer, 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 data set 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 that was 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 on the basis of their contribution to within-sample model fit and out-of-sample predictive ability.

Details of the variables that were 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 $PM_{2.5}$

Table 1. Variables included in each of five candidate models†

Variable	*Variables and effects in the following models:*									
	(i)		*(ii)*		*(iii)*		*(iv)*		*(v)*	
	Fixed	*Random*	*Fixed*	*Random*	*Fixed*	*Random*	*Fixed*	*Random*	*Fixed*	*Random*
Intercept	✓		✓	✓	✓	✓	✓	✓	✓	✓
X_1‡	✓		✓		✓		✓		✓	
X_2‡	✓		✓		✓		✓		✓	
X_3‡	✓		✓		✓		✓		✓	
X_4	✓		✓	✓	✓	✓	✓	✓	✓	✓
X_5	✓				✓	✓			✓	✓
X_6							✓		✓	
X_7							✓		✓	
X_8§			✓	✓	✓	✓	✓	✓	✓	✓
X_9							✓		✓	

†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 $PM_{2.5}$-estimates; 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.
‡Together with interaction with X_4 and X_5 where they are included within the model.
§Country level random effects are assigned a conditional auto-regressive prior.

Table 2. Summary of results from fitting five candidate models described in Table 1†

Model	R^2	*DIC*	*RMSE* ($\mu g\, m^{-3}$)	*PwRMSE* ($\mu g\, m^{-3}$)
(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)

†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 fits, R^2 and the deviance information criterion DIC are given and, for out-of-sample predictive ability, the root-mean-squared error RMSE and population-weighted root-mean-squared error PwRMSE.

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 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 that was used in the 2013 GBD (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)–(v) are variants of the model that was presented in Section 3.1.

For evaluation, cross-validation was performed using 25 combinations of training (80%) and validation (20%) data sets. 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, and 100 μg m^{-3} or more) 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, the 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 with 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 Fig. 6 which shows PwRMSE for each model by super-region. It can be seen that using any of the hierarchical models based on the structure that was described in Section 3.1 provides an immediate improvement in all metrics when compared with the linear model, with a single global calibration function, used in the 2013 GBD. 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 μg m^{-3} for RMSE and population-weighted RMSE respectively. This improvement can be seen in each of the super-regions (Fig. 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

