ENVIRONMETRICS
Environmetrics (2008)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/env.965

An evaluation of European air pollution regulations for particulate matter monitored from a heterogeneous network

Sujit K. Sahu[1*,†] and Orietta Nicolis[2]

[1]*School of Mathematics, S3RI, University of Southampton, UK*
[2]*Department of IIMM, University of Bergamo, Italy*

SUMMARY

Statistical methods are needed for evaluating many aspects of air pollution regulations increasingly adopted by many different governments in the European Union. The atmospheric particulate matter (PM) is an important air pollutant for which regulations have been issued recently. A challenging task here is to evaluate the regulations based on data monitored on a heterogeneous network where PM has been observed at a number of sites and a surrogate has been observed at some other sites. This paper develops a hierarchical Bayesian joint space–time model for the PM measurements and its surrogate between which the exact relationship is unknown, and applies the methods to analyse *spatio*-temporal data obtained from a number of sites in Northern Italy. The model is implemented using MCMC techniques and methods are developed to meet the regulatory demands. These enable full inference with regard to process unknowns, calibration, validation, predictions in time and space and evaluation of regulatory standards. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: Bayesian inference; hierarchical model; Markov chain Monte Carlo; separable *spatio*-temporal process; stationarity

1. INTRODUCTION

In order to improve the air quality and to protect the human health, the European community has defined some regulations on limit values of air pollutants, including PM_{10} concentration levels (pollution particles measuring 10 micron or less in diameter and measured in $\mu g/m^3$ units). These limits are being gradually adopted by the member states of the European Union. For example, a directive has been issued by the Italian government in 2002 which states that

1. the daily average PM_{10} concentration should not be over 50,
2. the daily average should not exceed 50 for more than 7 days in a year until the year 2010,
3. the annual average should not be over 40.

Statistical space–time models for data monitored over a set of monitored sites are required to examine compliance with respect to the above directives. Moreover, sound statistical methods, based on a suitable

*Correspondence to: S. K. Sahu, School of Mathematics, S3RI, University of Southampton, UK.
†E-mail: S.K.Sahu@soton.ac.uk

Received 2 April 2008
Accepted 11 September 2008

model, must be developed to evaluate the above regulatory conditions for any particular site in a study region. The primary motivation for the current paper is the lack of development of such a model and model-based methods since the issuance of the directives in 2002. The focus here is to develop methods for data monitored on a heterogeneous network in North Italy over a year.

The reference sampling and measurement of PM_{10} concentrations are based on the collection on a filter of the PM_{10} fraction of ambient particulate matter (PM) and the gravimetric mass determination which we acronymise *low volume sampler gravimetric* (LVG) method and/or instrument. However, in our study region of North Italy the local authority, due to technical, administrative and historical reasons, use a different instrument to measure PM_{10}. They maintain a dense network of automatic monitors based on a *tapered-element oscillating microbalance* (TEOM). These monitors are known to underestimate the true PM_{10} levels given by the reference LVG method. Moreover, there is heterogeneity between the two sets of measurements. Thus there is an urgent need to correct the TEOM measurements so that those are comparable with the LVG measurements.

This paper develops a joint space time model for data provided by the above two heterogeneous instruments, LVG and TEOM. Except for one station, called Consolata, the LVG and TEOM were observed at completely different sites. This gives rise to a problem of spatial misalignment. Modelling the two measurements in each site using a bivariate distribution will not be satisfactory since such an approach will require imputation of the unobserved measurement at each site. This requires imputation of at least 50% missing data since the sites where only LVG has been observed will have missing TEOM measurements and the sites where only TEOM has been observed will have missing LVG measurements. Our modelling strategy avoids the problem by incorporating a latent space–time process common to both types of measurements. This induces dependence between the two measurements in space and time, and enables learning of the common underlying *spatio*-temporal process using the heterogeneous measurements. The model also includes seasonality effects often found in PM_{10} data. The full Bayesian model, implemented using MCMC, enables: calibration for the station where both were measured, validation at a number of sites and spatial interpolation and forecasting of PM_{10}.

The model for daily PM_{10} levels allows us to aggregate to any desired spatial and temporal summary of PM_{10}. In particular, the annual averages and the total number of days in a year for which the daily average exceeded 50 and their associated uncertainties are estimated. Moreover, the probability that the annual average is greater than 40 at any particular site is spatially interpolated in our Bayesian setup. Non-model-based interpolation of these annual averages and the extremes encourages too much smoothing and will lead to biased results, possibly without proper estimation of the associated uncertainties, see for example Sahu *et al.* (2007).

Short term space–time statistical modelling for PM_{10} has also been considered by Shaddick and Wakefield (2002) and by Sun *et al.* (2000) from the hierarchical Bayesian point of view. Zidek *et al.* (2002) developed predictive distributions on non-monitored PM_{10} concentrations in Vancouver, Canada. Cressie *et al.* (1999) compared Kriging and Markov Random fields models in the prediction of PM_{10} concentrations around Pittsburgh, USA. For the $PM_{2.5}$, Smith *et al.* (2003) proposed a *spatio*-temporal model using a nonparametric approach. Kibria *et al.* (2002) presented a multivariate spatial prediction methodology in a Bayesian context for the prediction of $PM_{2.5}$ in Philadelphia, USA. Sahu and Mardia (2005) present a short-term forecasting analysis of $PM_{2.5}$ data in New York City during 2002: within a Bayesian hierarchical structure, they model the spatial structure with principal kriging functions and the time component is modelled by a vector random-walk process. Sahu *et al.* (2007) develop methods for assessing trend in ozone levels using high-resolution space–time modelling.

Hauck *et al.* (2004) studied the calibration problem for different measuring instruments using separate regression models for different sites and seasons. Thus their method does not allow spatial interpolation.

DOI: 10.1002/env

The unified *spatio*-temporal model presented in this paper, however, allows spatial interpolation and temporal aggregation. McBride and Clyde (2003), Fassò and Nicolis (2005) and Fassò *et al.* (2007) also consider the problem of calibration using geo-statistical state-space approaches. Our methods, however, are fully Bayesian and in some particular cases produces better results than those from Fassò and Nicolis (2005) when applied to the same dataset from North Italy. In addition, they did not develop methods for making inference for the annual summaries.

Sahu *et al.* (2006) propose a hierarchical space-time model for $PM_{2.5}$ that introduces two *spatio*-temporal processes, one capturing rural or background effects, the second adding extra variability for urban/suburban locations. By weighting these two processes using population density surfaces they obtain models with non-stationary covariance structures. They, however, model weekly averages obtained from a single network of monitoring sites. Thus, they did not consider the problem of modelling data from a heterogeneous network and the associated problem of calibration. However, they also developed methods for approximating the annual averages, but here we use exact calculations based on MCMC.

The structure of the paper is as follows. In Section 2, we provide a description of the data with summary tables and exploratory graphics. Section 3 develops the Bayesian *spatio*-temporal model that accounts for monitor type, seasonality and random effects. Bayesian prediction methods are detailed in Section 4. Model-based data analyses are presented in Section 5. Some concluding remarks are given in Section 6. The Appendix contains the conditional posterior distributions needed for Gibbs sampling and predictions.

2. THE DATASET

We consider the PM_{10} daily concentrations for $T = 365$ days in the year 2003. The study region covers approximately an area of 400 by 200 km grid and the monitors are located in three regions in North Italy, see Figure 1. The monitoring sites were near city centres and in rural areas covered

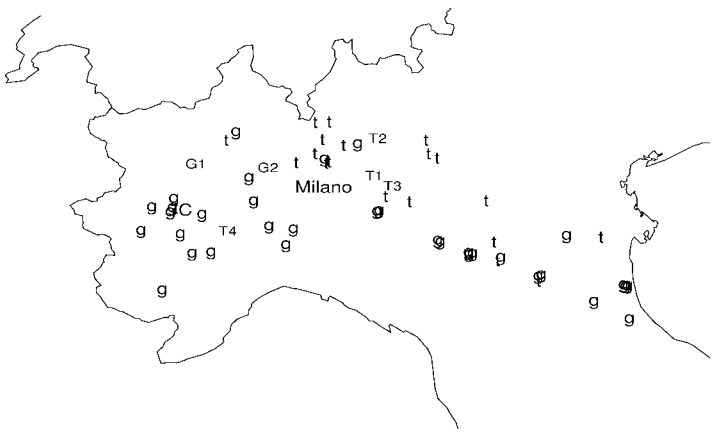

Figure 1. The 54 monitoring sites in North Italy. The 34 LVG sites are denoted bg 'g' and the 20 TEOM sites are plotted as 't'. The six validation sites are marked as G1, G2 (LVG) and T1, T2, T3, T4 (TEOM). The site Consolata is denoted by 'C'

DOI: 10.1002/env

by the main roads. The network of sites is characterised by instrument heterogeneity: some regions in the north have many TEOM sites and few LVG sites whilst the opposite holds for others. The regions covering the TEOM and LVG sites, however, are not disjoint, see, for example, the sites near Milano and further to the east. We analyse data from $n = 54$ stations composed of $n_1 = 34$ LVG monitors and $n_2 = 20$ TEOM monitors. Out of these 19 710 (54×365) observations, less than 5% are missing.

