Biostatistics (2016), **0**, 0, *pp.* 1–15
doi:10.1093/biostatistics/Leebiostatssupp

Supplementary material to *A rigorous statistical framework for spatio-temporal pollution prediction and estimation of its long-term impact on health*

DUNCAN LEE[1*], SABYASACHI MUKHOPADHYAY[2], ALASTAIR RUSHWORTH[3],

SUJIT K. SAHU[2]

[1] *School of Mathematics and Statistics, University of Glasgow, UK,*
[2] *Mathematical Sciences, University of Southampton, UK.*

[3] *Department of Mathematics and Statistics, University of Strathclyde, UK.*

Duncan.Lee@glasgow.ac.uk

1. Introduction

This supplementary material accompanies the main paper and contains the following content. Section 2 provides additional exploratory data analysis for the pollution data used in the England study. Section 3 provides brief details of the implementation of the pollution model, while Section 4 provides similar details for the disease model. Finally, Sections 5 and 6 provide additional results from the pollution and disease modelling respectively.

2. Additional exploratory analysis of the pollution data

As discussed in the main paper, pollution data from a monitoring network are often prone to large numbers of missing values, due to factors such as instrument malfunction, discontinuation

[*] To whom correspondence should be addressed.

of some sites, introduction of new sites during the study period, or the fact that not all sites monitor all pollutants. This missingness, broadly defined, for the AURN network is summarised in Table 1 below, which shows largely constant levels of missingness over time except for $PM_{2.5}$ which decreases after 2008 due to more sites measuring this pollutant. This increase in sites may be due to the introduction of the EU Directive on Ambient Air Quality and Cleaner Air for Europe (2008/50/EC) in 2008, which included a new focus on $PM_{2.5}$ due to a recognition of a lack of evidence about this pollutant.

Numerical and graphical summaries of the observed pollution data are given by Table 2 and Figures 1 and 2, which respectively present the data by either site type (16 Rural, 80 Urban and 46 RKS) or year. The site type figure shows greater average concentrations and greater levels of variation for RKS sites for NO_2, $PM_{2.5}$ and PM_{10}, where as the converse is true for O_3 in terms of mean concentrations. The concentrations of all four pollutants show little variation by year, with no discernible changes in average concentrations or the levels of variation. Finally, Figure 3 displays scatter plots of the measured against the bilinearly interpolated AQUM modelled pollution concentrations for each month and site, and shows moderate correlations ranging between 0.37 and 0.69 depending on the pollutant. The figure also shows clear bias in the AQUM output, as the modelled concentrations are almost always lower than the measured concentrations. This is not surprising for RKS sites, as the AQUM outputs are modelled background concentrations, but is more surprising for the Urban and Rural sites.

3. Implementation of the pollution model

The details for fitting the independence and GP models are provided in Bakar and Sahu (2015). Following are the details for implementing the space-time non-stationary model (3.3) and (3.4) discussed in the main paper. The logarithm of the full posterior distribution is given by:

$$\begin{array}{rcl}\log\left(\pi\left(\mathbf{S}_m^*, \boldsymbol{\eta}, \boldsymbol{\theta}|\mathbf{z}\right)\right) & \propto & -\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^n \sum_{t=1}^T \left(z(\mathbf{s}_i,t) - \mu(\mathbf{s}_i,t) - \tilde{\eta}(\mathbf{s}_i,t)\right)^2 \\ & - & \frac{nT}{2}\log(\sigma_\epsilon^2) + \sum_{j=1}^m \log(\pi(\mathbf{s}_j^*)) + \log\left(\pi(\boldsymbol{\eta}_0^*)\right) \\ & - & \frac{mT}{2}\log(\sigma_\eta^2) - \frac{T}{2}\log\left|\mathbf{H}_{\eta^*}(\phi,\nu)\right| \\ & - & \frac{1}{2\sigma_\eta^2}\sum_{t=1}^T \left(\boldsymbol{\eta}_t^* - \varrho\boldsymbol{\eta}_{t-1}^*\right)^\top \mathbf{H}_{\eta^*}(\phi,\nu)^{-1}\left(\boldsymbol{\eta}_t^* - \varrho\boldsymbol{\eta}_{t-1}^*\right) \\ & + & \log(\pi(\boldsymbol{\theta})),\end{array}$$

where $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \varrho, \sigma_\epsilon^2, \sigma_\eta^2, \phi, \nu)^T$ denotes the model parameters and $\pi(\boldsymbol{\theta})$ thus denotes its joint prior distribution. Inference uses Markov chain Monte Carlo (MCMC) simulation, via a combination of Gibbs sampling and Metropolis-Hastings steps. Our implementation follows that described in Section 3.2 of Guhaniyogi *and others* (2011), except that we take a different approach to updating the knot-locations $\mathbf{S}_m^*$. We adopt the method developed by Sahu and Mukhopadhyay (2015), where we simulate m proposed knots from the prior outlined in the main paper without replacement, and then use a Metropolis-Hastings step to accept or reject the proposed knots. The starting configuration of the knots is taken to be according to a space filling design, and we tune the algorithm to have 15-40% acceptance rate as is common in practice.