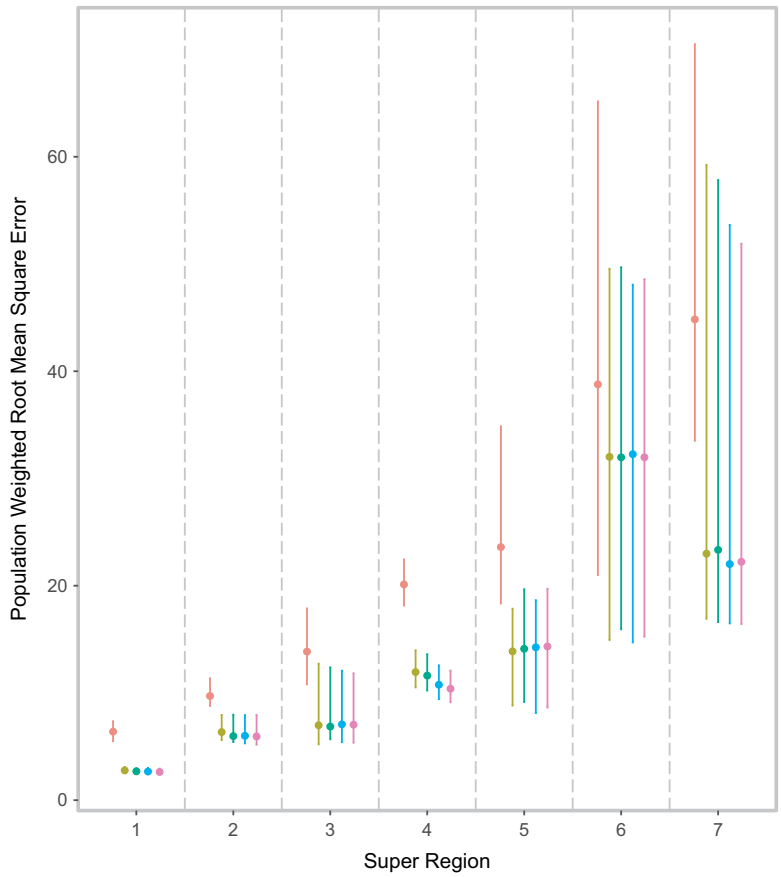
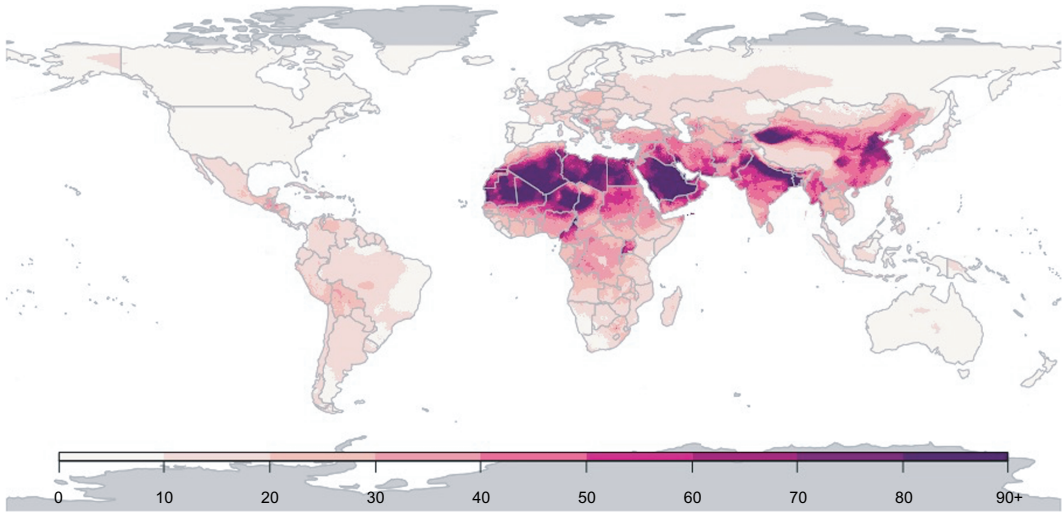


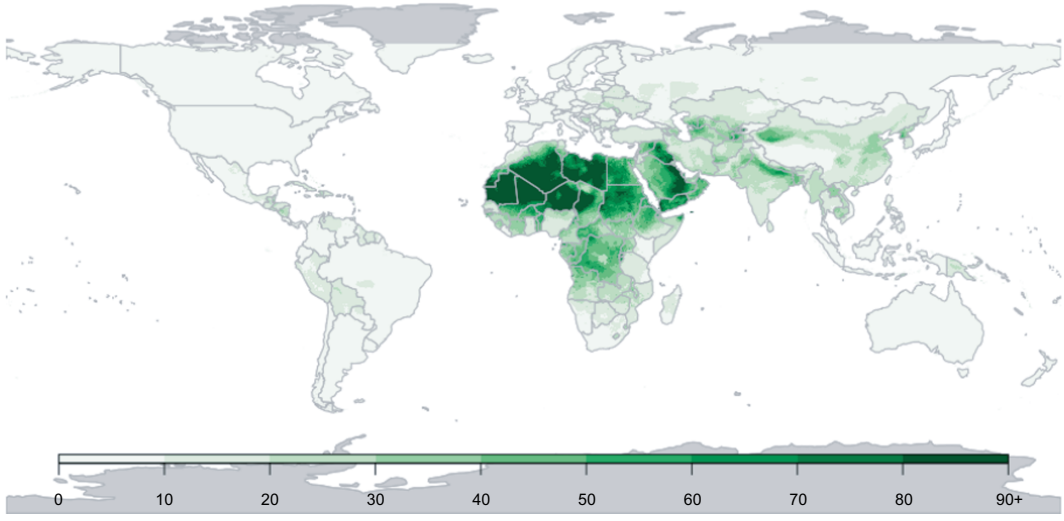
Fig. 6. Summaries of predictive ability of the 2013 GBD model (i) (●) and four candidate models (ii) (●), (iii) (●), (iv) (●) and (v) (●), for each of seven super-regions (for each model, population-weighted root-mean-squared errors (μg m^{-3}) are given with dots denoting the median of the distribution from 25 training–evaluation sets and the vertical lines the range of values): 1, high income; 2, Central Europe, Eastern Europe and Central Asia; 3, Latin America and the Caribbean; 4, South-east Asia, East Asia and Oceania; 5, North Africa–Middle East; 6, sub-Saharan Africa; 7, South Asia