In one of the 20 TEOM monitoring stations, Consolata, denoted by C in Figure 1, we also have the LVG measurements, but this site has not been included among the 34 LVG sites. The LVG measurements from this site are used for model validation since the objective of this paper is to make prediction and inference for the LVG measurements. This is because pollution standards are calculated using LVG measurements, as mentioned in the introduction, so it is more important to validate the model and prediction methods for the LVG measurements.

Data from six additional stations, two LVG and four TEOM labelled respectively by G1, G2, T1, T2, T3 and T4 in Figure 1, are used for validation of the model. The station G1 is not very close to any data monitoring site and the station T4 is a TEOM monitoring site in the middle of several LVG sites—the nearest modelled TEOM site, Consolata, is quite a distance away. Only 1444 (65.93%) out of 2190 (6×365) observations are available from these six stations. These observations are used for validation rather than estimation due to this high percentage of missingness caused by instrument malfunctioning.

Figure 2 provides boxplots of the data grouped by months and instrument type. The plot shows that both the LVG and TEOM measurements are affected similarly by strong seasonal effects. The measurements in the five winter months, November, December, January, February and March have higher levels with higher variability. Table 1 provides the variances of LVG and TEOM measurements on the original, square root and logarithmic scales for data from two seasons: summer and winter. On the log-scale, which we shall adopt for modelling, there are no significant

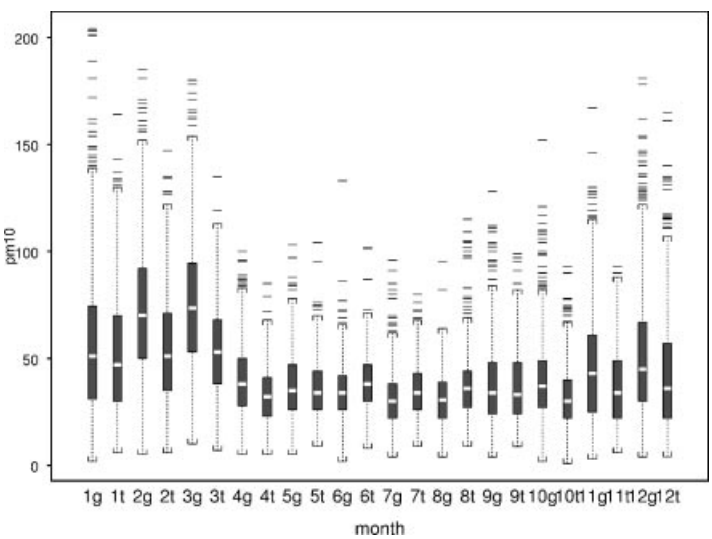

Figure 2. Boxplot of the PM_{10} concentrations by months and instrument type. The LVG measurements in month i are denoted by 'ig' and the TEOM measurements are denoted by 'it', $i = 1, \ldots, 12$

DOI: 10.1002/env

Table 1. The variances of LVG and TEOM measurements on three different scales

	Original		Square-root		Logarithm	
	LVG	TEOM	LVG	TEOM	LVG	TEOM
Summer	253.73	204.50	1.73	1.42	0.23	0.18
Winter	1053.30	659.16	4.50	3.46	0.39	0.36

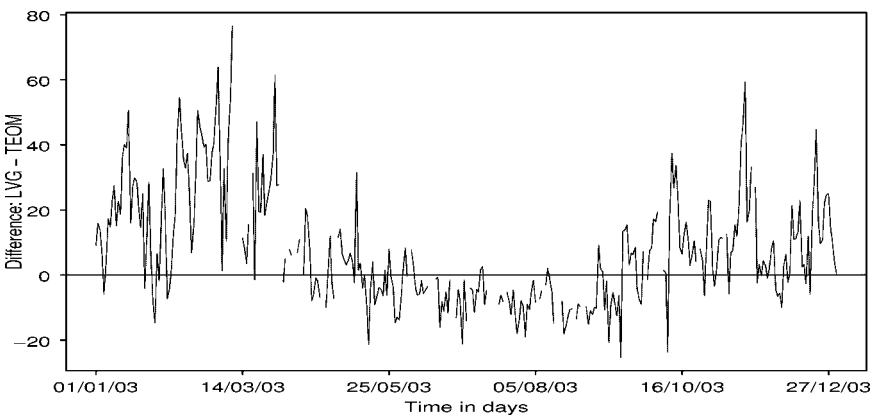

Figure 3. Time series plot of the differences between LVG and TEOM measurements in the site Consolata

differences between the summer and winter variances (the ratio of the larger to the smaller is less than three).

The time series plot of the difference between the LVG and TEOM measurements from the station Consolata is given in Figure 3. The plot shows that the LVG measurements are higher in the winter months but are lower in the three summer months. In fact, this plot shows that an ad-hoc rule, often used in practise, that LVG measurements are on average 1.3 times the TEOM measurements is unlikely to be true. These facts justify the separate models for mean developed in Section 3.

Inspection of the data shows that variances increase with the mean levels, see Figure 4 where the variance for each station is plotted against the mean, using data on the original, square root and logarithmic scales. The log-scale is seen to be the best transformation for stabilising the variance (see also Table 1) and removing the mean–variance relationships. Thus, we choose to model data on the logarithmic scale, however, report all predictions and their standard deviations on the original scale for ease of interpretation. Modelling of the data using the square-root transformation gave considerable poorer model fit and prediction.

3. MODELLING DETAILS

Let $Z_g(\mathrm{s}, t)$ denote the logarithm of the observed LVG measurement at a location s and at time t. Recall that we have LVG measurements from $n_1 = 34$ stations, $\mathrm{s}_1, \ldots, \mathrm{s}_{n_1}$, at each of $T = 365$ days. Let

DOI: 10.1002/env

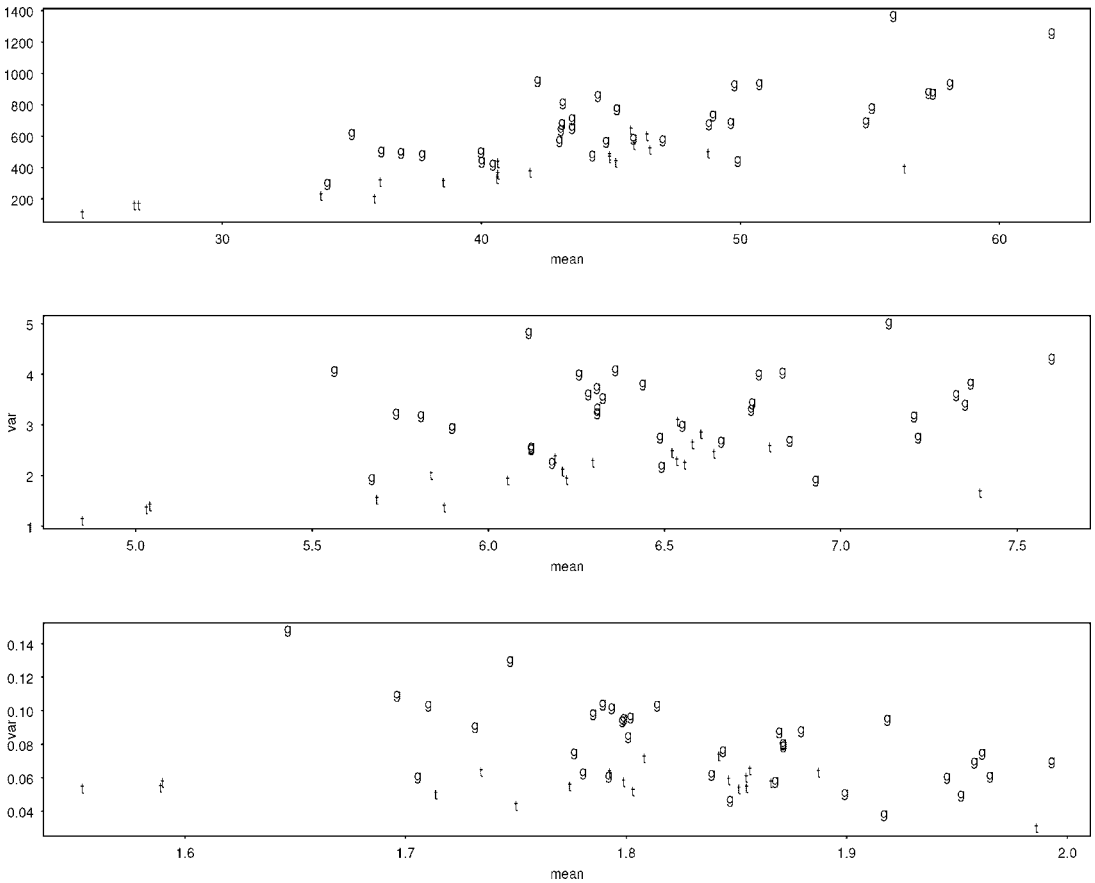

Figure 4. The variance against the mean of PM concentration levels in each of the 54 sites. Top panel is on the original scale, middle panel is on the square-root scale and bottom panel is on the log scale. The LVG sites are denoted by 'g' and the TEOM sites are denoted 't'

$Z_h(\mathrm{s}, t)$ denote the logarithm of the observed TEOM measurement at a location s and at time t. There are TEOM data from $n_2 = 20$ stations, $\mathrm{s}_{n_1+1}, \ldots, \mathrm{s}_{n_1+n_2}$, at each of T days.