4. Implementation of the disease model

The major challenge in implementing the disease model defined by (3.7), (3.8) and (3.12) is computational, because there are 19,380 spatio-temporal observations (for example the study by Elliott *and others*, 2007 used around 800 data points, while the largest study region used in Lee *and others*, 2009 was around 1500). The random effects $\boldsymbol{\psi}$ with the GMRF prior are updated in the MCMC routine using computationally efficient algorithms written in C++, which utilise the sparsity of the neighbourhood matrix $\mathbf{W}$ via its triplet form representation. The elements of $\mathbf{W}$ are proposed in a computationally efficient manner by updating $\boldsymbol{g}^+$ in blocks of size 10. In the acceptance ratio for each block proposal, a new determinant, $\det(\boldsymbol{Q}(\boldsymbol{W},\rho))$, is required, which was calculated efficiently by updating the (sparse) Cholesky decomposition, $\boldsymbol{L}$, so that $\boldsymbol{L}\boldsymbol{L}^\top = \boldsymbol{Q}(\boldsymbol{W},\rho)$, and as a result $\det(\boldsymbol{Q}(\boldsymbol{W},\rho)) = \det(\boldsymbol{L})^2 = \prod_{i=1}^K L_{ii}^2$, since $\mathbf{L}$ is triangular.

The health model, without the ecological bias correction or the uncertainty propagation, can be implemented in the R package CARBayesST, which can also implement similar models with binomial and Gaussian likelihoods.

The other major computational challenge is updating the set of spatio-temporal pollution concentrations Z_{kt} from (3.11), in the model where pollution is treated as unknown. These concentrations should not be updated independently for each areal unit and time period, because the pollution model from stage one produces posterior predictive realisations of the spatio-temporal pollution surface that are correlated in space and time. However, randomly selecting one of the L posterior predictive samples (for all 19,380 data points) at random as a proposal, and then accepting or rejecting it via a Metropolis-Hastings step is not feasible due to the curse of dimensionality, which results in very poor acceptance rates. Therefore we propose a multivariate Gaussian approximation to the posterior predictive distribution for each time period separately, that is:

$$\pi(Z_{1t},\ldots,Z_{Kt}|\mathbf{z})^\top = \mathrm{N}\left(\hat{\mathbf{z}}_t, \hat{\mathbf{\Sigma}}_t\right) \qquad \text{for} \quad t=1,\ldots,T,$$

where the kth element of $\hat{\mathbf{z}}_t$ is given by $\hat{z}_{kt} = \frac{1}{Ln_k}\sum_{j=1}^{n_k}\sum_{\ell=1}^{L} z^{(\ell)}(\mathbf{v}_{kj},t)$, while the k and k'th element of $\hat{\mathbf{\Sigma}}_t$ is given by:

$$\left(\hat{\mathbf{\Sigma}}_t\right)_{kk'} = \frac{1}{L-1}\sum_{\ell=1}^{L}(\hat{z}_{kt}^{(\ell)} - \hat{z}_{kt})(\hat{z}_{k't}^{(\ell)} - \hat{z}_{k't}).$$

While this specification disregards the temporal dependence between exposures that occur for a single region, the spatial dependence structure is largely preserved. The Gaussian approximation allows the univariate prior conditional distributions, $\pi(Z_{kt}|\mathbf{Z}_{-kt})$ for $k=1,\ldots,K$ (where $\mathbf{Z}_{-kt}$ denotes the vector with the kth element removed) to be easily computed, making Metropolis-Hastings updating one element at a time straightforward. We note that a full spatio-temporal approximation could also be used, but that this would require a $KT \times KT$ covariance matrix to be constructed, from which it would be computationally demanding to calculate conditional distributions.

5. Additional pollution modelling results

Scatter plots of the predictions from the *GPP* model against the observed values are displayed in Figures 4 and 5 for all four pollutants, the latter being on the square root scale on which the data are modelled. The figures show generally good agreement between the predicted and observed values, with most lying close to the line of equality. The exceptions are that the models cannot predict the very low O_3 and the very high NO_2 concentrations that well, which is evident from the non-linearities in the plots.

