

Web Working Papers
by
The Italian Group of Environmental Statistics



Gruppo di Ricerca per le Applicazioni della Statistica
ai Problemi Ambientali

www.graspa.org

Comparing air quality statistical models

Michela Cameletti, Rosaria Ignaccolo, Stefano Bande

GRASPA Working paper n.40, November 2010

Comparing air quality statistical models

Michela Cameletti^a, Rosaria Ignaccolo^b,
Stefano Bande^c

^aDip. di Matematica, Statistica, Informatica e Applicazioni,
Università di Bergamo^{*}

^bDip. di Statistica e Matematica Applicata, Università di Torino[†]

^cDip. Tematico Sistemi Previsionali, Qualità dell'aria, ARPA Piemonte

Abstract

Air pollution is a great concern because of its impact on human health and on the environment. Statistical models play an important role in improving knowledge of this complex spatio-temporal phenomenon and in supporting public agencies and policy makers. We focus on the class of hierarchical models that provides a flexible framework for incorporating spatio-temporal interactions at different hierarchical levels. The challenge is to choose a model that is satisfactory in terms of goodness of fit, interpretability, parsimoniousness, prediction capability and computational costs. In order to support this choice, we propose a comparison approach based on a set of criteria summarized in a table that can be easily communicated to non-statisticians.

^{*}Corresponding address: Via dei Caniana 2, 24127 Bergamo (Italy).

E-mail: michela.cameletti@unibg.it

[†]R. Ignaccolo is also affiliated with Statistics Initiative, Collegio Carlo Alberto, Italy.

Our proposal - simple in principle but articulated in practice - holds true for many environmental phenomena where a hierarchical structure is suitable, a large-scale trend is included and a spatio-temporal covariance function has to be chosen.

We illustrate the details of our proposal through a case study concerning particulate matter concentrations in Piemonte region (Italy) during the cold season October 2005-March 2006. From the evaluation of the proposed criteria for our case study we draw some conclusions. First, a model with a complex hierarchical structure is globally preferable to one with a complex spatio-temporal covariance function. Moreover, in the absence of suitable computational resources, a model simple in structure and with a simple covariance function can be chosen, since it shows good prediction performance at reasonable computational costs.

Keywords: Particulate Matter PM_{10} , hierarchical models, spatial mapping, spatio-temporal covariance function, prediction performance indexes.

1 Introduction

Air quality is jeopardized by the presence of several pollutants. Particulate matter (PM) is one of the most critical air pollutants in Europe and, despite the improvements thanks to European Union legislation, it still has a heavy toll on human life ([Harrison *et al.*, 2008](#)). From a statistical perspective, many articles have been proposed for modelling the concentration of PM (and of other pollutants) and understanding its underlying complex

spatio-temporal dynamics. In particular, almost all the works propose spatial prediction techniques in order to obtain concentration maps useful for evaluating the health risk and assessing compliance with European and national directives, even in places where no measurement stations are located (e.g. [Fassò *et al.*, 2007](#); [Pollice and Jona Lasinio, 2010b](#); [Sahu *et al.*, 2006](#); [Sahu and Nicolis, 2007](#); [Zidek *et al.*, 2002](#)). Moreover, some works also develop methods for temporal forecasting (such as, for example, [Sahu and Mardia, 2005](#) and [Smith and Kolenikov, 2003](#)) or consider more critical pollutants at the same time (e.g. [Al-Awadhi and Al-Awadhi, 2006](#); [Shaddick and Wakefield, 2002](#); [Pollice and Jona Lasinio, 2010a](#)). When explanatory variables are available, measured by a monitoring network or simulated by a deterministic model, they contribute to the mean structure of the model, also known as large-scale component in geostatistical literature (see, for example, [Cressie, 1993](#)). To this regard sensitivity analysis techniques can be used to understand the role of covariates with respect to the output uncertainty, as shown in [Cocchi *et al.* \(2007\)](#) and [Fassò and Cameletti \(2010\)](#).

The common characteristic of many models found in literature is their hierarchical structure. This means that they are constructed by putting together conditional sub-models defined at each hierarchical stage. With reference to likelihood, this corresponds to taking a conditional point of view for which the joint probability distribution of a spatio-temporal process can be expressed as the product of certain simpler conditional distributions. This property makes it possible to deal with the complexity of spatio-temporal processes in a straightforward way which is the reason why hierarchical models have become so popular for modelling environmental processes, especially

from a Bayesian perspective ([Clark, 2005](#); [Wikle *et al.*, 1998](#); [Wikle, 2003](#)).

With particular reference to air pollution, covariates, such as meteorological and orographical variables, play an important role in seizing the large-scale variation of data, the influence of meteorology and geographical factors on pollutant concentration being well-known. Residual variability, also called small-scale variation, is modelled by a spatio-temporal process defined at a particular level of the hierarchy and by a spatio-temporal covariance function. In this context, an interesting question arises concerning the best hierarchical structure that can be combined with the spatio-temporal covariance function in order to seize the spatial and temporal dynamics of the considered process. For example, is it preferable to have a two-level model with a complex nonseparable covariance function or a model with a more complicated hierarchical structure but a simpler covariance function? Finding an answer to this question might be useful in order to provide environmental agencies with an effective statistical model for building reliable PM concentration maps, equipped with the corresponding uncertainty measure.

The goal of this paper is to provide an instrument to answer to this question. Therefore, we propose certain criteria that are as objective as possible for comparing spatio-temporal models for air quality data. Such criteria consider both goodness of fit and model complexity as well as computational costs. Actually, a model comparison for space-time models is proposed by [Huang *et al.* \(2007\)](#) through AIC and BIC criteria, which are investigated in their practical behaviour since asymptotic theory of these indexes for geostatistical models is substantially missing in literature. Moreover, these authors use mean squared prediction error at a fixed time to compare prediction ca-

pability. Specifically for fine particulate matter, [Pang *et al.* \(2010\)](#) compare, on practical grounds, ordinary kriging with Bayesian maximum entropy technique, as implemented in SEKS-GUI software through averaged estimation errors and error variances at four validation sites ([Kolovos *et al.*, 2006](#)).

Hence, in the absence of a space-time model selection theory, we first discuss the hierarchical models considered and then propose a set of empirical criteria to compare their intrinsic and computational complexity and their spatial prediction capability (through suitable indexes for air quality models opportunely summarized in a qualitative index). To implement our proposal, we deal with a case study concerning particulate matter in the Piemonte region (Italy). Here, we compare models on air quality real data; nonetheless, the proposed comparison approach holds true for general environmental phenomena where a hierarchical structure is suitable, a large-scale trend is included and a spatio-temporal covariance function has to be chosen.

The paper is organized as follows. In [Section 2](#) we introduce six models to be compared, specifying their hierarchical structure and spatio-temporal covariance function. The models, which constitute an extensive class of spatio-temporal hierarchical models, are fully discussed in [Section 3](#). [Section 4](#) describes the criteria used to establish which model best describes the data. In particular, we consider the intrinsic and computational complexity as well as the spatial prediction capability of each model. The paper also features an application proposed in [Section 5](#) regarding PM data measured in Piemonte during the 2005-2006 winter season. Unlike [Huang *et al.* \(2007\)](#) we implement all the models in a fully Bayesian framework via Markov Chains Monte Carlo (MCMC) methods. The paper ends with a discussion of the

results and the conclusions. The Appendices contain details about the full conditional and predictive distributions involved in the model estimation and spatial prediction procedures which may help.