First, we assume the hierarchical models:

$$Z_g(\mathrm{s}_i, t) = Y_g(\mathrm{s}_i, t) + \epsilon_g(\mathrm{s}_i, t), i = 1, \ldots, n_1, t = 1, \ldots, T \tag{1}$$

and

$$Z_h(\mathrm{s}_i, t) = Y_h(\mathrm{s}_i, t) + \epsilon_h(\mathrm{s}_i, t), i = n_1 + 1, \ldots, n_1 + n_2, t = 1, \ldots, T \tag{2}$$

where $Y_g(\mathrm{s}, t)$ and $Y_h(\mathrm{s}, t)$ are true space–time processes for LVG and TEOM measurements respectively; $\epsilon_g(\mathrm{s}_i, t)$ and $\epsilon_h(\mathrm{s}_i, t)$ are independent white noise processes assumed to follow $N(0, \sigma_g^2)$ and $N(0, \sigma_h^2)$, respectively.

DOI: 10.1002/env

We suppose that the *spatio*-temporal processes $Y_g(\mathrm{s}, t)$ and $Y_h(\mathrm{s}, t)$ have different mean structures but are governed by a single latent *spatio*-temporal PM pollution process $u(\mathrm{s}, t)$. This zero mean *spatio*-temporal process introduces dependence between the Y_g and Y_h processes, and these in turn influence dependencies between the observation processes $Z_g(\mathrm{s}, t)$ and $Z_h(\mathrm{s}, t)$. Thus we assume that

$$Y_g(\mathrm{s}_i, t) = \mu_g(\mathrm{s}_i, t) + u(\mathrm{s}_i, t), i = 1, \ldots, n_1, t = 1, \ldots, T \tag{3}$$

and

$$Y_h(\mathrm{s}_i, t) = \mu_h(\mathrm{s}_i, t) + u(\mathrm{s}_i, t), i = n_1 + 1, \ldots, n_1 + n_2, t = 1, \ldots, T \tag{4}$$

We model the means $\mu_g(\mathrm{s}, t)$ and $\mu_h(\mathrm{s}, t)$ with monthly seasonal effects. We define monthly seasonal indicators, $v(t, m)$ as follows:

$$v(t, m) = \begin{cases} 1 & \text{if the time } t \text{ is in the } m\text{th month} \\ 0 & \text{otherwise} \end{cases}$$

for $t = 1, \ldots, 365$, $m = 1, \ldots, 12$. Thus we have

$$\mu_g(\mathrm{s}, t) = \mathbf{x}_t'\boldsymbol{\beta}_g, \text{ and } \mu_h(\mathrm{s}, t) = \mathbf{x}_t'\boldsymbol{\beta}_h$$

where the $p(=12)$-dimensional vector $\mathbf{x}_t'$ is given by $(1, v(t, 2), \ldots, v(t, 12))'$, $\boldsymbol{\beta}_g = (\beta_g(1), \ldots, \beta_g(12))'$ and $\boldsymbol{\beta}_h = (\beta_h(1), \ldots, \beta_h(12))'$. Note that for identifiability purposes we do not include the dummy for January, $v(t, 1)$, in the model. Thus, the parameters $\beta_g(1)$ and $\beta_h(1)$ represent the overall means and $\beta_g(k)$ and $\beta_h(k)$ represent the differences between the kth month and January, $k = 2, \ldots, 12$.

Fassò and Nicolis (2005) consider the pairs of model

$$Z_g(\mathrm{s}_i, t) = \mu(\mathrm{s}_i, t) + \epsilon_g(\mathrm{s}_i, t), Z_h(\mathrm{s}_i, t) = \alpha(t) + \beta\mu(\mathrm{s}_i, t) + \epsilon_h(\mathrm{s}_i, t)$$

where $\alpha(t)$ is an auto-regressive process. The unobserved mean process $\mu(s, t)$ was estimated by principal fields (components) decomposition (see e.g. Mardia *et al.*, 1998; Wikle and Cressie, 1999). However, prediction using this approach introduces a lot of more variability through the temporally varying random auto-regressive intercept $\alpha(t)$ as we shall see in Section 5. Moreover, by assigning different mean functions in Equations (3) and (4) we introduce more flexibility in the proposed model.

Some further remarks regarding the chosen mean function are appropriate. In an attempt to improve the mean function we have included some linear and spline functions of site characteristics such as the latitude and longitude. These modifications did not improve the validations and forecasting at all. Moreover, some formal Bayesian model choice criteria selected the simpler model for the mean function adopted.

Let $\mathbf{U}$ be given by

$$\mathbf{U}_{T\times n} = \begin{pmatrix} u(\mathrm{s}_1, 1) & \cdots & u(\mathrm{s}_{n_1}, 1) & u(\mathrm{s}_{n_1+1}, 1) & \cdots & u(\mathrm{s}_n, 1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u(\mathrm{s}_1, T) & \cdots & u(\mathrm{s}_{n_1}, T) & u(\mathrm{s}_{n_1+1}, T) & \cdots & u(\mathrm{s}_n, T) \end{pmatrix}$$

DOI: 10.1002/env

Let the ith column of $\mathbf{U}$ be denoted by $\mathbf{u}_i$, so that $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$. We also let $\mathbf{u}$ denote the vector obtained by the stacking the columns of the matrix $\mathbf{U}$.

Let Σ_s and Σ_t denote the spatial and temporal correlation matrices of the $u(\mathrm{s}, t)$ process. That is, for $i, j = 1, \ldots, n$ and $k, l = 1, \ldots, T$, we have

$$(\Sigma_s)_{ij} = \rho_s(\mathrm{s}_i - \mathrm{s}_j; \phi_s), (\Sigma_t)_{kl} = \rho_t(k - l; \phi_t)$$

where we assume the exponential covariance structure $\rho(d; \phi) = \exp(-\phi|d|)$. Also for convenience, we assume a separable covariance structure, see for example, Mardia and Goodall (1993), for the latent $u(\mathrm{s}, t)$ process. The prior specification is given by

$$\mathbf{u} \sim N\big(\mathbf{0}, \sigma_u^2 \Sigma_s \otimes \Sigma_t\big)$$

where $\otimes$ denotes the Kronecker product. Formal non-Bayesian tests of separability, see for example, Fuentes (2006) and Mitchell *et al.* (2005), are available for smaller datasets with replications in both space and time. For our large dataset (about 19 000 data points) without replications in space and time such tests are not feasible—those may require us to store nT by nT matrices. Besides our separability and stationarity assumptions are for the latent $u(\mathrm{s}, t)$ process and not for the observed data process $z_g(\mathrm{s}, t)$ or $z_h(\mathrm{s}, t)$. Nevertheless, we validate the assumptions as part of the full Bayesian model specification by a large number out of sample predictions with set-aside data in Section 5.

Ideally, $\boldsymbol{\phi} = (\phi_s, \phi_t)'$, should be estimated within the Bayesian model as well. However, in a classical inference setting it is not possible to consistently estimate all the parameters ϕ and σ^2 in a typical model for spatial data with a covariance function belonging to the Matèrn family, see Zhang (2004). Moreover, Stein (1999) shows that spatial interpolation is sensitive to the product $\sigma^2\phi$ but not to either one individually. Moreover, in our Bayesian inference setup using Gibbs sampling joint estimation is often poorly behaved due to weak identifiability and extreme slow-mixing of the associated Markov chains under vague prior distributions for $\boldsymbol{\phi}$. In addition, the full conditional distribution of $\boldsymbol{\phi}$ is not conjugate and sampling those in a Gibbs sampler requires expensive likelihood evaluations at each iteration. In Section 5 we shall choose optimal values of $\boldsymbol{\phi}$ using a validation mean square error criterion and estimate the variances conditional on those values. Note that the full conditional distributions of the variances are conjugate under the assumption of conjugate prior distributions.