Figure 6 displays a scatter plot of the widths of the 95% prediction intervals for the GP and GPP models for all four pollutants. The figure clearly shows that the intervals from the GP models are almost always much wider than those from the GPP models, with the exceptions for NO_2 being mainly 3 sites (60 months of predictions each) that are near the boundary of the study region and thus far away from the predictive knot locations. The very large interval widths for the GP model is why the former have a near 100% coverage. This is because its spatio-temporal process $\eta(\mathbf{s}_i, t)$ is overly flexible in space and time with no temporal autocorrelation and a separate random effect for each spatial location. In contrast, the *GPP* model is autocorrelated in time and uses a reduced rank spatial predictive process, resulting in more borrowing of strength in the estimation and reduced uncertainty. The differences in widths are substantial, with for example the average widths for $PM_{2.5}$ being around 6 (GPP) and 45 (GP) respectively. Given the scale of the $PM_{2.5}$ data ranging from 2.66 to 36.45, a width of 45 is overly large and gives very little information on the true predictive uncertainty.

6. Additional disease modelling results

Here we present posterior inference for $\mathbf{w}^+$ from the localised smoothing model given by (3.12), specifically, $\varrho_{kj} = \mathbb{P}(w_{kj} < 0.5|\mathbf{Y})$, the posterior probability that each adjacency element w_{kj} is less than 0.5 which is the mid-point of the allowable unit interval. Recall that if $w_{kj} \in \mathbf{w}^+$ is estimated as zero then the random effects (ψ_{kt}, ψ_{jt}) for all times t are conditionally independent given all other random effects, while a w_{kj} value close to one suggests strong partial autocorrelation between the random effects. Thus using a cut-off value of 0.5 when computing ϱ_{kj} illustrates whether the balance of probability corresponds to partial autocorrelation or close to conditional independence. A histogram of the posterior probabilities $\{\varrho_{kj}\}$ over all 861 adjacency elements in $\mathbf{w}^+$ is displayed in Figure 7, and shows a bimodal distribution, with most values having either a very low or a very high posterior probability. To illustrate the locations of these localised conditional independences, Figure 8 displays the locations where $\varrho_{kj} > 0.99$, that is the posterior probability of w_{kj} being less than 0.5 is greater than 0.99. These locations are shown as white lines, and are superimposed on the estimated average (over the 60 months) random effects surface. The figure shows that the majority of white lines correspond to borders between areas with very different random effect values as expected. However, we note these are conditional rather than marginal independences, and thus an assumed conditional independence between two regions will also affect other neighbouring regions. The Figure contains 205 such highlighted borders corresponding to 23.8% of the total number of borders in the study region, which suggests the presence of widespread localised spatial autocorrelation in the random effects.

References

Bakar, K. Shuvo and Sahu, Sujit K. (2015). `spTimer`: Spatio-temporal bayesian modelling using `r`. *Journal of Statistical Software* **63**(15).

Elliott, P, Shaddick, G, Wakefield, J, Hoogh, C and Briggs, D. (2007). Long-term associations of outdoor air pollution with mortality in Great Britain. *Thorax* **62**, 1088–1094.

Guhaniyogi, Rajarshi, Finley, Andrew O., Banerjee, Sudipto and Gelfand, Alan E. (2011). Adaptive gaussian predictive process models for large spatial datasets. *Environmetrics* **22**(8), 997–1007.

Lee, D, Ferguson, C and Mitchell, R. (2009). Air pollution and health in Scotland: a multicity study. *Biostatistics* **10**, 409–423.