2 Hierarchical spatio-temporal models

Let $z(s_i, t)$ be the scalar spatio-temporal process observed at site s_i and at time t where $i = 1, \dots, d$ and $t = 1, \dots, T$.

We assume the following *measurement equation* as the first level of the hierarchical models:

$$z(s_i, t) = u(s_i, t) + \varepsilon(s_i, t) \quad (1)$$

where $\varepsilon(s_i, t) \sim N(0, \sigma_\varepsilon^2)$ is the measurement error defined by a Gaussian white-noise process, serially and spatially uncorrelated. The term $u(s_i, t)$ is the so-called state process and can be defined, in turn, by other sub-levels giving rise to different hierarchical models described in the following subsections.

2.1 Model A

Model A is a two-level regression model characterized by a large-scale term and a spatio-temporal process for the residual small-scale component. In particular, quantity $u(s_i, t)$ of Eq.(1) is given by the sum of a trend $\mu(s_i, t)$ and a random process $\omega(s_i, t)$, as follows:

$$u(s_i, t) = \mu(s_i, t) + \omega(s_i, t). \quad (2)$$

Trend $\mu(s_i, t)$ is a function of k covariates and is given by

$$\mu(s_i, t) = X(s_i, t) \beta \quad (3)$$

where $X(s_i, t) = (X_1(s_i, t), \dots, X_k(s_i, t))$ denotes the covariate vector for site s_i at time t and $\beta = (\beta_1, \dots, \beta_k)'$ is the coefficient vector.

The zero-mean Gaussian process $\omega(s_i, t)$ of Eq.(2) is the residual process whose spatio-temporal covariance function depends on the parameter vector θ , namely

$$\text{Cov}(\omega(s_i, t), \omega(s_j, t')) = \sigma_\omega^2 C_\theta(h, l) \quad \forall i \neq j, t \neq t' \quad (4)$$

where σ_ω^2 is the constant in time and space variance of the process and $C_\theta(., .)$ is the spatio-temporal correlation function parameterized by θ , with $h = \|s_i - s_j\|$ the Euclidean distance between sites i and j , $l = |t - t'|$ the temporal lag between t and t' . As the covariance function (4) only depends upon h and l , the process is supposed to be second-order stationary and spatially isotropic. We consider three different forms for the covariance function, which give rise to the following models.

Model A1 Under the hypothesis that the process $\omega(s_i, t)$ is i.i.d. over time, it holds that

$$\text{Cov}(\omega(s_i, t), \omega(s_j, t')) = \begin{cases} 0 & \text{if } t \neq t' \\ \sigma_\omega^2 C_\theta(h) & \text{if } t = t' \end{cases} \quad (5)$$

where $C_\theta(h)$ is a purely spatial correlation function. To this regard

many functions defining isotropic second-order stationary spatial processes can be found in [Banerjee *et al.* \(2004, Ch.1\)](#).

Model A2 Adopting a separable approach, the space-time covariance function factors into a purely spatial and a purely temporal component and is given by

$$C_{\theta}(h, l) = C_{\theta_1}(l) C_{\theta_2}(h) \quad (6)$$

with $\theta = (\theta_1, \theta_2)$. In this case there is no interaction between space and time.

Model A3 Adopting a nonseparable approach, a space-time correlation function can be defined in a general form, as introduced in [Gneiting \(2002\)](#):

$$C_{\theta}(h, l) = \frac{1}{\psi(|l|^2)} \varphi\left(\frac{h^2}{\psi(|l|^2)}\right). \quad (7)$$

The function $\psi(x), x \geq 0$, is any completely monotone function while $\varphi(x), x \geq 0$, is any positive function with a completely monotone derivative. In particular, we consider the two cases described in [Table 1](#), corresponding to **Model A3-1** and **Model A3-2** respectively, where $0 < b \leq 1$, $\nu > 0$, the smoothness parameters α and γ take values in $(0, 1]$ and the space-time interaction parameter τ of Model A3-2 is defined in $[0, 1]$. Finally, the scaling parameters of time and space a and c are positive. It follows that $\theta_{A3-1} = (\alpha, b, \gamma, a, c)$ for Model A3-1 and $\theta_{A3-2} = (\alpha, \tau, \gamma, \nu, a, c)$ for Model A3-2.

2.2 Model B

Differently from Model A, Model B has a three-level hierarchical structure due to an additional equation required for modeling the temporal dynamics of a constant in space latent process. This means that for Model B the term $u(s_i, t)$ of Eq.(1) is given by the sum of a trend $\mu(s_i, t)$, of a unidimensional latent process $y(t)$ and of a purely spatial process $\omega(s_i)$, as follows:

$$u(s_i, t) = \mu(s_i, t) + y(t) + \omega(s_i), \quad (8)$$

where $y(t)$ and $\omega(s_i)$ are uncorrelated. In Eq.(8) the trend $\mu(s_i, t)$ is defined as in Eq.(3), while the term $y(t)$ refers to a constant in space unidimensional process with a temporal dynamics, with an autoregressive coefficient ρ , given by

$$y(t) = \rho y(t-1) + \eta(t) \quad (9)$$

where $y(0) \sim N(0, \sigma_B^2)$ and $\eta(t) \sim N(0, \sigma_\eta^2)$ are uncorrelated. Finally the spatial process $\omega(s_i)$ is assumed to follow a Gaussian distribution $N(0, \sigma_\omega^2)$ whose spatio-temporal covariance function is $Cov(\omega(s_i), \omega(s_j)) = \sigma_\omega^2 C_\theta(h)$.

2.3 Model C

Model C is a three-level hierarchical model defined by a spatio-temporal residual process with temporal dynamics. Specifically the term $u(s_i, t)$ of Eq.(1) is given by a trend $\mu(s_i, t)$ and a spatio-temporal process $y(s_i, t)$, namely

$$u(s_i, t) = \mu(s_i, t) + y(s_i, t). \quad (10)$$

The trend $\mu(s_i, t)$ has the same structure as in Model A and B (see Eq.(3)) while $y(s_i, t)$ is a spatio-temporal process that changes in time according to the following dynamics with autoregressive coefficient ρ :

$$y(s_i, t) = \rho y(s_i, t - 1) + \omega(s_i, t) \quad (11)$$

where $y(s_i, 0) \sim N(0, \sigma_C^2)$. Finally, $\omega(s_i, t) \sim N(0, \sigma_\omega^2)$ is a spatio-temporal process i.i.d. over time, so that its spatio-temporal covariance function is given by Eq.(5). Obviously y and ω are uncorrelated.

3 Model discussion

All the models are characterized by the same measurement equation given by (1) and by the trend term in Eq.(3) defined as a function of some covariates that can change in space and time (e.g. meteorological variables) or can be constant in time (e.g. spatial coordinates). The six models we consider differ in the way the residual detrended process is modelled and in how the spatio-temporal correlation is treated.