Denote the unknown parameters by $\boldsymbol{\theta} = (\boldsymbol{\beta}_g', \boldsymbol{\beta}_h', \sigma_g^2, \sigma_h^2, \sigma_u^2)'$. We assume that, *a priori*, the $\boldsymbol{\beta}_g \sim N(0, A^2 I)$ where I denotes the identity matrix and A^2 is a large positive constant so that the prior specification is flat. Similarly, we assume $\boldsymbol{\beta}_h \sim N(0, A^2 I)$. For the three variance parameters σ_g^2, σ_h^2 and σ_u^2 we assume independent proper inverse gamma prior distributions, $IG(a, b)$ (with $a > 1$, hence, mean $b/(a-1)$) to avoid having an improper posterior distribution. In our numerical example we set $a = 2$ and $b = 1$ so that the resulting prior distribution has mean 1 and infinite variance.

Let $\mathbf{Z}$ and $\mathbf{W}$ denote respectively the observed and missing data points. The log-likelihood is given by

$$l(\boldsymbol{\theta}, \mathbf{U}, \mathbf{W}; \mathbf{z}) \propto -\frac{n_1 T}{2}\log(\sigma_g^2) - \frac{1}{2\sigma_g^2}\sum_{i=1}^{n_1}\sum_{t=1}^{T}\{z_g(\mathrm{s}_i, t) - \mu_g(\mathrm{s}_i, t) - u(\mathrm{s}_i, t)\}^2$$

$$-\frac{n_2 T}{2}\log(\sigma_h^2) - \frac{1}{2\sigma_h^2}\sum_{i=n_1+1}^{n}\sum_{t=1}^{T}\{z_h(\mathrm{s}_i, t) - \mu_h(\mathrm{s}_i, t) - u(\mathrm{s}_i, t)\}^2$$

DOI: 10.1002/env

Hence the log of the joint posterior distribution is given by

$$\log\left\{\pi(\boldsymbol{\theta}, \mathbf{U}, \mathbf{W}|\mathbf{z})\right\} \propto l(\boldsymbol{\theta}, \mathbf{U}, \mathbf{W}; \mathbf{z}) + \log(\pi(\boldsymbol{\beta}_g, \boldsymbol{\beta}_h)) + \log\left\{\pi(\sigma_g^2, \sigma_h^2, \sigma_u^2)\right\}$$
$$-\frac{nT}{2}\log(\sigma_u^2) - \frac{T}{2}\log|\Sigma_s| - \frac{n}{2}\log|\Sigma_t| - \frac{1}{2\sigma_u^2}\mathbf{u}'(\Sigma_s^{-1} \otimes \Sigma_t^{-1})\mathbf{u}$$

where $\pi(\boldsymbol{\beta}_g, \boldsymbol{\beta}_h)$ and $\pi(\sigma_g^2, \sigma_h^2, \sigma_u^2)$ denote the prior distributions.

4. PREDICTION DETAILS

4.1. Calibration

For calibration purposes we want to predict LVG (or TEOM) at one of the sampled locations, $\mathrm{s}_{n_1+1}, \ldots, \mathrm{s}_n$ (or $\mathrm{s}_1, \ldots, \mathrm{s}_{n_1}$). To predict the LVG at a location s at a time t we see from Equations (1) and (3) that

$$Z_g(\mathrm{s}, t) \sim N\left(\mu_g(\mathrm{s}, t) + u(\mathrm{s}, t), \sigma_g^2\right) \tag{5}$$

The predictions are obtained using the distribution:

$$\pi(Z_g(\mathrm{s}, t)|\mathbf{z}) = \int \pi(Z_g(\mathrm{s}, t)|\boldsymbol{\theta}, \mathbf{U}, \mathbf{W})\pi(\boldsymbol{\theta}, \mathbf{U}, \mathbf{W}|\mathbf{z})d\mathbf{W}d\mathbf{U}d\boldsymbol{\theta} \tag{6}$$

We perform this integration using the draws from the posterior distribution $\pi(\boldsymbol{\theta}, \mathbf{U}, \mathbf{W}|\mathbf{z})$. In particular, at each MCMC iteration we also sample from the distribution (5). If, instead we want to predict TEOM at one of the LVG sites $\mathrm{s}_1, \ldots, \mathrm{s}_{n_1}$, we use the above methodology with obvious modifications; in particular, we simulate a new $z_h(\mathrm{s}, t)$ from $Z_h(\mathrm{s}, t) \sim N(\mu_h(\mathrm{s}, t) + u(\mathrm{s}, t), \sigma_h^2)$ at each MCMC iteration. We then exponentiate the realisations to obtain the values in original scale.

4.2. Prediction at new locations

Analogous to Equation (5), at a new location s′ and a time point t', $Z_g(\mathrm{s}', t')$ follows $N(\mu_g(\mathrm{s}', t') + u(\mathrm{s}', t'), \sigma_g^2)$. The posterior predictive distribution of $Z_g(\mathrm{s}', t')$ given $\mathbf{z}$ is

$$\pi(Z_g(\mathrm{s}', t')|\mathbf{z}) = \int \pi\left(Z_g(\mathrm{s}', t')|\boldsymbol{\theta}, u(\mathrm{s}', t')\right)\pi\left(u(\mathrm{s}', t')|\mathbf{U}, \sigma_u^2\right)\pi(\boldsymbol{\theta}, \mathbf{U}, \mathbf{W}|\mathbf{z})du(\mathrm{s}', t')d\mathbf{W}d\mathbf{U}d\boldsymbol{\theta} \tag{7}$$

When using MCMC methods to draw samples from the posterior, the predictive distribution (7) is sampled by composition; draws from the posterior, $\pi(\boldsymbol{\theta}, \mathbf{U}, \mathbf{W}|\mathbf{z})$ enable draws from $\pi(u(\mathrm{s}', t')|\mathbf{U}, \sigma_u^2)$ and thus draws for $Z_g(\mathrm{s}', t')$. (The distribution $\pi(u(\mathrm{s}', t')|\mathbf{U}, \sigma_u^2)$ is derived in the Appendix.) As before we exponentiate the realisations and calculate the summaries to obtain the predictions in the original scale, denoted by $\hat{Z}_{g,\mathrm{orig}}(\mathrm{s}', t')$.

DOI: 10.1002/env

4.3. Annual summaries

Using the above details we are able to predict $Z_g(\mathrm{s}, t)$ at any location s and any time point t. Let $Z^{(j)}_{g,\mathrm{orig}}(\mathrm{s}, t)$ denote the jth MCMC iterate of the LVG value at location s and time t. The annual average at the jth MCMC iteration is obtained by

$$\bar{Z}^{(j)}_{g,\mathrm{orig}}(\mathrm{s}) = \frac{1}{365}\sum_{t=1}^{365} Z^{(j)}_{g,\mathrm{orig}}(\mathrm{s}, t)$$

In order to estimate the probability that the annual average is greater than 40 at a site s we simply calculate the indicator random variables $I(\bar{Z}^{(j)}_{g,\mathrm{orig}}(\mathrm{s}) > 40)$. The number of days the daily average exceeding 50 at the jth MCMC iteration is obtained by calculating

$$N^{(j)}_g(\mathrm{s}) = \sum_{t=1}^{365} I\left(Z^{(j)}_{g,\mathrm{orig}}(\mathrm{s}, t) > 50\right)$$

The summaries of the MCMC iterates $\bar{Z}^{(j)}_{g,\mathrm{orig}}(\mathrm{s})$, $I(\bar{Z}^{(j)}_{g,\mathrm{orig}}(\mathrm{s}) > 40)$ and $N^{(j)}_g(\mathrm{s})$ provide the model-based prediction and the associated uncertainties of the annual average, the probability that the annual average is greater than 40, and the number of days the daily average exceeding 50 at a particular location s, respectively.

5. MODEL-BASED ANALYSIS

The decay parameters $\boldsymbol{\phi} = (\phi_s, \phi_t)$ are selected by a validation criterion. We consider data from six sites (see Section 2) for the validation of the model. For each pair of values of ϕ_s and ϕ_t on a two-dimensional grid we evaluate the mean square error:

$$\mathrm{MSE} = \frac{1}{k}\sum_{i=1}^{6}\sum_{t=1}^{365}\left(Z_{\mathrm{g,\ orig}}(\mathrm{s}_i^*, t) - \hat{Z}_{\mathrm{g,\ orig}}(\mathrm{s}_i^*, t)\right)^2 I\left(Z_{\mathrm{g,\ orig}}(\mathrm{s}_i^*, t)\right)$$

where $I\left(Z_{\mathrm{g,\ orig}}(\mathrm{s}_i^*, t)\right) = 1$ if $Z_{\mathrm{g,\ orig}}(\mathrm{s}_i^*, t)$ has been observed and 0 otherwise and k is the total number of available observations for validation, 1444 for our dataset, see Section 2. The optimal values of ϕ_s and ϕ_t are found to be 0.02 and 0.75, respectively. The MSE criterion is not very sensitive to changes in values of $\boldsymbol{\phi}$ near this optimal value. These optimal values suggest that spatial correlation decays at a distance of about 150 km while the temporal correlation decays in about 4 days—both of these are plausible since there may be high spatial correlation in daily data and temporal correlation may persist for 3–5 days to account for weekend/weekday effects. See Sahu *et al.* (2006) for more detailed discussions regarding the choice of the decay parameters $\boldsymbol{\phi}$.