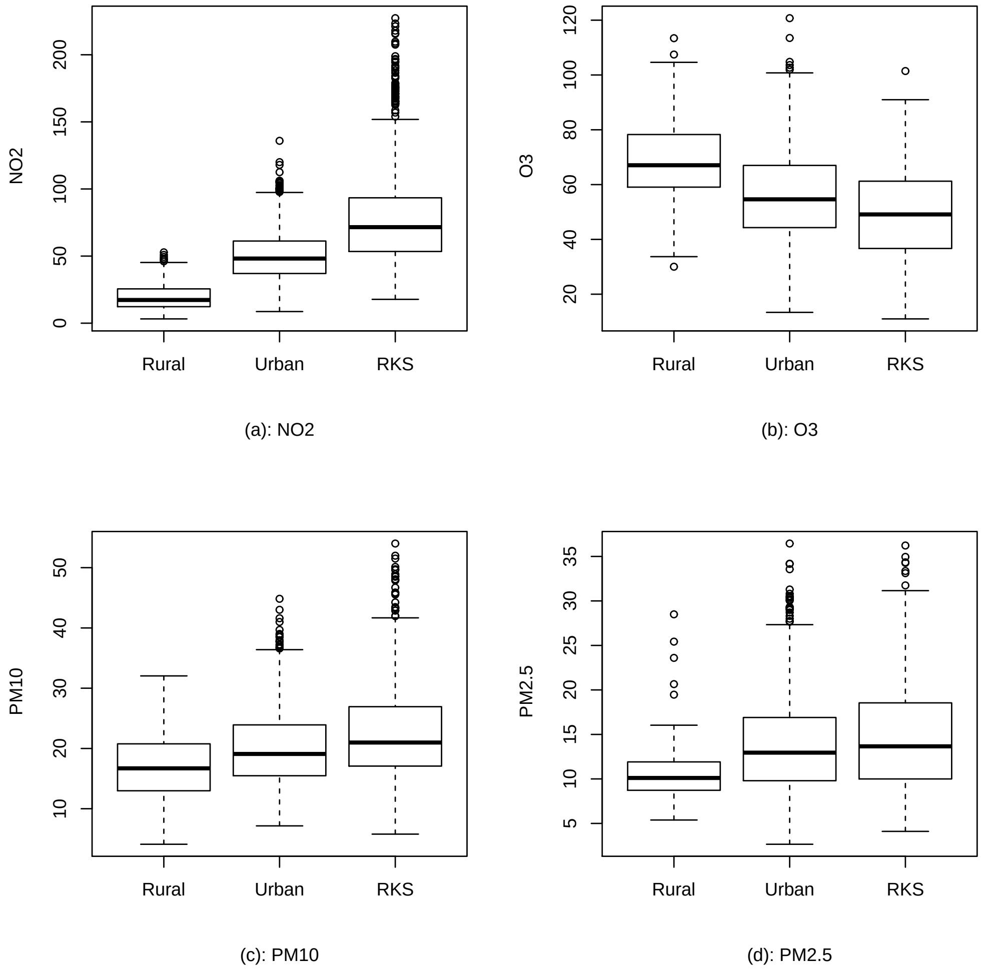
Sahu, S. K. and Mukhopadhyay, S. (2015). On generating a flexible class of anisotropic spatial models using gaussian predictive processes. *Technical Report, University of Southampton*.

	2007	**2008**	**2009**	**2010**	**2011**
NO_2	38.38	38.08	36.97	37.21	34.62
O_3	52.26	59.14	60.09	58.68	58.33
PM_{10}	64.96	66.26	69.37	71.48	69.13
$PM_{2.5}$	96.53	91.43	65.96	64.32	64.03

Table 1. Percentage of missing monthly observations at the 142 sites in each year for each of the four pollutants.

Type	**Min**	**Median**	**Mean**	**Max**	**Sd**	***N***
			NO_2			
Rural (16)	3.21	17.30	19.53	52.79	9.68	696
Urban (80)	8.67	48.18	49.81	135.90	17.92	2952
RKS (46)	17.77	71.55	76.54	227.30	33.27	1775
All (142)	3.21	50.46	54.67	227.30	29.61	5423
			O_3			
Rural (16)	30.03	67.06	68.60	113.40	13.29	788
Urban (80)	13.38	54.62	56.00	120.70	16.01	2523
RKS (46)	10.99	49.12	48.58	101.40	17.64	322
All (142)	10.99	57.72	58.08	120.70	16.69	3633
			PM_{10}			
Rural (16)	4.10	16.70	16.82	32.04	5.68	153
Urban (80)	7.16	19.09	20.03	44.84	6.24	1623
RKS (46)	5.79	20.99	22.74	54.01	7.92	988
All (142)	4.10	19.62	20.82	54.01	7.05	2764
			$PM_{2.5}$			
Rural (16)	5.38	10.11	10.72	28.49	3.64	110
Urban (80)	2.66	12.95	13.72	36.45	5.27	1321
RKS (46)	4.11	13.66	14.71	36.22	6.08	620
All (142)	2.66	12.87	13.86	36.45	5.53	2051

Table 2. Summary of the monthly pollution data for the four pollutants from the 16 Rural, 80 Urban and 46 RKS sites over the 5 years. All pollutants are measured in $\mu g m^{-3}$. Here Sd stands for standard deviation and N is the number of available monthly averages on which the summaries have been calculated.