Model A, described in Section 2.1, features a very simple structure characterized by a unique residual spatio-temporal process for which three different covariance functions are considered. The most complex case is represented by Model A3, which is characterized by a nonseparable spatio-temporal covariance function. Considering that space and time interact together, from a physical point of view it is reasonable to adopt a nonseparable approach even if it has certain computing drawbacks. In fact, the size of the variance-

covariance matrix of the $\omega(s_i, t)$ process is $(dT \times dT)$ and, especially in the case of a dense network or a long monitoring period, matrix operations become infeasible from a computational point of view. Model A1 and Model A2 introduce some simplifications: Model A1 supposes that the process $\omega(s_i, t)$ is i.i.d. over time, which leads to a purely spatial covariance function (see Eq.(5)), while Model A2 is based on the separability hypothesis (see Eq.(6)). The choice of a spatio-temporal covariance function to be used depends on certain considerations. First of all, as stated in [Cressie and Huang \(1999\)](#), separable models are often chosen for convenience rather than for their ability to properly fit the data; basically, the same holds true for the i.i.d. over time hypothesis case. This is mainly related to the computational advantages in implementing and estimating a model with these simplified functions, which usually depend on a small number of parameters and involve smaller matrices. Generally speaking, a spatial covariance function that does not depend on time - the i.i.d. case - can be used when it is possible to show, for example by means of daily variograms, that the spatial correlation does not change significantly in time. If this is not the case, a separability test (e.g. [Fuentes, 2006](#)) should be performed in order to verify if the separability hypothesis can be assumed. Otherwise, a nonseparable covariance function should be used.

Model B and C differ from Model A for their three-level structure: in both cases, in fact, an equation is introduced for modelling the temporal dynamics of a latent process. In particular, for Model B this is given by a purely temporal AR(1) process (defined in (9)) which is supposed to be constant in space, while for Model C it has an AR(1) structure with innovations i.i.d.

over time (see Eq.(11)). It is important to point out that the size of the latent process is always unidimensional for Model B while for Model C it is defined by the number of spatial sites d . From a computational perspective this means that estimation procedure costs are higher for Model C than for Model B, even if Model B has an extra parameter to estimate, i.e. σ_η^2 .

It is interesting to point out that both the spatio-temporal covariance functions of Model B and Model C can be rewritten in a separable form, additive and multiplicative respectively (for further details see Appendix A).

4 Model comparison

We are interested in determining which is the most effective model for fitting the data. To achieve this goal, we compare the six models using a set of empirical criteria which explore the model complexity and prediction capability, as described hereafter.

4.1 Intrinsic and computational model complexity

The intrinsic complexity of a model can be roughly defined as the number of parameters to be estimated. Generally, as the number of parameters increases the estimation procedure becomes more complex since, in the fully Bayesian framework we adopt, steps are added to the Gibbs sampling algorithm. Moreover, if the parameter estimation requires the use of the Metropolis-Hastings (MH) algorithm, since no closed-form full conditional posterior distributions are available, the algorithm can become unstable and requires a larger number of iterations in order to reach convergence. This underscores out how a

richer model, in terms of parameters and hierarchical structure, is necessarily more complex from a computational point of view. To this regard, we also consider the size of the biggest matrix to be inverted for each model. Considering that matrix inversion is of order n^3 in computation (where n is the total number of data), a considerable size can give rise to massive computational loads infeasible to be carried out. To this regard, just to have an idea, consider that Matlab¹ on an Intel Core 2 Duo Mac (2.4 Ghz, 4GB RAM) takes about 15 and 76 seconds to invert - using the Cholesky factorization - a full symmetric matrix with size (4500×4500) and (8000×8000) , respectively. The same operations require about 5 and 21 seconds on an Intel Xeon 8 CPU cluster (2.66 Ghz, 8 GB RAM). In the hypothetical case of one parameter and 100000 iterations required for the convergence of the algorithm, at least 5 or 24 days would be necessary for the implementation of the 4500 or 8000-dimensional cases, respectively. Obviously, these computing times increase (in a non linear way) as the parameter set becomes larger.

This information about intrinsic and computational model complexity is shown in the first rows of Table 6 (pag. 53)². It is clear that the nonseparable models, Model A3-1 and Model A3-2, are the most complex ones since they have the biggest parameter vectors (respectively 7 and 8 parameters, excluding the β 's), all estimated using the MH algorithm. Moreover, the size of the variance-covariance matrix is $(dT \times dT)$. This means that, from a computational point of view, nonseparable models are extremely expensive and their implementation is expected to be severely time consuming.

¹We use Matlab R2009b with the Parallel Computing Toolbox.

²We suppose that for Model A1, B and C the parameter vector θ is unidimensional and that for Model A2 the parameter vectors θ_1 and θ_2 are unidimensional.

Model A2 has 4 parameters to be estimated and, thanks to the separability of the spatio-temporal covariance function, it enjoys the properties of the Kronecker product. This results in certain computational advantages regarding the inverse and the determinant of the $(dT \times dT)$ variance-covariance matrix, because we deal with $(d \times d)$ and $(T \times T)$ matrices separately (for details see Section B.2). In Table 6 the size of the biggest matrix to be inverted is $(T \times T)$ since we consider a small monitoring network, such that $d \ll T$.

In terms of complexity Model A1, Model B and Model C are more suitable because they have small parameter vector, make use of the MH algorithm in a limited way and are characterized by $(d \times d)$ -dimensional variance-covariance matrix. This means much more computationally manageable models.

With respect to the computational complexity, the models are compared also considering the computing time required to estimate the parameters and for performing the spatial predictions over the validation stations. The times are evaluated per iteration of the MCMC run and are computed using the above-defined Intel Xeon 8 CPU cluster. Generally, it is clear that, keeping all the rest equal, a model that can be quickly implemented is more desirable.

4.2 Spatial prediction capability

As the aim of the modelling is prediction, we compare models on the basis of their spatial prediction capability which is evaluated using certain performance indexes computed on validation stations. In particular, we consider five indicators based on the differences between predicted and observed data

together; moreover, we compute the observed coverage probability. More precisely, together with the usual root mean square error (RMSE) and the correlation coefficient ρ , we adopt the Normalised Mean Bias Factor (NMBF) recently introduced by [Yu *et al.* \(2006\)](#) and two indexes, named WNNR and NNR, proposed by [Poli and Cirillo \(1993\)](#) and defined on the Normalized Ratios between the predicted and the observed values.

For a fixed site s_i , let z_t be the observed time series and $\hat{z}_t$ the predicted time series with $t = 1, \dots, T$ (see [Appendix C](#) for details about prediction); moreover, denote with $\bar{z}$ and $\bar{\hat{z}}$ the corresponding mean values. The Normalised Mean Bias Factor is defined by

$$\text{NMBF} = \begin{cases} \frac{\sum_t \hat{z}_t}{\sum_t z_t} - 1 & \text{if } \bar{\hat{z}} \geq \bar{z} \\ 1 - \frac{\sum_t z_t}{\sum_t \hat{z}_t} & \text{if } \bar{\hat{z}} < \bar{z} \end{cases}.$$

NMBF is defined on $\mathbb{R}$ and has the advantage of both avoiding inflation and asymmetry, two problems discussed in [Yu *et al.* \(2006\)](#).

The Weighted Normalised mean square error of the Normalised Ratios is defined by

$$\text{WNNR} = \frac{\sum_t s_t^2 (1 - k_t)^2}{\sum_t s_t k_t},$$

while the non-weighted one is

$$\text{NNR} = \frac{\sum_t (1 - k_t)^2}{\sum_t k_t},$$

where $s_t = z_t/\bar{z}$ is the weight and $k_t = \exp -|\ln(\hat{z}_t/z_t)|$ is the normalised ratio. WNNR and NNR are both positive and have the advantage of taking

properly into account the peaks of observed data (see the discussion in [Poli and Cirillo, 1993](#)).