Tables 2 and 3 show the parameter estimates, their posterior standard deviations and the associated 95% credible intervals. The estimates of the variance components σ_g^2, σ_h^2 and σ_u^2 show that more variation is explained by the *spatio-* temporal effects than the measurements errors. Moreover, as expected σ_g^2 is estimated to be higher than σ_h^2. The overall mean parameter for LVG, $\beta_g(1)$, is significantly higher than

DOI: 10.1002/env

Table 2. The estimates of the variance components

	Mean	SD	95% interval
σ_g^2	0.042	0.001	(0.040, 0.044)
σ_h^2	0.003	0.001	(0.002, 0.004)
σ_u^2	0.181	0.003	(0.175, 0.188)

Table 3. The estimates of $\beta_g(k)$ and $\beta_h(k)$, $k = 1, \ldots, 12$

	$\beta_g(k)$			$\beta_h(k)$			$\beta_g(k) - \beta_h(k)$
	Mean	SD	95%CI	Mean	SD	95%CI	95%CI
Overall	3.82	0.04	(3.73, 3.90)	3.47	0.04	(3.37, 3.54)	(0.31, 0.36)
February	0.23	0.07	(0.11, 0.40)	0.10	0.07	(−0.03, 0.27)	(0.45, 0.53)
March	0.31	0.05	(0.21, 0.43)	0.23	0.05	(0.14, 0.34)	(0.40, 0.48)
April	−0.24	0.06	(−0.38, −0.15)	−0.12	0.06	(−0.28, −0.03)	(0.19, 0.28)
May	−0.36	0.05	(−0.46, −0.26)	−0.03	0.05	(−0.13, 0.06)	(−0.01, 0.07)
June	−0.41	0.06	(−0.54, −0.31)	0.04	0.06	(−0.09, 0.14)	(−0.14, −0.06)
July	−0.50	0.06	(−0.59, −0.36)	0.00	0.06	(−0.08, 0.14)	(−0.18, −0.10)
August	−0.56	0.05	(−0.66, −0.45)	0.00	0.05	(−0.09, 0.13)	(−0.25, −0.16)
September	−0.40	0.05	(−0.49, −0.29)	−0.07	0.04	(−0.14, 0.03)	(−0.02, 0.07)
October	−0.33	0.05	(−0.43, −0.22)	−0.21	0.05	(−0.29, −0.10)	(0.19, 0.28)
November	−0.26	0.07	(−0.38, −0.10)	−0.17	0.08	(−0.29, 0.00)	(0.22, 0.30)
December	−0.22	0.06	(−0.34, −0.11)	−0.18	0.06	(−0.31, −0.09)	(0.28, 0.36)

the same for TEOM, $\beta_h(1)$ (their 95% credible intervals do not overalap), as expected. Moreover, the 95% credible intervals for the differences, $\beta_g(k) - \beta_h(k)$, are provided in the last column of Table 3. These differences and the estimates of the parameters $\beta_g(k)$ and $\beta_h(k)$ for $k = 2, \ldots, 12$ show the overall pattern in LVG and TEOM levels seen in Figure 2.

To test whether any two particular months k_1 and k_2, say, can be collapsed in the seasonal model, we form the contrasts $\beta_g^{(j)}(k_1) - \beta_g^{(j)}(k_2)$ for LVG and $\beta_h^{(j)}(k_1) - \beta_h^{(j)}(k_2)$ for TEOM at each MCMC iteration, $j \geq 1$ and test the significance at the end of the MCMC run. The results are provided in Table 4. Since many differences are significant, a simpler summer–winter collapsed model either for LVG or TEOM will not be suitable here.

We now consider the calibration problem for the station Consolata. As mentioned in Section 2 this station measured both LVG and TEOM but we only modelled the TEOM measurements. Using the calibration methods described in Section 4 we have predicted the LVG measurements, see Figure 5. The 95% prediction intervals (not shown) contain 94% of the actual observations and the sharp drop in actual LVG values in the beginning of the summer is matched by the model. The predictions from the Fassò and Nicolis (2005) model are plotted as dashed lines. Clearly the proposed Bayesian *spatio*-temporal model shows a great deal of flexibility and perform much better than their non-Bayesian state-space type regression model.

Next, we consider validation for the 1444 data points from the six reserved sites. The plot is provided in Figure 6. The two plots in the top row of this figure are for two sites, G1 and G2, validating LVG values and the remaining four plots are for the TEOM sites, T1–T4. The actual site-wise coverage

DOI: 10.1002/env

Table 4. Significant (S) and non-significant (N) differences between the monthly contrasts formed using the months in row and column

	January	February	March	April	May	June	July	August	September	October	November	December
January	—	S	S	S	S	S	S	S	S	S	S	S
February	N	—	N	S	S	S	S	S	S	S	S	S
March	S	S	—	S	S	S	S	S	S	S	S	S
April	S	S	S	—	S	S	S	S	S	N	N	N
May	N	S	S	S	—	N	S	S	N	N	N	S
June	N	N	S	S	N	—	N	N	N	N	S	S
July	N	N	S	S	N	N	—	N	N	S	S	S
August	N	N	S	S	N	N	N	—	S	S	S	S
September	N	S	S	N	N	N	N	N	—	N	S	S
October	S	S	S	N	S	S	S	S	S	—	N	S
November	S	S	S	N	S	S	N	S	N	N	—	N
December	S	S	S	N	S	S	S	S	S	N	N	—

The upper triangle is for the LVG monthly coefficients (β_g) and the lower triangle is for the TEOM monthly coefficients (β_h)

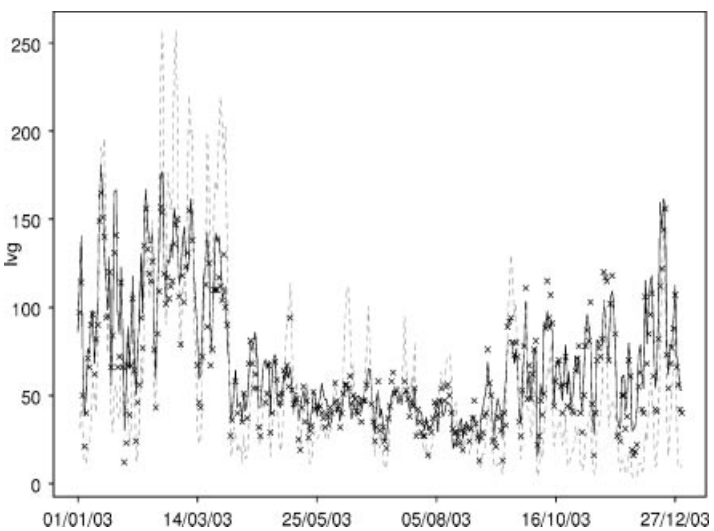

Figure 5. Calibration plot for the site Consolata. The observations are the crosses in the graph; the predictions are plotted as solid lines; the predictions from the Fassò and Nicolis model are plotted as dashed lines

values, labelled in the plots, are all greater than 90% and the coverage for the 1444 data points is 93.24%. Thus the proposed model performs very well for out of sample predictions both for the LVG and TEOM values. These plots also confirm the validity of the separable and exponential correlation function assumptions made in Section 3 since departure from those assumptions would have led to considerable poorer performance in model fit and validation.