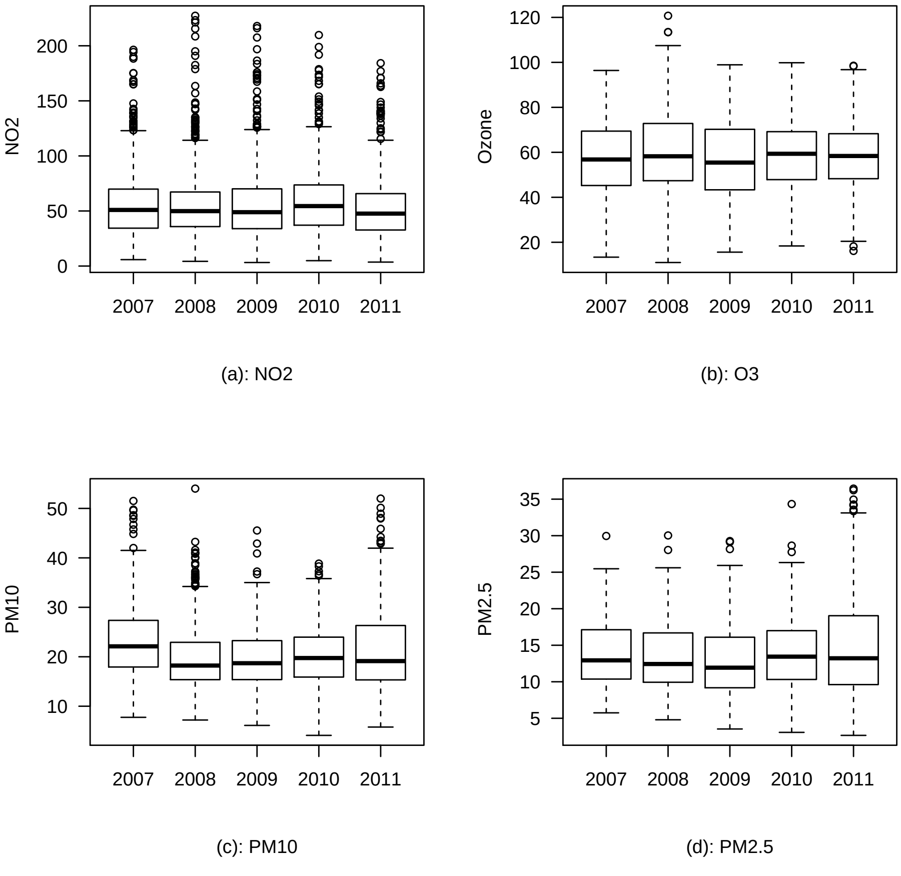
(a): NO2

(b): O3

(c): PM10

(d): PM2.5

Fig. 1. Boxplots of the monthly average concentrations for each pollutant by site type. The whiskers extend to the most extreme data point that is no more than 1.5 times the interquartile range away from the box.