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 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 because the variables that were used in model (iv) are for 2014 whereas the estimates from TM5-FASST are from 2010. Considering the lack of improvement in predictive ability and the increased complexity and computational burden that are involved when incorporating an additional set of random effects, these estimates are not included in the final model (model (iv)).



(a)



(b)

Fig. 7. Estimates of annual $PM_{2.5}$-averages (μg m^{-3}) for 2014 together with associated uncertainty for each grid cell (0.1° × 0.1° resolution) using a Bayesian hierarchical model (see the text for details): (a) medians of posterior distributions; (b) half the width of 95% posterior credible intervals

Predictions from the final model (model (iv)) can be seen in Figs 7(a) and 7(b). The point estimates that are shown in Fig. 7(a) 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 various ways of presenting this uncertainty and Fig. 7(b) 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 the areas of North Africa and the Middle East.

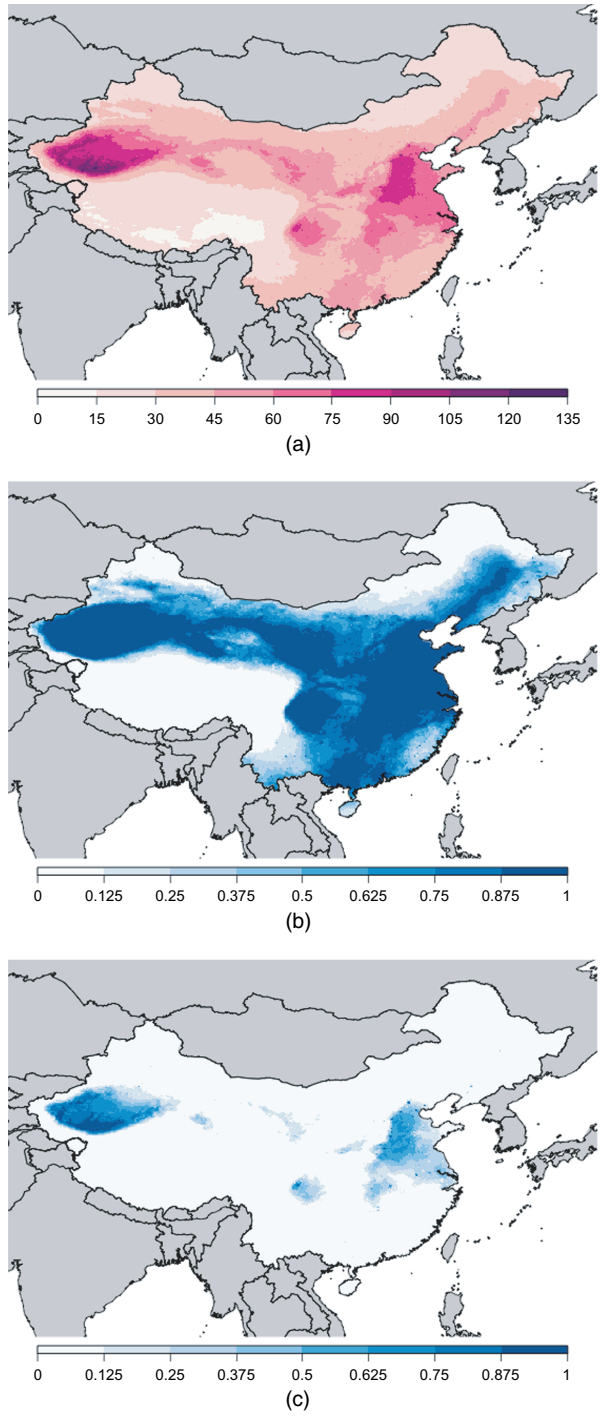


Fig. 8. Estimates of annual mean $PM_{2.5}$-concentrations (μg m^{-3}) for 2014 together with exceedance probabilities by using a Bayesian hierarchical model (see the text for details) for each grid cell (0.1° × 0.1° resolution) in China: (a) medians of posterior distributions; (b) probability of exceeding 35 μg m^{-3}; (c) probability of exceeding 75 μg m^{-3}