The adopted model is now used to perform predictive inference. The maps of predicted annual average and the predicted number of days the daily average exceeding 50 are shown in panel (a) of Figures 7 and 8, respectively. The two maps agree qualitatively that on average the north-west part of

DOI: 10.1002/env

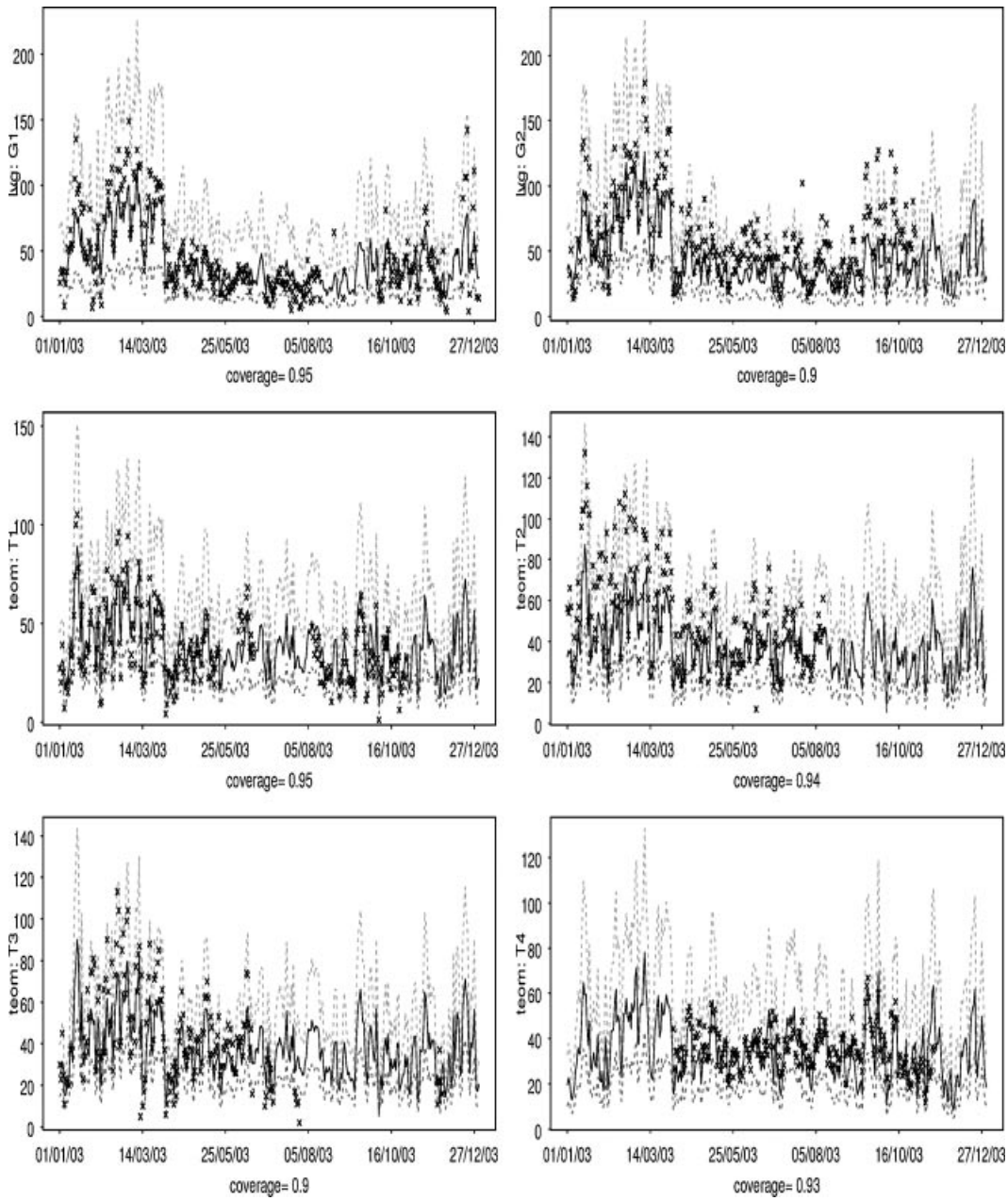

Figure 6. Validation plots for the six reserved sites. The observations are the crosses in the graph; the predictions are plotted as solid lines and the 95% prediction intervals are plotted as dotted lines

DOI: 10.1002/env

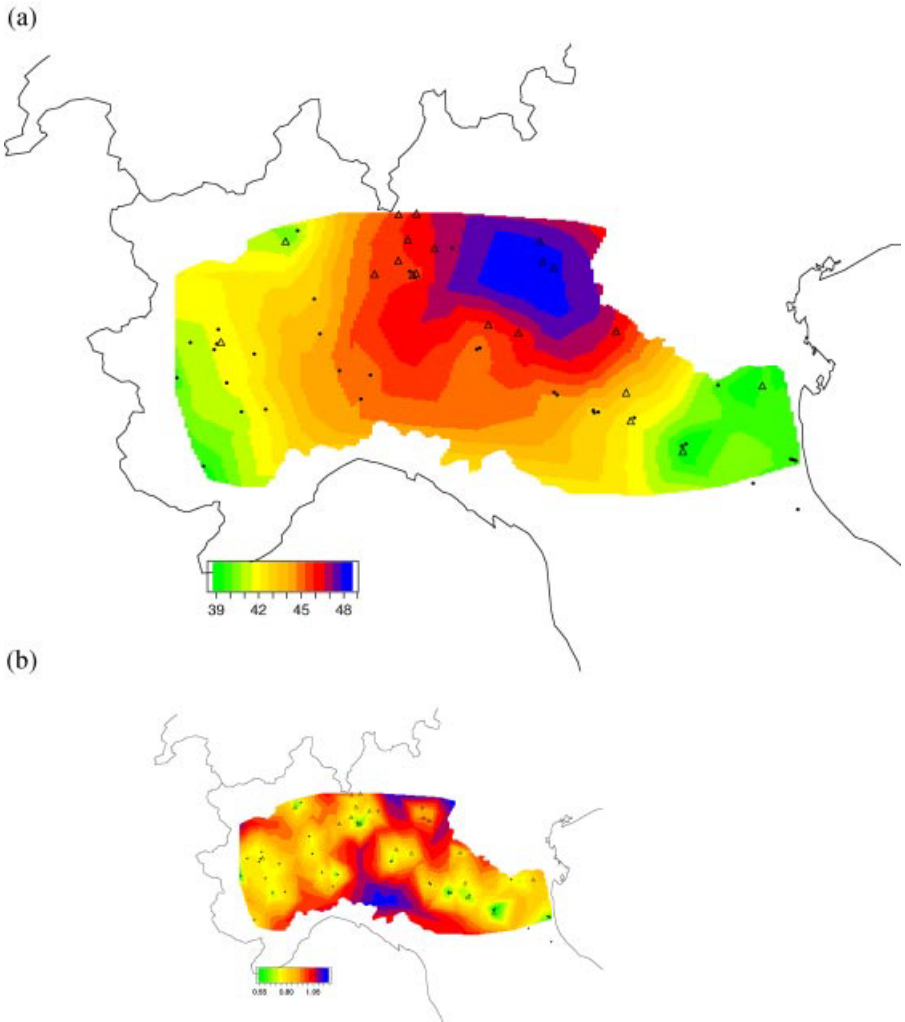

Figure 7. (a) Map of annual predictions for 2003; points denote the LVG sites and the △ denotes the TEOM sites (b) map of standard deviations of the annual predictions. This figure is available in colour online at www.interscience.wiley.com/journal/env

the study region has higher LVG levels than the other regions. This is likely since the model showed very good performance in validation at the three sites T1, T2 and T3 located in this region. In addition, we have calculated the root mean square error between the observed annual summaries at the 34 LVG stations and the predictions in the corresponding nearest sites to be 8.1. That is, the predictions and the data values differ by only 8.1 on the average. Thus, there is very good agreement between the annual summaries and the model-based predictions. (However, we do recognise that due to the presence of missing data it is not possible to compare the model-based estimates with exact observed annual values.) The standard deviation map of the annual predictions are shown in panel (b) of Figures 7 and 8. As