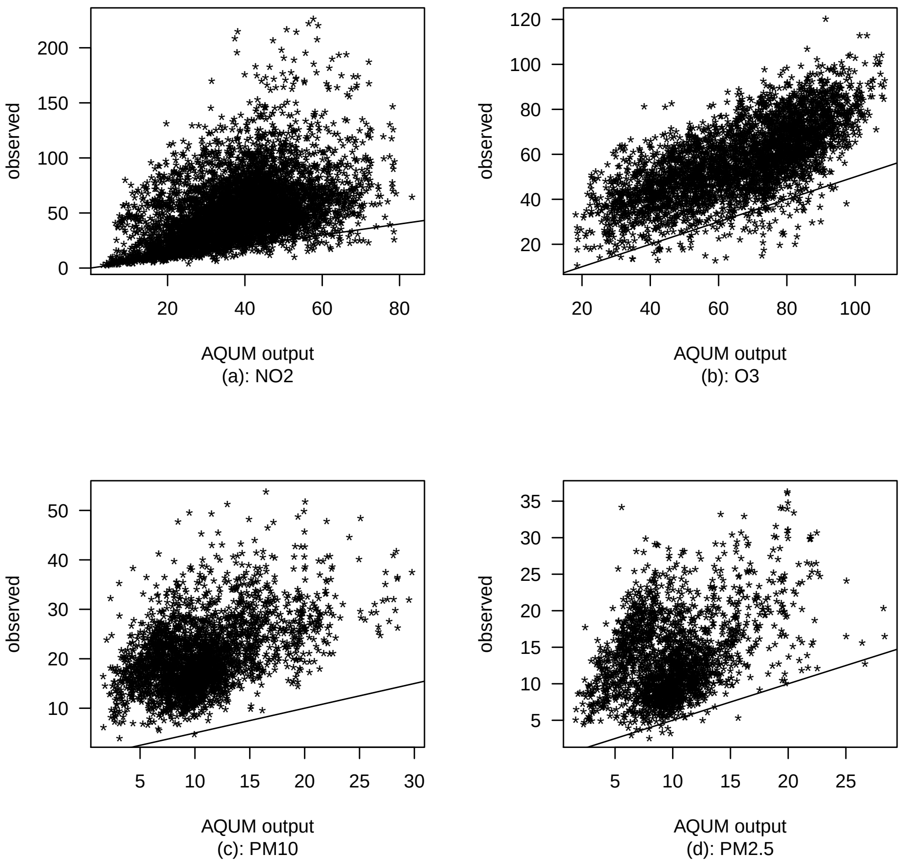
(a): NO2

(b): O3

(c): PM10

(d): PM2.5

Fig. 2. Boxplots of the monthly average concentrations for each pollutant by year. The whiskers extend to the most extreme data point that is no more than 1.5 times the interquartile range away from the box.



(a): NO2

(b): O3

(c): PM10

(d): PM2.5

Fig. 3. Scatter plot of the observed concentrations against the bilinearly interpolated AQUM model output for each of the four pollutants. The line $y = x$ is superimposed.

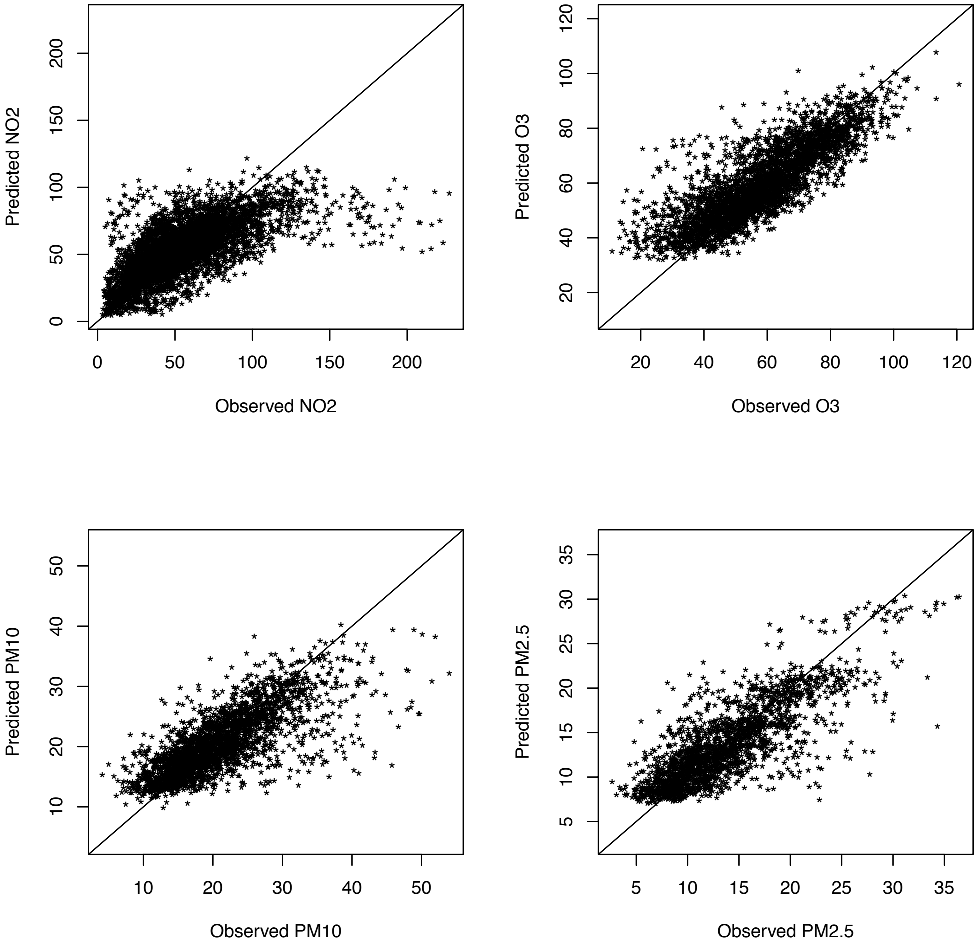


Fig. 4. Scatterplot of the predicted concentrations against the observed concentrations on the original scale for all four pollutants. The line $y = x$ is superimposed.