These 5 indexes and the observed coverage probability are computed for all the validation stations and for each model. Successively, these 6 performance measures are summarized over stations in a qualitative index based on “stars category”, so that more stars correspond to a better prediction capability of the model (see Table 6 on page 53).

5 Model comparison for PM₁₀ in Piemonte

In order to compare the models described in Section 2, below we consider particulate matter concentration with an aerodynamic diameter of less than 10 μm (PM₁₀, in $\mu g/m^3$) measured in the Piemonte region (Italy) during the October 2005 - March 2006 winter season. Piemonte is situated in the western part of the Po river basin, surrounded on three sides by the Alps. The pollutant dispersion is strongly influenced concurrently by the shelter effect of the Alps and by meteorological features which depend on the complex orography of the region. So, for example, we can have weak winds and stagnation in the central part of the region or breezes and foehn winds in mountains and valleys. For these reasons, we usually observe lower PM₁₀ concentration near the Alps, whereas higher pollution levels are detected in plains closer to urban areas. Actually, as expected, air quality is worse in urbanized areas where the most important emission sources, in other words industrial sites and high traffic levels, are located.

5.1 Data description

We analyze daily PM_{10} data measured by a network of $d = 24$ sites (see red triangles in Figure 1 and the corresponding labels in Figure 2) for $T = 182$ days (data are provided by an information system called *AriaWeb Regione Piemonte*). Moreover, we set aside data from 10 sites for validation purposes (see blue dots in Figure 1). The 24 sites are selected so that the amount of missing data does not exceed 20% and the missing data is not sequential. This guarantees good spatial coverage of the monitoring network so that stations can also be found in rural plain areas as well as in urbanized alpine valleys.

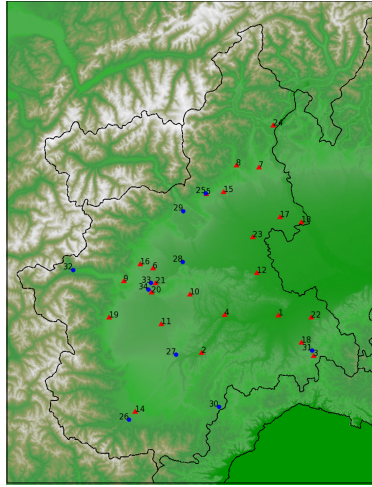


Figure 1: Locations of the 24 PM_{10} monitoring sites (red triangles) and 10 validation stations (blue dots). The complete names of the 24 stations are given in the x-axis labels of Fig. 2. The labels of the 10 validation sites are: 25 Biella - Largo Lamarmora, 26 Borgo San Dalmazzo, 27 Bra, 28 Chivasso, 29 Ivrea, 30 Saliceto, 31 Serravalle Scrivia, 32 Susa, 33 Torino - Piazza Rivoli, 34 Torino - Via Gaidano.

Figure 2 shows the distribution of the PM_{10} concentration by stations.

We can see that most of the stations have average and median concentration levels above $50 \mu g/m^3$, which is the threshold set by the European Commission (2008/50/EC directive) that can be exceeded no more than 30 days a year. All the distributions are positive skewed due to the occurrence of extremely polluted days. This kind of situation is quite common in the Po Valley, especially during the winter season when relatively stable atmospheric conditions, associated with a reduced washout of particulate matter, give rise to higher concentration.

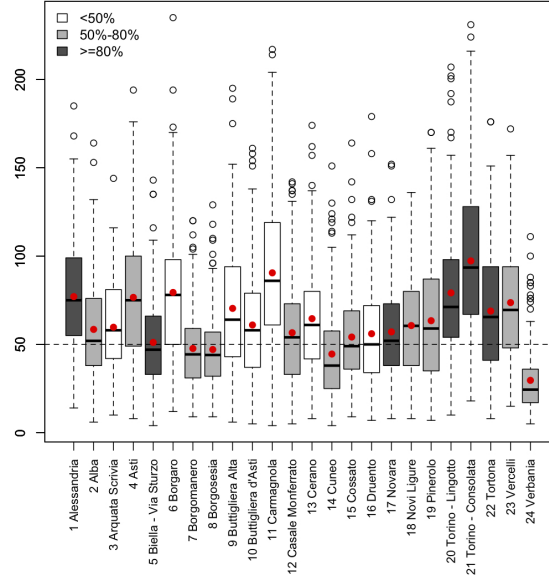


Figure 2: PM₁₀ concentration distribution over stations (the bold horizontal line corresponds to the median while the bold dot to the mean). The colours of the boxplots are given by the area type, i.e. the percentage of built-up surface (see the legend).

According to European legislation, each site is classified by area-type which has three categories (rural, suburban and urban) depending on the level of urbanisation of the area. More precisely, in Piemonte a site is classi-

fied rural if the percentage of built-up surface within a 1 km radius around it is less than 50%, suburban if that percentage is between 50% and 80%, urban if otherwise. Looking at Figure 2, area-type does not seem to have a visible effect on PM_{10} concentration. In fact, we observe rural stations with high levels of pollution as well as urban locations with lower concentration. This can seem unusual since area-type can be considered as a proxy of anthropogenic activities, but it could happen since PM_{10} pollution is a complex phenomenon strongly related to meteorology and orography, especially during the winter season.

Indeed, our models share a common trend component where some meteorological and orographical variables appear. The first ones are time-varying covariates obtained from a nested system of deterministic computer-based models implemented by the environmental agency ARPA Piemonte (Bande *et al.*, 2007; Finardi *et al.*, 2008). Such models provide the estimates on a regular $4 \text{ km} \times 4 \text{ km}$ grid of some meteorological variables, turbulence and chemicals parameters. By means of a preliminary regression analysis using Akaike's information criteria (AIC) and parameter significance, we choose the following covariates: daily maximum mixing height (HMX, in m), daily total precipitation (PREC, in mm), daily mean wind speed (WS, m/s), daily mean temperature (TEMP, in K) and daily emission rates of primary aerosols (EMI, in g/s). The mixing height is one of the fundamental parameters that characterize the structure of the atmosphere near the ground. Low mixing heights mean that the air is generally stagnant with very little vertical motion, and therefore pollutants are usually trapped near the ground surface. High mixing heights allow vertical mixing within a deep

layer of the atmosphere and good dispersion of pollutants. Thus a negative relationship is expected between PM_{10} and HMIIX. Precipitation plays a crucial role in explaining PM_{10} variations taking into account the wet removal of aerosol suspended in the atmosphere. Moreover, the daily mean values of wind speed are used in order to take into account the pollutant removal due to strong wind episodes, often combined with foehn conditions, that frequently occur in Piemonte during the winter time. Mean temperature influences the dispersion and accumulation of pollutants. On the one hand, it is related to atmospheric photochemical reactions and, consequently, to the production of secondary aerosols. On the other, low temperatures near the ground are often related to strong thermal inversion, one of the atmospheric features responsible for heavy pollutant events in urban area. Moreover, it is well-known that low temperatures cause an increase in particulate emissions from vehicle traffic sources. Finally, emissions take into account information about anthropogenic activities (e.g. energy production, domestic and industry production, road transport) which are the main sources of primary pollutants and the precursors of secondary pollutants. Together with these time-varying covariates, we also consider altitude (A , in m) and coordinates (UTMX and UTM Y , in km) in order to take into account the orography of the region.