DOI: 10.1002/env

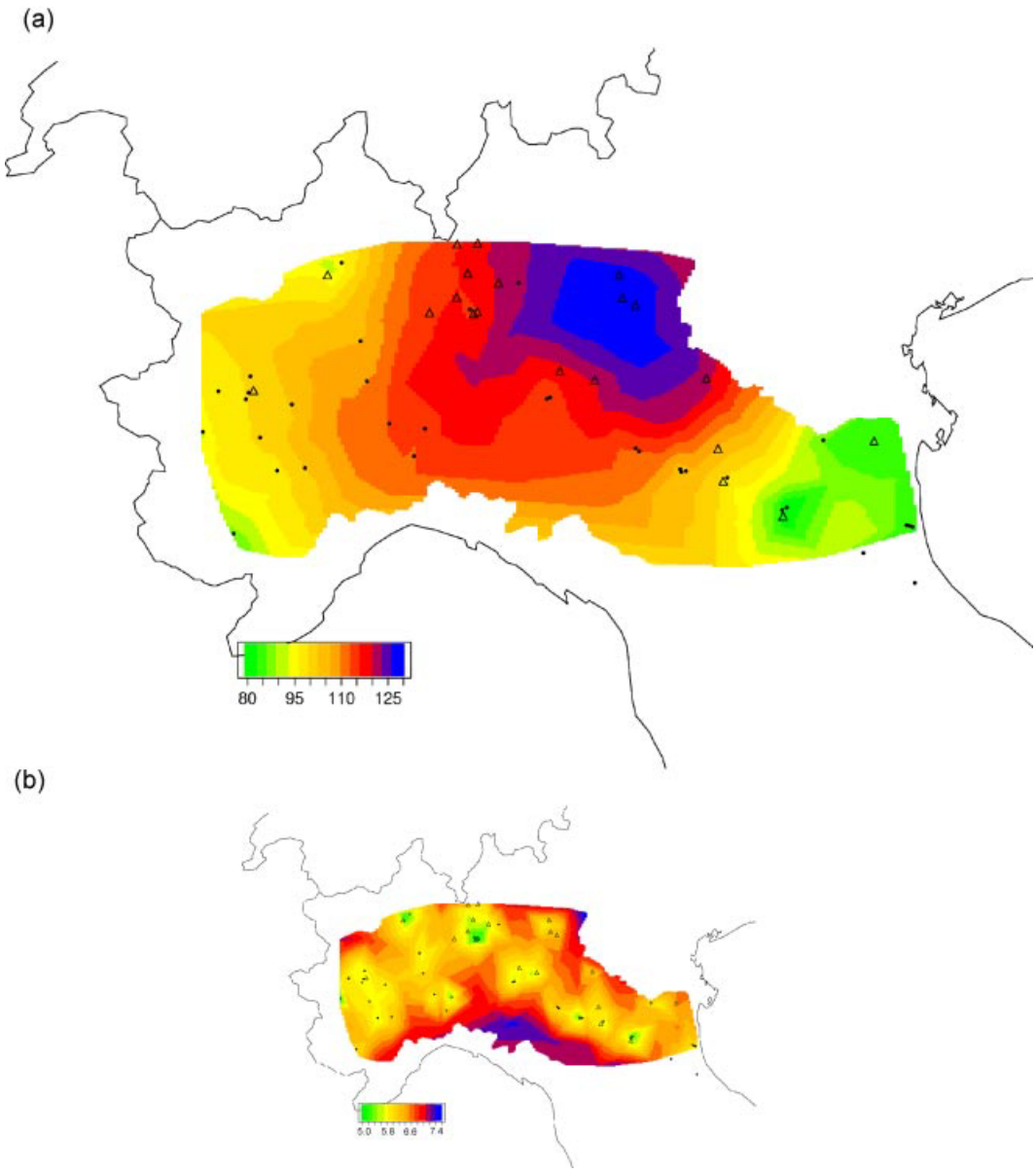

Figure 8. (a) Map of the predicted number of days in the year 2003 when LVG is greater than 50; points denote the LVG sites and the △ denotes the TEOM sites. (b) Map of standard deviations of the predictions plotted in (a). This figure is available in colour online at www.interscience.wiley.com/journal/env

DOI: 10.1002/env

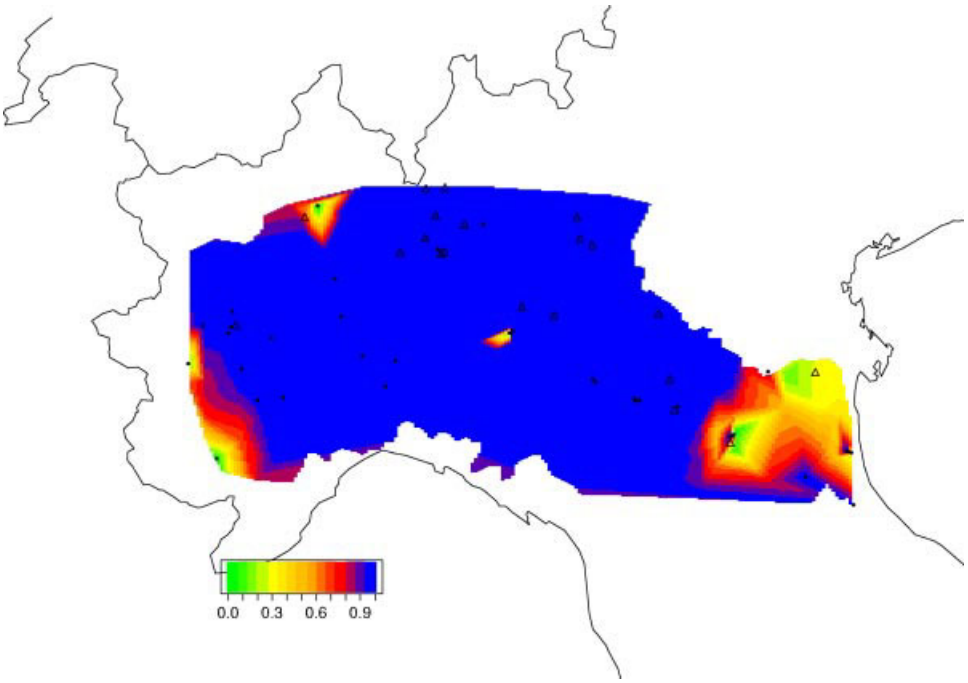

Figure 9. The estimated probability that the annual average for 2003 is greater than 40; points denote the LVG sites and the △ denotes the TEOM sites. This figure is available in colour online at www.interscience.wiley.com/journal/env

expected standard deviations are smaller near the observation sites. It is clear that the annual average in 2003 exceeded 40 in almost all areas and the number of days for which the daily average exceeded 50 was a great deal more than 7. In Figure 9, we plot the estimated probability that the annual average is greater than 40. As expected, we see that the probability that the annual average exceed 40 is more than 0.90 for most areas.

6. CONCLUSION

We have introduced a spatio temporal model for predicting daily PM_{10} concentrations in North Italy and calibrating the surrogate TEOM measurements. From the results, it emerges that seasonal components have a strong effect both on the LVG and TEOM concentrations. The common spatio temporal component is able to improve the forecasting methods and to adjust the TEOM values. Our model and the associated assumptions of separability, stationarity and exponential covariance structure for the latent space–time process have been validated by a large number of out of sample predictions at the highest (daily) level of temporal resolution.

The proposed high-resolution space–time model enables us to infer about temporal and spatial summaries of PM_{10} at any desired spatial and temporal resolution. In this paper, we have developed methods for predicting the annual summaries which are of interest to regulatory bodies. In addition, we are able to attach uncertainty to all of these predictions, derived from the model fitting. We are also

DOI: 10.1002/env

able to demonstrate the benefit of fitting models when predicting the annual averages as opposed to interpolating the observations themselves. Our results show that the annual summaries for 2003 do not comply with the European regulations currently in force.

Further work will seek to introduce some exogenous variables in order to improve the mean function. Further information about the monitoring sites, such as urban/rural levels, the population density and the elevation information may improve the predictions.

ACKNOWLEDGEMENTS

This work was supported partially by PRIN 2004 grant. The authors are also grateful to Dott.ssa R. Ignaccolo, Dott. F. Greco and Ing. V. Gianelle for the Po Valley data and to Professor A. Fassò for useful comments.

REFERENCES

Cressie N, Kaiser MS, Daniels MJ, Aldworth J, Lee J, Lahiri SN, Cox L. 1999. Spatial analysis of particulate matter in an urban environment. In *GeoEnv II: Geostatistics for Environmental Applications*, Gmez-Hernndez J, Soares A, Froidevaux R (eds). Kluwer: Dordrecht; 41–52.

Fuentes M. 2006. Testing for separability of spatial-temporal covariance functions. *Journal of Statistical Planning and Inference* **136**: 447–466.

Fassò A, Cameletti M, Nicolis O. 2007. Air quality monitoring using heterogeneous networks. *Environmetrics* **18**: 245–264.

Fassò A, Nicolis O. 2005. Space-time integration of heterogeneous networks in air quality monitoring. *Atti SIS*, 21–23 September, Messina (Italy).

Hauck H, Berner A, Gomiscek B, Stopper S, Puxbaum H, Kundi M, Preining O. 2004. On the equivalence of gravimetric PM data with TEOM and beta-attenuation measurements. *Journal of Aerosol Science* **35**: 1135–1149. doi:10.1016/j/jaerosci.2004.04.004

Kibria B, Golam M, Sun L, Zidek JV, Le ND. 2002. Bayesian spatial prediction of random space-time fields with application to mapping $PM_{2.5}$ exposure. *Journal of the American Statistical Association* **97**: 112–124.

Mardia KV, Goodall C. 1993. Spatial-temporal analysis of multivariate environmental monitoring data. In *Multivariate Environmental Statistics*, Patil GP, Rao CR (eds). Amsterdam: Elsevier; 347–386.

Mardia KV, Goodall C, Redfern EJ, Alonso FJ. 1998. The Kriged Kalman filter (with discussion). *Test* **7**: 217–252.

McBride SJ, Clyde MA. 2003. Hierarchical Bayesian calibration with reference priors: an application to airborne particulate matter monitoring data. *Working paper number 03/23*, Department of Statistical Science, Duke University.