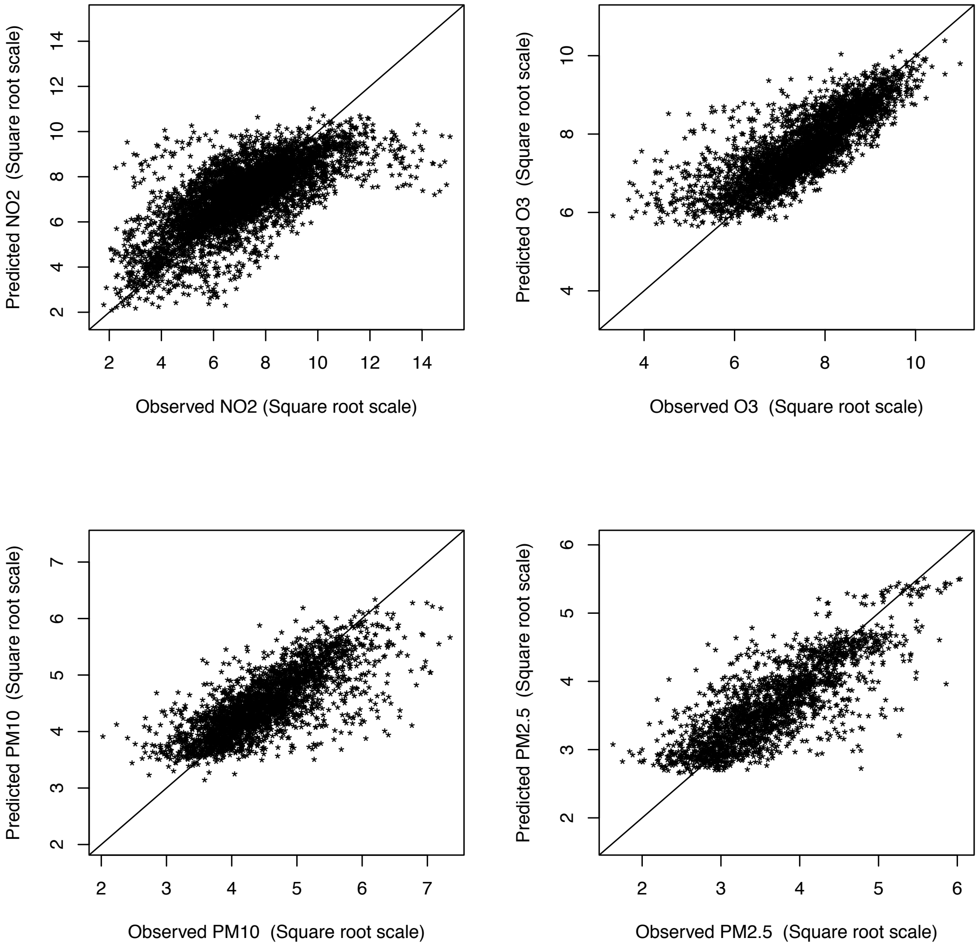


Fig. 5. Scatterplot of the predicted concentrations against the observed concentrations on the modelled square-root scale for all four pollutants. The line $y = x$ is superimposed.