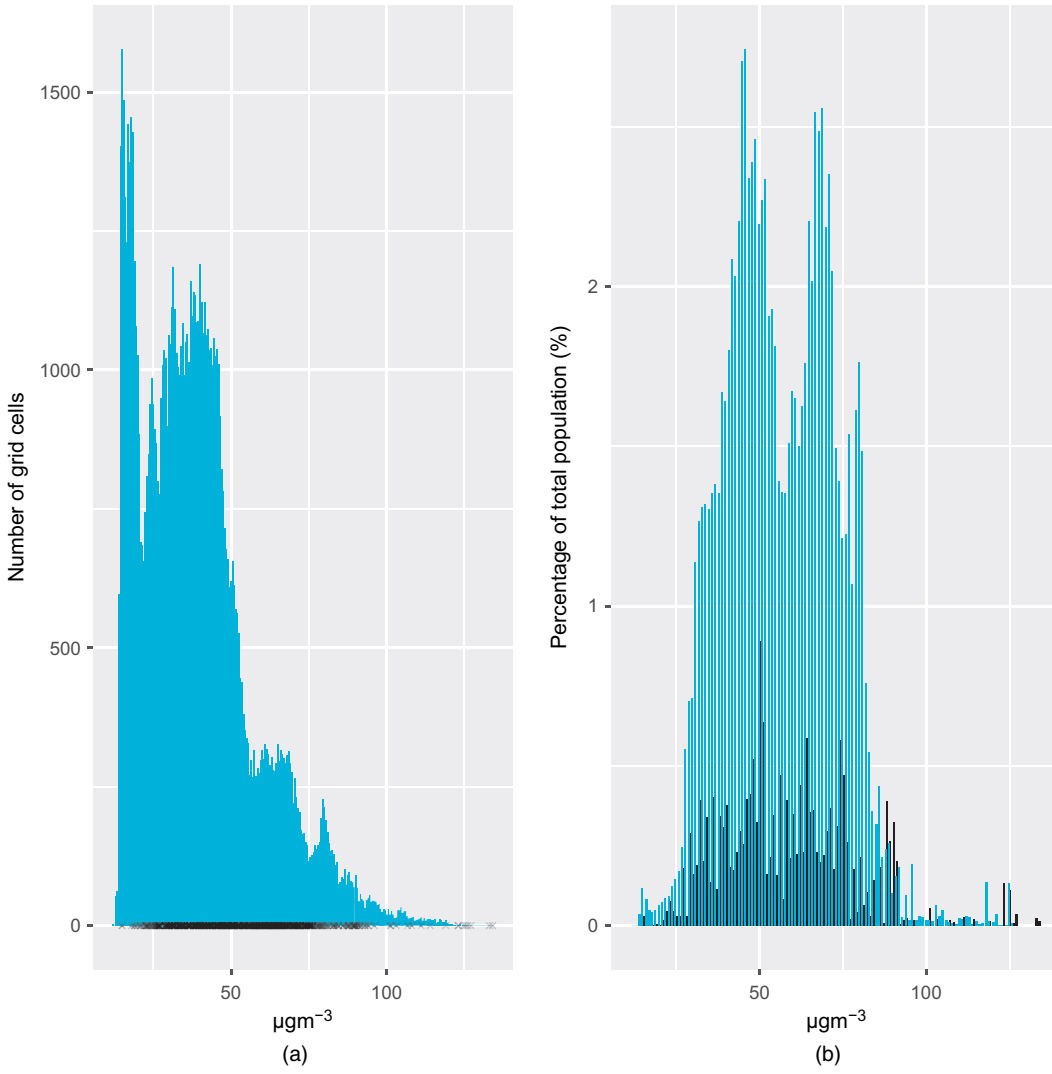


Fig. 9. Distributions of annual mean concentrations and population level exposures for $PM_{2.5}$ (μg m^{-3}) in China: (a) estimated annual average $PM_{2.5}$-concentrations by grid cell (0.1° × 0.1° resolution) (×, annual averages recorded at ground monitors with the level of transparency denoting the density of monitors at each concentration); (b) estimated population level exposures (|) and, for cells containing at least one monitoring station, population-weighted measurements from ground monitors (‖)

The distributions for each cell can also be used to examine the probabilities of exceeding particular thresholds. Fig. 8 shows an example of this and contains predicted concentrations for China (Fig. 8(a)) together with the probability for each cell that the value exceeds 35 μg m^{-3} (Fig. 8(b)) and 75 μg m^{-3} (Fig. 8(c)). 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 Fig. 3(c). The distribution of estimated exposures shown in the map of median values (of the marginal

posterior distributions) shown in Fig. 8(a) can also be seen in Fig. 9(a) for which the profile of air pollution ($PM_{2.5}$) in this country contains three distinct components:

(a) a land mass with low levels of air pollution;
(b) a much larger proportion of the total land mass with (comparatively) high levels;
(c) 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. Fig. 9(b) 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 resides in areas with the lowest concentrations with the vast majority of the population experiencing much higher $PM_{2.5}$-levels.

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 2013 GBD 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 utilized, but only provided an informal analysis of the uncertainty that is 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 is 'borrowed' from the surrounding region and super-region where local monitoring data are inadequate for stable estimation of the coefficients in the calibration models. This is achieved by 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 $PM_{2.5}$-levels 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 means, together with measures of uncertainty, e.g. 95% credible intervals. They can also be used to estimate exceedance probabilities, e.g. the probability of exceeding air quality guidelines. On the basis of 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 μg m^{-3}, which is greater than the 87% that was reported in Brauer *et al*. (2015) for 2013.

In addition to the hierarchical approach to modelling that was used here, the increased availability of ground monitoring data has been utilized in the analysis. Ground measurements were available from 6003 locations (compared with 4073 for the 2013 GBD) 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 data sets (each containing 80% of the overall data with stratified sampling to ensure that samples were representative in terms of the distribution of concentrations within each super-region) and predictions compared with 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 TM5-FASST used in the 2013 GBD 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 predictive ability and are included in the model.

The model that is presented here has been shown to offer improved $PM_{2.5}$-estimates 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 Figs 7 and 8, would be to guide where future monitoring efforts might be focused. It may also be possible to utilize 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 close to each other where there are differences over time would be to use data from only the most recent years. However, such data are often not available in precisely the regions where ground measurements are most needed to produce accurate calibration functions.