In order to stabilize the variances, which increase with the mean values, and make the distribution of PM_{10} data approximately normal, we adopt the logarithm transformation of PM_{10} data. Subsequently, in order to investigate spatial and temporal correlation, we fit a simple regression model over all the 4368 ($=24 \times 182$) PM_{10} data with the covariates - standardized - described

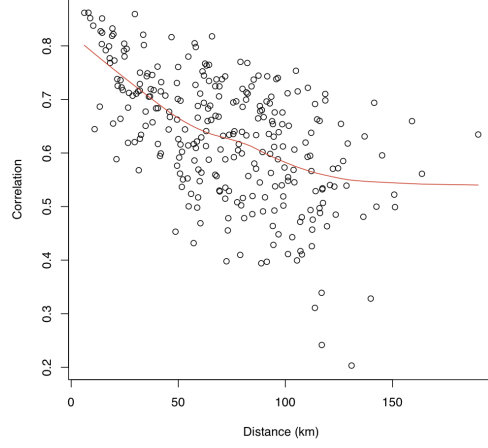


Figure 3: Spatial residual correlation cloud and *lowess* curve.

above. Figure 3 illustrates the residual correlation cloud, that is the set of spatial correlations between sites at different distances. It shows that, even after removing the so-called large-scale component given by the covariates, a spatial correlation still remains: the lowess curve decreases slowly with distance and settles around 0.6 at 100 km. Figure 4 shows the boxplots of empirical autocorrelations calculated on the residuals - over stations - and it stands out that a temporal structure also remains, with a temporal correlation of about 0.6 at the first time lag. These results show great evidence of the importance of using a spatio-temporal model for catching the complex structure and dynamics of the phenomenon.

5.2 Model implementation details

The models presented in Section 2 are now fitted on the PM_{10} concentration data described in the previous section.

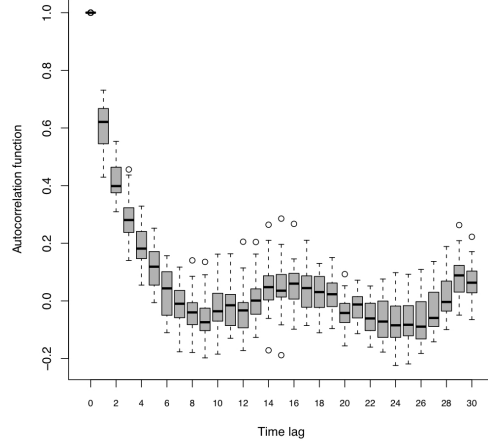


Figure 4: Boxplots of the residual empirical autocorrelations computed over the 24 stations.

The covariate vector $X(s_i, t)$ of Eq.(3), which also includes a constant for the intercept term, is given by

$$X(s_i, t) = \left(1, A(s_i), UTMX(s_i), UTMY(s_i), WS(s_i, t), \right. \\ \left. HMI X(s_i, t), TEMP(s_i, t), PREC(s_i, t), EMI(s_i, t) \right)$$

where $i = 1, \dots, 24$ and $t = 1, \dots, 182$.

With reference to the covariance functions, we adopt a purely spatial exponential form given by $C_\theta(h) = \exp(-\theta h)$ for Model A1, B and C and a double exponential structure for Model A2, $C_\theta(h, l) = \exp(-\theta_1 l) \exp(-\theta_2 h)$. For Model A3-1 and A3-2 we refer to the nonseparable functions defined in Table 1.

As regards inference, i.e. parameter estimation and spatial prediction, we

adopt a fully Bayesian approach via Monte Carlo Markov Chain (MCMC) methods. In particular, we use the Metropolis-within-Gibbs algorithm implemented in Matlab by ad-hoc written code. The complete details of the full conditional and predictive distributions are given in Appendices [B](#) and [C](#). Convergence is diagnosed by monitoring the mixing of the chains by means of traceplots together with autocorrelation and kernel density plots.

With reference to the prior specification, we assume Normal independent priors $N(0, 100)$ for each component of the β vector and Inverse Gamma distributions $IG(2, 1)$ for the variance parameters $(\sigma_\omega^2, \sigma_\varepsilon^2 \text{ and } \sigma_\eta^2)$. A Uniform prior distribution $U(0, 1)$ is chosen for the parameter θ of Model A1, B and C, and the parameter θ_2 of Model A2: this corresponds to a spatial correlation between 0.37 and 1 at 1 *km* of distance and between 0 and 1 at the maximum distance of 190 *km*. For the temporal decay parameter θ_1 of Model A2 we adopt a Uniform prior distribution $U(0.3, 3)$ whose support corresponds to a temporal range between 1 and 10 days (the range is determined using the relationship $\exp(-\theta_1 l) \approx 0.05$). For the parameters of the nonseparable correlation function of Model A3-1 and A3-2, we suppose vague prior distributions, that is $U(0, 1)$ for α, b, γ, τ and $U(0, 10)$ for a, c, γ .

Note that for Model A2 it is not possible to marginalize the posterior distribution over the latent process $u(s_i, t)$ (see Appendix [B](#) for details); thus, for sampling it we employ the *en bloc* procedure described in [Sahu et al. \(2006\)](#). Finally, for Model B and C the latent process y is sampled from its full conditional distribution by the Forward Filtering Backward Sampling (FFBS) algorithm as described in [West and Harrison \(1997\)](#).

5.3 Results and discussion

Table 2 displays the posterior estimates for the covariate coefficient vector β . Note that they are robust with respect to the model specification, meaning that the choice of a spatio-temporal covariance function as well as the introduction of a temporal dynamics equation do not affect strongly the estimates. In particular, the posterior mean of the intercept is around 3.9 on the log scale, which corresponds to an average pollution level of about $49.4 \mu g/m^3$. As expected, a significant and positive relationship can be seen between emissions (EMI) and PM_{10} concentration. Moreover, the significance of the coefficients of WS, HMIX, TEMP and PREC confirms the importance of meteorological variables on air quality. The UTMX and UTMY coefficient estimates are both negative, indicating a decreasing spatial trend from West to East and from South to North. Finally, altitude (A) has a significant effect in reducing PM_{10} concentration.

With regards to the variance estimates in Table 3, we observe that more variation is explained by the spatial or spatio-temporal term rather than by the measurement error. For Model A1, A3-1, A3-2 and B, the estimate of σ_ω^2 is larger - between 1 and 3 times - than the estimate of σ_ε^2 while for Model A2 and C it is, respectively, 16 and 74 times larger. The fact that for Model C the measurement error variance is so small with respect to the spatio-temporal error variance can be ascribed to its complex structure, clearly useful for explaining variability of the spatio-temporal process.

Table 4 reports the posterior estimates of the correlation function coefficients. For Model A1, the estimate of θ is 0.0033 giving rise to a strong spatial

correlation which decreases slowly with distance. In fact, at a maximum distance of 190 *km* the spatial correlation is about 0.54. For Model A2, we find $\theta_1 = 0.492$ and $\theta_2 = 0.032$ that correspond to a temporal and spatial range of, respectively, 6 days and 95 *km* (the spatial range is determined using the relationship $\exp(-\theta_2 h) \approx 0.05$). Instead, for Model B we estimate a range of about 60 *km*, indicating that the spatial correlation decreases rapidly and that at 190 *km* there is no correlation. On the other hand, for Model B the estimate of the AR(1) temporal correlation coefficient ρ is 0.8313, confirming the short-term temporal persistence of particulate matter. Model C seems to have a spatial correlation similar to the one of Model A1 (the posterior mean of θ is 0.0022 with a correlation of about 0.66 at 190 *km*), while its temporal correlation is weaker (estimate of ρ equal to 0.6535) than the one of Model B.