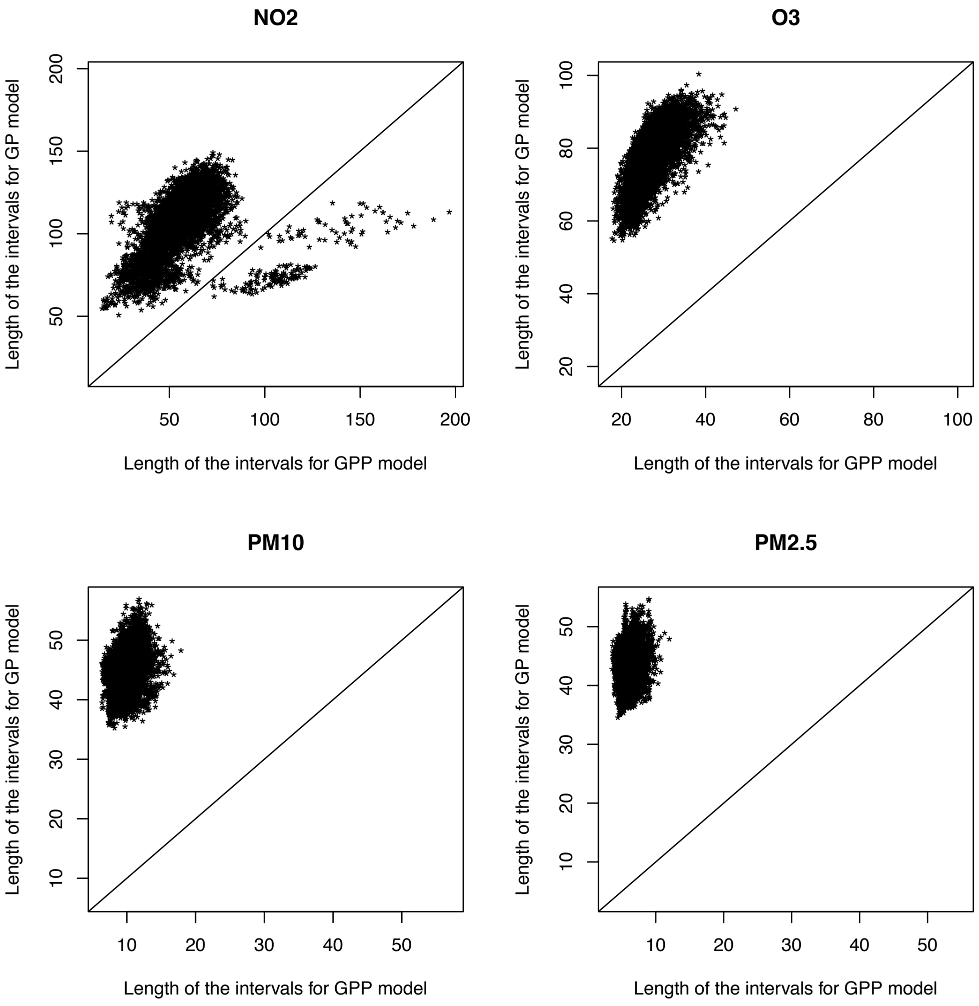


Fig. 6. Scatterplot of the widths of the 95% prediction intervals for the GP and GPP models for all four pollutants. The line $y = x$ is superimposed.

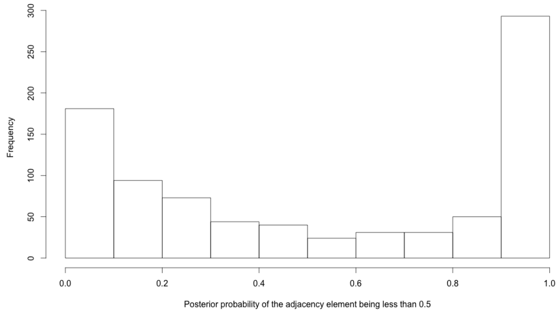


Fig. 7. Histogram of the posterior probabilities of $\{w_{jk}\}$ being less than 0.5.

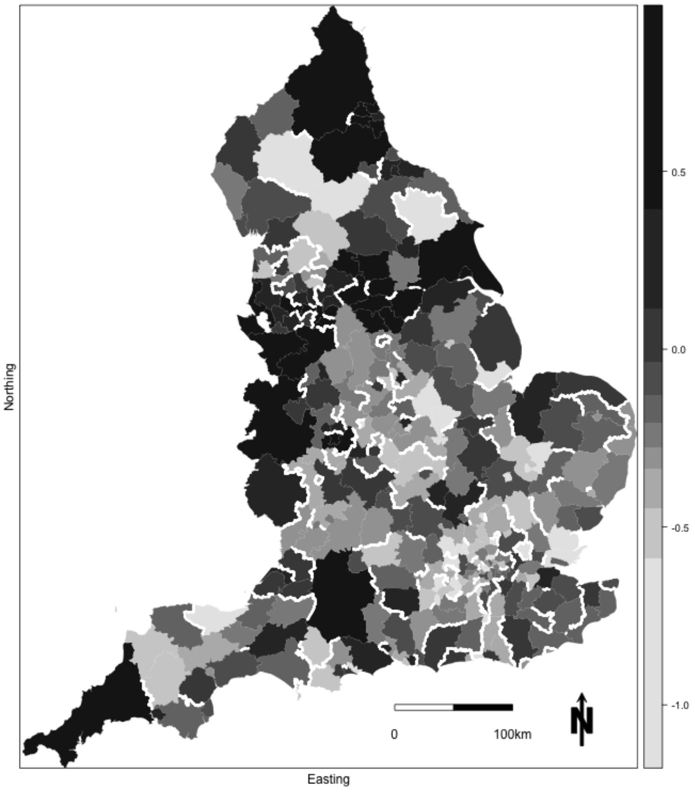


Fig. 8. Map of the posterior median random effects surface averaged over all 60 time periods. The white lines depict borders between areal units (k, j) where $\varrho_{kj} > 0.99$.