Mitchell MW, Genton MG, Gumpertz ML. 2005. Testing for separability of space-time covariances. *Environmetrics* **16**: 819–831.

Sahu SK, Gelfand AE, Holland DM. 2006. Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural Biological and Environmental Statistics* **11**: 61–86.

Sahu SK, Gelfand AE, Holland DM. 2007. High resolution space-time ozone modeling for assessing trends. *Journal of the American Statistical Association* **102**: 1221–1234.

Sahu SK, Mardia KV. 2005. A Bayesian Kriged-Kalman model for short-term forecasting of air pollution levels. *Journal of the Royal Statistical Society, Series C* **54**: 223–244.

Shaddick G, Wakefield J. 2002. Modelling daily multivariate pollutant data at multiple sites. *Journal of the Royal Statistical Society, Series C* **51**: 351–372.

Smith RL, Kolenikov S, Cox LH. 2003. Spatio-temporal modelling of $PM_{2.5}$ data with missing values. *Journal of Geophysical Research-Atmospheres* **108**(D24): 9004, doi:10.1029/2002JD002914.

Stein M. 1999. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Verlag, New York.

Sun L, Zidek JV, Le ND, Ozkaynak H. 2000. Interpolating Vancouver's daily ambient PM10 field. *Environmetrics* **11**: 651–663.

Wikle CK, Cressie N. 1999. A dimension-reduced approach to space-time Kalman filtering. *Biometrika* **86**: 815–829.

Zhang H. 2004. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association* **99**: 250–261.

Zidek JV, Sun L, Le N, Ozkaynak H. 2002. Contending with space-time interaction in the spatial prediction of pollution: Vancouver's hourly ambient PM_{10} field. *Environmetrics* **13**: 595–613.

DOI: 10.1002/env

APPENDIX : CONDITIONAL DISTRIBUTIONS

Straightforward calculation yields the following complete conditional distributions:

$$\frac{1}{\sigma_g^2} \sim G\left(\frac{n_1 T}{2} + a, b + \frac{1}{2}\sum_{i=1}^{n_1}\sum_{t=1}^{T}\left\{z_g(\mathrm{s}_i, t) - \mu_g(\mathrm{s}_i, t) - u(\mathrm{s}_i, t)\right\}^2\right)$$

$$\frac{1}{\sigma_h^2} \sim G\left(\frac{n_2 T}{2} + a, b + \frac{1}{2}\sum_{i=n_1}^{n}\sum_{t=1}^{T}\left\{z_h(\mathrm{s}_i, t) - \mu_h(\mathrm{s}_i, t) - u(\mathrm{s}_i, t)\right\}^2\right)$$

$$\frac{1}{\sigma_u^2} \sim G\left(\frac{nT}{2} + a, b + \mathbf{u}'(\Sigma_s^{-1} \otimes \Sigma_t^{-1})\mathbf{u}\right)$$

We sample $\boldsymbol{\beta}_g$ and $\boldsymbol{\beta}_h$ en-bloc. Define the vectors $\mathbf{z}_g = (z_g(\mathrm{s}_1, 1), z_g(\mathrm{s}_1, 2), \ldots, z_g(\mathrm{s}_{n_1}, T))'$ and $\mathbf{z}_h = (z_h(\mathrm{s}_{n_1+1}, 1), z_h(\mathrm{s}_{n_1+1}, 2), \ldots, z_h(\mathrm{s}_n, T))'$. Similarly we form the matrices X_g and X_h of order $n_1 T \times p$ and $n_2 T \times p$, respectively using the covariate dummies $\mathbf{x}_t$ appropriately. We also define $\mathbf{u}_g = (\mathbf{u}_1', \ldots, \mathbf{u}_{n_1}')'$ and $\mathbf{u}_h = (\mathbf{u}_{n_1}', \ldots, \mathbf{u}_n')'$. Now the conditional distribution of $\boldsymbol{\beta}_g$ is normal with mean $\Lambda\boldsymbol{\zeta}$ and covariance Λ where

$$\Lambda = \left(\frac{1}{\sigma_g^2} X_g' X_g + A^{-2} I\right)^{-1}, \text{ and } \boldsymbol{\zeta} = \frac{1}{\sigma_g^2} X_g'(\mathbf{z}_g - \mathbf{u}_g)$$

The conditional distribution of $\boldsymbol{\beta}_h$ is normal with mean $\Lambda\boldsymbol{\zeta}$ and covariance Λ where

$$\Lambda = \left(\frac{1}{\sigma_h^2} X_h' X_h + A^{-2} I\right)^{-1}, \text{ and } \boldsymbol{\zeta} = \frac{1}{\sigma_h^2} X_h'(\mathbf{z}_h - \mathbf{u}_h)$$

We sample the *spatio*-temporal process $u(\mathrm{s}_i, t)$ en-bloc as follows. The prior complete conditional distribution of $\mathbf{u}_j$ for $j = 1, \ldots, n$ given all other columns $i \neq j, i = 1, \ldots, n$ is normal with mean $\boldsymbol{\zeta}_j$ and covariance Λ_j where

$$\boldsymbol{\zeta}_j = -\sum_{i \neq j, i=1}^{n} \frac{(\Sigma_s)_{ij}^{-1}}{(\Sigma_s)_{jj}^{-1}} \mathbf{u}_i, \text{ and } \Lambda_j = \sigma_u^2 \frac{1}{(\Sigma_s)_{jj}^{-1}} \Sigma_t$$

The likelihood contribution for $\mathbf{u}_j$ is also normal with mean $\boldsymbol{\xi}_j$ and covariance χ_j where for $j = 1, \ldots, n_1$, $\boldsymbol{\xi}_j = (z_g(\mathrm{s}_j, 1) - \mu_g(\mathrm{s}_j, 1), \ldots, , z_g(\mathrm{s}_j, T) - \mu_g(\mathrm{s}_j, T))'$, $\chi_j = \sigma_g^2 I$, and for $j = n_1 + 1, \ldots, n$, $\boldsymbol{\xi}_j = (z_h(\mathrm{s}_j, 1) - \mu_h(\mathrm{s}_j, 1), \ldots, z_h(\mathrm{s}_j, T) - \mu_h(\mathrm{s}_j, T))'$, $\chi_j = \sigma_h^2 I$. The posterior complete conditional distribution is now seen to be normal with

$$\text{mean} = \Omega_j\left(\chi_j^{-1}\boldsymbol{\xi}_j + \Lambda_j^{-1}\boldsymbol{\zeta}_j\right) \text{ and covariance } \Omega_j = \left(\chi_j^{-1} + \Lambda_j^{-1}\right)^{-1}$$

DOI: 10.1002/env

To derive the distribution $\pi(u(\mathrm{s}', t')|\mathbf{U}, \sigma_u^2)$

$$\begin{pmatrix} u(\mathrm{s}', t') \\ \mathbf{u} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \sigma_u^2 \begin{pmatrix} 1 & \Sigma_s'(\mathrm{s} - \mathrm{s}') \otimes \Sigma_t'(\mathbf{t} - t') \\ \Sigma_s(\mathrm{s} - \mathrm{s}') \otimes \Sigma_t(\mathbf{t} - t') & \Sigma_s \otimes \Sigma_t \end{pmatrix}\right]$$

where $\Sigma_s(\mathrm{s} - \mathrm{s}')$ is an $n \times 1$ column vector with the ith entry given by $\sigma(\mathrm{s}_i - \mathrm{s}') = \rho_s(\mathrm{s}_i - \mathrm{s}'; \phi_s)$ and $\Sigma_t(\mathbf{t} - t')$ is a $T \times 1$ column vector with the kth entry given by $\sigma(k - t') = \rho_t(k - t'; \phi_t)$. Hence

$$u(\mathrm{s}', t')|\mathbf{U} \sim N\left(\sum_{j=1}^{n}\sum_{k=1}^{T} b_{jk}(\mathrm{s}', t')u(\mathrm{s}_j, k), \sigma_u^2 C(\mathrm{s}', t')\right)$$

where

$$b_{jk}(\mathrm{s}', t') = \sum_{i=1}^{n}\sum_{l=1}^{T} \sigma(\mathrm{s}_i - \mathrm{s}')\sigma(l - t')(\Sigma_s)_{ij}^{-1}(\Sigma_t)_{lk}^{-1}$$

and

$$C(\mathrm{s}', t') = 1 - \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{T}\sum_{k=1}^{T} \sigma(\mathrm{s}_i - \mathrm{s}')\sigma(l - t')(\Sigma_s)_{ij}^{-1}(\Sigma_t)_{lk}^{-1}\sigma(\mathrm{s}_j - \mathrm{s}')\sigma(k - t')$$

DOI: 10.1002/env