In the calibration approach that was used here there is an implicit assumption that the covariates are error free, which may be untenable in practice. When integrating data from many sources, each source will have its own error structures and spatially varying biases. For example, the $PM_{2.5}$-estimates 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 provides a coherent framework in which data from different sources at different levels of aggregation can be integrated, and enables 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 by using MCMC sampling, because of 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 considering grid cells containing only measurement locations within the estimation, after which prediction at unknown locations is relatively straightforward (Chang, 2016). In its current incarnation, using MCMC sampling, Bayesian melding is computationally infeasi-

ble 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 forwards 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 that is associated with the individual data sources. This information can also be incorporated within a health effects model leading to improved characterization 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.

Acknowledgements

The model was developed by a multidisciplinary 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 the first, fourth–ninth and 12th–16th authors of this paper together with members of the WHO (the 10th, 11th and 17th authors). The views that are expressed in this paper 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 was presented and reviewed at the second meeting of the 'Global platform for air quality' meeting, Geneva, August 2015. Matthew Lloyd Thomas is supported by a scholarship from the Engineering and Physical Sciences Research Council Centre for Doctoral Training in Statistical Applied Mathematics at Bath, under 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. *J. Agric. Biol. Environ. Statist.*, **15**, 176–197.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc.* B, **36**, 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. *J. Geophys. Res. Atmos.*, **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. *Environ. Sci. Technol.*, **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. *Environ. Sci. Technol.*, **50**, 79–88.

Brook, R., Rajagopalan, S., Pope, C. R., Brook, J., Bhatnagar, A., Diez-Roux, A., Holguin, F., Hong, Y., Luepker, R., Mittleman, M., Peters, A., Siscovick, D., Smith, S. J., Whitsel, L. 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.

Center for International Earth Science Information Network (2016) *Gridded Population of the World, Version 4 (GPWv4): Administrative Unit Center Points with Population Estimates*. New York: Center for International Earth Science Information Network.

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), ch. 16, pp. 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 *Proc. 11th Int. Conf. Harmonisation within Atmospheric Dispersion Purposes,*

Cambridge (eds D. J. Carruthers and C. A. McHugh). Cambridge: Cambridge Environmental Research Consultants.

Forouzanfar, M. H., Alexander, L., Anderson, H. R., Bachman, V. F., Biryukov, S., Brauer, M., Burnett, R., Casey, D., Coates, M. M., Cohen, A., *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. *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 Sens. Environ.*, **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.

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. *Atmos. Chem. Phys.*, **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. *Environ. Hlth*, **12**, article 43.

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. *Geoscient. Modl Devlpmnt*, **3**, 445–473.

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. *Atmos. Environ.*, **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) The carcinogenicity of outdoor air pollution. *Lancet Oncol.*, **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. (2015) Expert position paper on air pollution and cardiovascular disease. *Eur. Hrt J.*, **36**, no. 2, 83–93.

Poole, D. and Raftery, A. E. (2000) Inference for deterministic simulation models: the Bayesian melding approach. *J. Am. Statist. Ass.*, **95**, 1244–1255.

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

Rue, H., Martino, S. and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *J. R. Statist. Soc.* B, **71**, 319–392.

Rue, H., Martino, S. and Lindgren, F. (2012) The R-INLA Project. Norwegian University for Science and Technology, Trondheim. (Available from `http://www.r-inla.org`.)

Sava, F. and Carlsten, C. (2012) Respiratory health effects of ambient air pollution: an update. *Clin. Chest Med.*, **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 Conf. Abstr.*, **16**, abstract 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. *Environ. Sci. Technol.*, **50**, 3762–3772.

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. *Stoch. Environ. Res. Risk Assessmnt*, **21**, 183–194.

World Health Organization (2013) *Review of Evidence on Health Aspects of Air Pollution—REVIHAAP Project*. Geneva: World Health Organization.

World Health Organization (2016a) *Ambient Air Pollution: a Global Assessment of Exposure and Burden of Disease*. Geneva: World Health Organization.

World Health Organization (2016b) *WHO Global Urban Ambient Air Pollution Database (Update 2016)*. Geneva: World Health Organization.

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