Posterior means and credible intervals of the nonseparable correlation function parameters (Model A3-1 and A3-2) are shown in Table 5. Using these estimates, the correlation functions defined in Eq.(7) and Table 1 are plotted in Figure 5 (with 0-10 days of temporal lag and 0-100 *km* of spatial distance). The two surfaces resemble each other, in other words the spatio-temporal correlation decreases similarly with respect to both the temporal lag and the spatial distance.

As discussed in Section 3, the first rows of Table 6 provide information about intrinsic and computational model complexity. Moreover, the table contains estimation and prediction times (in seconds per iteration). As expected, the estimation computing time increases with the number of parameters and the size of the biggest matrix to be inverted: in particular, Model

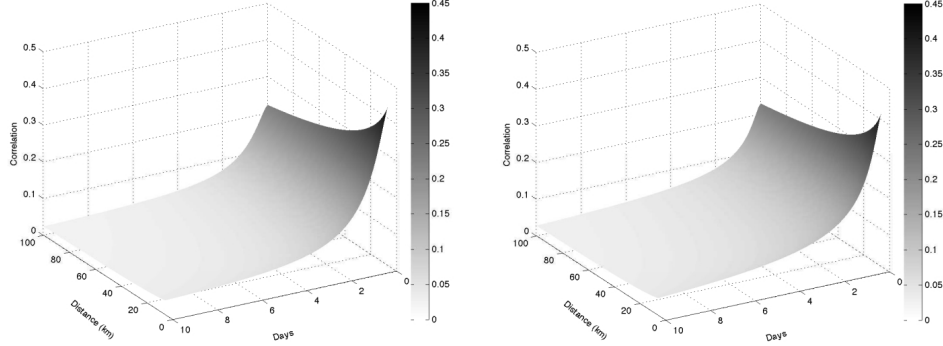


Figure 5: Nonseparable spatio-temporal correlation function for Model A3-1 (left) and Model A3-2 (right) with the estimated parameters in Table 5.

A3-1 and A3-2 are the most time consuming (about 45 and 47 seconds, respectively) while the simple Model A1 is the fastest to be estimated (0.014 seconds per iteration). Focusing on Model B and C, both characterized by an additional equation for the temporal dynamics requiring the use of the FFBS algorithm (see (9) and (11)), we observe that Model B is less time-consuming than Model C, even if it has one additional parameter to estimate, σ_η^2 , and adopts more extensively the MH algorithm. This can be explained by considering that, while for Model B the FFBS algorithm involves 1-dimensional terms, for Model C it deals with d -dimensional terms. Model A2 requires more time (about 4 seconds) than Model A1, B and C because it uses *en bloc* sampling for $u(s_i, t)$, that is an iterative computationally expensive procedure.

Regarding prediction time, Model A1 and B have very similar performances since the prediction procedure is almost the same (see Appendix C). Model C is slightly slower because it has a more complicated temporal dynamics requiring the introduction of an additional step for the composition

sampling procedure given by Eq.(24). The prediction requires more time for Model A3-1 and A3-2 given that the routine involves variance-covariance matrices with size $(dT \times dT)$. The 27 seconds per iteration necessary for Model A2 could be a cause for worry, but they are justified by the fact that drawing from the Normal distribution, as done in Eq.(22), slows down the composition sampling procedure.

The last row of Table 6 summarizes in “stars category” (from 1 to 3) the prediction capability of our models. The stars are based on the values of the performance measures defined in Section 4.2 and computed for the 10 validation sites. Figure 6 shows through boxplots, for each model, the distribution of the indexes and of the observed coverage probability over station. The boxplots of Model A2 show the greatest variability and the largest values of RMSE, WNNR and NNR, meaning that the prediction is poorer where the observed time series have peaks. Thus, overall Model A2 has the worst prediction performance and it is classified with one star. Conversely, Model C boxplots show less variability, especially looking at NMFB, RMSE and the observed coverage probability. The other models seem to have very similar prediction performances, except for the observed coverage probability distributions of Model A3-1 and A3-2 that are left skewed (their first quartiles are around 0.87). For this reason we assign three stars to Model C and two stars to the remaining models.

The entire Table 6 provides the set of criteria useful for choosing the most effective model in our case study. First of all, we discard Model A2 because of its poor prediction performance and long computing time for spatial prediction. Also, we can not suggest implementing Model A3-1 and

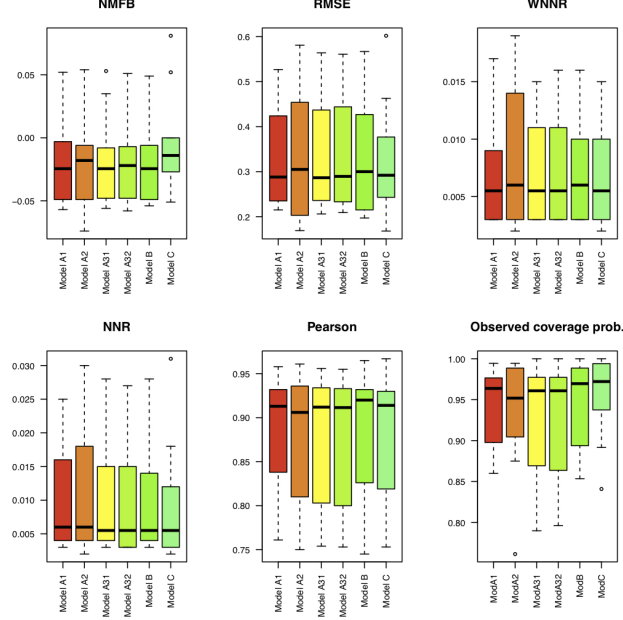


Figure 6: Boxplots of the performance measure distributions computed for each model over the 10 validation stations.

A3-2 - with the complex nonseparable covariance function - for reasons of computational costs that are not offset by a significant improvement of the prediction capability. Among the remaining models, Model B is the less appealing one since it has less stars than Model C and is slower than Model A1 in the estimation routine. Finally, the choice is between Model A1 - simple in structure and covariance function - and Model C, which differs in the hierarchy for an additional equation explicating the temporal dynamics. As observed, Model C has a slightly better prediction capability, which is offset by a larger computational cost. In the absence of suitable computational resources, Model A1 has a good performance at a reasonable cost.

6 Conclusions

In this work we propose a comparison of six models for air quality data, taking into account different levels of complexity either in the hierarchical structure or in the spatio-temporal covariance function. This makes it possible to analyze a wide range of models suitable for complex environmental phenomena characterized by both a large and small scale of variation, as well as by spatial relationships, temporal dynamics and spatio-temporal interactions. In this work we focus on PM concentration, comparing models on the basis of certain criteria that take into account intrinsic complexity, computational costs and spatial prediction capability. These criteria are particularly important because environmental agencies need to evaluate air quality status, predicting pollutant concentration at unmonitored sites, at a reasonable computational cost. The table that summarizes the proposed set of criteria - Table 6 in our case study - has the merit of being easy to communicate to non-statisticians, who may be environmental agency practitioners or policy makers, responsible for drawing up air quality legislations. Moreover, these spatial prediction maps can be used in risk assessment analysis and ecological risk models, as continuous exposure levels, or as validation criteria in a network design study.

The compared models share a large-scale trend component whose estimated coefficient are robust as the model changes. Furthermore, this property allows models to be compared only through the residual detrended process. First, this component is modelled by directly specifying certain spatio-temporal correlation functions with increasing complexity, from the i.i.d. over

time case to the nonseparable form (Model A1, A2, A3-1 and A3-2). Then, in Model B the detrended process is defined as the sum of a constant in space process, with an AR(1) temporal dynamics, and a purely spatial process. Model C, on the other hand, features a spatial process which evolves in time according to an AR(1) equation with innovations i.i.d. over time.

Our case study, concerning the prediction of PM_{10} in Piemonte, allows us to identify the best combination of hierarchical structure and spatio-temporal covariance function based on the evaluation of the proposed criteria. The application highlights that complex spatio-temporal covariances (in Model A3-1 and A3-2) require computational costs which are too high with respect to the observed prediction capability. The separable Model A2 is discarded because of its poor performance in the predicting procedure (one star). Model B, on the other hand, is comparable to Model A1 with respect to the prediction capability (two stars) but it is slower from a computational point of view. Model C is the only three-star predictor: this suggests that, in our case study, a model with a complex hierarchical structure is globally preferable to one with a complex spatio-temporal covariance function. Thus, the last comparison is between Model A1 and C: since the latter requires small additional computational cost in order to be awarded an additional star, our final suggestion is to choose according to the available computational resources.

Obviously, the conclusions and the final suggestion in our case study do not constitute the general answer to the question about the best combination of hierarchical structure and spatio-temporal covariance function. For example, Model A3-1 and A3-2 could have a better prediction performance in a case study with a larger spatial domain and a longer time period, as well as

a thicker monitoring network. Moreover, their computational costs could be reduced by simplifying the structure of the dense covariance matrices, using for example covariance tapering techniques (Furrer *et al.*, 2006). Nonetheless, the proposed set of criteria summarized in a simple table - jointly with the model discussion - makes the strategy applied in our PM₁₀ application portable to other environmental studies, where decision makers choose which model to implement.

The choice of the model is also related to the size of the dataset used to estimate the parameters as well as the grid resolution for spatial mapping, since we adopt computing intensive statistical methods. The increasing availability of large spatio-temporal datasets in many environmental fields gives rise to the so-called “big n” problem and to the infeasibility of matrix operations whose complexity increases in cubic order. To tackle this issue, we need to adopt efficient computational strategies. A solution may be given by the implementation of parallel linear algebra algorithms based on graphical processors (GPU), which is part of our ongoing research.

Acknowledgements

The authors would like to thank Veronica J. Berrocal, Alan E. Gelfand, Giovanna Jona Lasinio for the useful discussions and comments.

References

- Al-Awadhi, F. and Al-Awadhi, S. (2006). Spatial-temporal model for ambient air pollutants in the state of Kuwait. *Environmetrics*, **17**, 739–752.
- Bande, S., Clemente, M., De Maria, R., Muraro, M., Picollo, M., Arduino, G., Calori, G., Finardi, S., Radice, P., Silibello, C., and Brusasca, G. (2007). The modelling system supporting Piemonte region yearly air quality assessment. *Proceedings of 6th International Conference on Urban Air Quality, Limassol, Cyprus, 27-29 March 2007*.
- Banerjee, S., Carlin, B., and Gelfand, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Monographs on Statistics and Applied Probability. Chapman and Hall, New York.
- Clark, J. (2005). Why environmental scientists are becoming Bayesians. *Ecology Letters*, **8**, 2–14.
- Cocchi, D., Greco, F., and Trivisano, C. (2007). Hierarchical space-time modelling of PM₁₀ pollution. *Atmospheric environment*, **41**, 532–542.
- Cressie, N. (1993). *Statistics for Spatial Data*. Wiley, New York.
- Cressie, N. and Huang, H. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, **94**, 1330–1340.
- Fassò, A. and Cameletti, M. (2010). A unified statistical approach for simulation, modelling, analysis and mapping of environmental data. *Simulation*, **86**(3), 139–154.

- Fassò, A., Cameletti, M., and Nicolis, O. (2007). Air quality monitoring using heterogeneous networks. *Environmetrics*, **18**(3), 245–264.
- Finardi, S., De Maria, R., D’Allura, A., Cascone, C., Calori, G., and Lollobrigida, F. (2008). A deterministic air quality forecasting system for Torino urban area, Italy. *Environmental Modelling and Software*, **23**(3), 344–355.
- Fuentes, M. (2006). Testing for separability of spatial-temporal covariance functions. *Journal of Statistical Planning and Inference*, **136**, 447–466.
- Furrer, R., Genton, M., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, **15**(3), 502–523.
- Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, **97**, 590–600.
- Harrison, R., Stedman, J., and Derwent, D. (2008). New directions: Why are PM_{10} concentrations in Europe not falling? *Atmospheric Environment*, **42**, 603–606.
- Huang, H., Martinez, F., Mateu, J., and Montes, F. (2007). Model comparison and selection for stationary space-time models. *Computational Statistics and Data Analysis*, **51**, 4577–4596.
- Kolovos, A., Yu, H., and Christakos, G. (2006). *SEKS-GUI v.0.6 User Manual*. Department of Geography, San Diego State University: San Diego, CA.

- Pang, W., Christakos, G., and Wang, J. (2010). Comparative spatiotemporal analysis of fine particulate matter pollution. *Environmetrics*, **21**, 305–317.
- Poli, A. and Cirillo, M. (1993). On the use of the normalized mean square error in evaluating dispersion model performance. *Atmospheric Environment*, **27**, 2427–2434.
- Pollice, A. and Jona Lasinio, G. (2010a). A multivariate approach to the analysis of air quality in a high environmental risk area. *Environmetrics*. DOI: 10.1002/env.1059.
- Pollice, A. and Jona Lasinio, G. (2010b). Spatiotemporal analysis of the PM₁₀ concentration over the Taranto area. *Environmental Monitoring and Assessment*, **162**, 177–190.
- Sahu, S. and Mardia, K. (2005). A Bayesian Kriged-Kalman model for short-term forecasting of air pollution level. *Journal of the Royal Statistical Society, Series C*, **54**, 223–244.
- Sahu, S. and Nicolis, O. (2007). An evaluation of European air pollution regulations for particulate matter monitored from a heterogeneous network. *Environmetrics*, **20**, 943–961.
- Sahu, S., Gelfand, A., and Holland, D. (2006). Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural, Biological and Environmental Statistics*, **11**, 61–86.
- Shaddick, G. and Wakefield, J. (2002). Modelling daily multivariate pollutant data at multiple sites. *Journal of Applied Statistics*, **3**, 351–372.

- Smith, R. and Kolenikov, S. (2003). Spatiotemporal modeling of PM_{2.5} data with missing values. *Journal of Geophysical Research*, **108**(D24), 11–1,11–11.
- Tanner, M. A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Springer, New York.
- West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer.
- Wikle, C. K. (2003). Hierarchical models in environmental science. *International Statistical Review*, **71**, 181–199.
- Wikle, C. K., Berliner, L., and Cressie, N. (1998). Hierarchical bayesian space-time models. *Journal of Environmental and Ecological Statistics*, **5**, 117–154.
- Yu, S., Eder, B., Dennis, R., Chu, S., and Schwartz, S. (2006). New unbiased symmetric metrics for evaluation of air quality models. *Atmospheric Science Letters*, **7**, 26–34.
- Zidek, J., Sun, L., Le, N., and Ozkaynak, H. (2002). Contending with space-time interaction in the spatial prediction of pollution: Vancouver’s hourly ambient PM₁₀ field. *Environmetrics*, **13**, 595–613.

A Details on covariance separability

For Model B it holds that

$$\begin{aligned}
Cov(u(s_i, t), u(s_j, t')) &= E(u(s_i, t)u(s_j, t')) - E(u(s_i, t))E(u(s_j, t')) \\
&= E(y(t)y(t')) - E(y(t))E(y(t')) + E(\omega(s_i)\omega(s_j)) \\
&= Cov(y(t), y(t')) + Cov(\omega(s_i), \omega(s_j)) \\
&= \rho^l \frac{\sigma_\eta^2}{1 - \rho^2} + \sigma_\omega^2 C_\theta(h)
\end{aligned}$$

where $h = \|s_i - s_j\|$ and $l = |t - t'|$. This means that the spatio-temporal covariance function is given by the sum of a purely temporal and a purely spatial covariance function. In particular, the temporal term is defined by the AR(1) structure of y (see (9)) with $|\rho| < 1$.

Turning to Model C we have that $Cov(u(s_i, t), u(s_j, t'))$ is equal to $Cov(y(s_i, t), y(s_j, t'))$. Moreover, recalling that any AR(1) process can be rewritten using an infinite-order moving average representation, it follows that

$$\begin{aligned}
Cov(y(s_i, t), y(s_j, t')) &= Cov\left(\sum_{k=0}^{\infty} \rho^k \omega(s_i, t - k), \sum_{k'=0}^{\infty} \rho^{k'} \omega(s_j, t' - k')\right) \\
&= \sum_{k, k'=0}^{\infty} \left(\rho^{k+k'} Cov(\omega(s_i, t - k), \omega(s_j, t' - k'))\right) \\
&= \frac{\rho^l}{1 - \rho^2} \sigma_\omega^2 C_\theta(h).
\end{aligned}$$

Hence the spatio-temporal covariance function can be rewritten in a separable multiplicative form where $\frac{\rho^l}{1 - \rho^2}$ is a valid temporal correlation function if

$$|\rho| < 1.$$

B Derivation of posterior distributions

Let us denote with Ψ the parameter vector to be estimated. Generally speaking the joint posterior distribution is given by

$$f(\Psi, U | Z) \propto f(Z | U, \Psi) f(U | \Psi) f(\Psi) \quad (12)$$

where the notation $f(\cdot)$ is used for the probability density function and Z and U denote, respectively, the collections of data $z(s_i, t)$ and of the latent processes $u(s_i, t)$. Note that independent prior distributions are chosen for the parameters, so that $f(\Psi) = \prod_{i=1}^{\dim(\Psi)} f(\Psi_i)$. Conditionally on U the observations $z(s_i, t)$ are serially independent and Eq.(12) can be rewritten as follows:

$$f(\Psi, U | Z) \propto \prod_{t=1}^T f(Z_t | U, \Psi) f(U | \Psi) f(\Psi) \quad (13)$$

where $Z_t = (z(s_1, t), \dots, z(s_d, t))'$ and $f(Z_t | U, \Psi)$ can be derived through the measurement equation (1). The joint posterior distribution (13) is completely specified when $f(U | \Psi)$ is defined. Since this conditional distribution is specific for each model introduced in Section 2, we describe each case separately in the following subsections specifying in the subscript of Ψ the model to which it refers.

Note that, as described in [Banerjee *et al.* \(2004\)](#), in the case of hierarchical models it is preferable, where possible, to marginalize the posterior

distribution over U , thus obtaining the following posterior distribution

$$f(\Psi \mid Z) \propto \prod_{t=1}^T f(Z_t \mid \Psi) f(\Psi). \quad (14)$$

In the sequel, this marginal approach is used for all models except Model A2, where the marginalization prevents the use of the properties of the Kronecker product described in Section B.2.

In order to implement the Gibbs sampling algorithm we derive the full conditional distributions and, when these are not available in an exact closed-form, we introduce a Metropolis-Hastings (MH) sampling step (which is the so-called Metropolis-within-Gibbs algorithm).

B.1 Model A1

As specified in Section 2.1 the parameter vector for Model A1 is given by $\Psi_{A_1} = (\beta, \sigma_\varepsilon^2, \theta, \sigma_\omega^2)$. Moreover, the following Gaussian d -dimensional conditional distribution holds

$$f(Z_t \mid \Psi_{A_1}) \sim N_d(X_t \beta, \Sigma_{\omega+\varepsilon})$$

where $X_t = (X(s_1, t)', \dots, X(s_d, t)')'$ is the $(d \times k)$ covariate matrix and the variance-covariance matrix is defined as

$$\Sigma_{\omega+\varepsilon} = \sigma_\omega^2 C_\theta(\|s_i - s_j\|)_{i,j=1,\dots,d} + \sigma_\varepsilon^2 I_d, \quad (15)$$

with $C_\theta(\|s_i - s_j\|)_{i,j=1,\dots,d} = C_\theta(\mathbf{h})$ with elements evaluated through the spatial correlation function shown in (2.5). Then, it follows that the posterior distribution is given by

$$f(\Psi_{A1} | Z) \propto \prod_{i=1}^{\dim(\Psi_{A1})} f(\Psi_{A1i}) \times |\Sigma_{\omega+\varepsilon}|^{-\frac{T}{2}} \exp \left[-\frac{1}{2} \sum_{t=1}^T (Z_t - X_t \beta)' \Sigma_{\omega+\varepsilon}^{-1} (Z_t - X_t \beta) \right].$$

Taking a Normal prior distribution for β , that is $\beta \sim N_k(0, \Sigma_0)$, straightforward calculation yields a Gaussian full conditional distribution for β with mean AB' and covariance A where $A = \left(\Sigma_0^{-1} + \sum_{t=1}^T X_t' \Sigma_{\omega+\varepsilon}^{-1} X_t \right)^{-1}$ and $B = \sum_{t=1}^T (Z_t' \Sigma_{\omega+\varepsilon}^{-1} X_t)$. In order to estimate the remaining parameters, the MH algorithm is used.

B.2 Model A2

Let us denote with $U_t = (U(s_1, t), \dots, U(s_d, t))'$ the latent process at time t . Moreover, let $U = \{U_1, \dots, U_T\}$ be the $(dT \times 1)$ random effects vector blocked by sites³ and $Z = \{Z_1, \dots, Z_T\}$ the $(dT \times 1)$ data vector. The distribution of U is Gaussian with mean vector given by $X\beta$ where $X = \{X_1, \dots, X_T\}$ is the $(dT \times k)$ array of covariates and β is the corresponding coefficient vector. The variance-covariance matrix of U is defined as

$$\Sigma_U = \sigma_\omega^2 C_\theta(\|s_i - s_j\|, |t - t'|)_{\substack{i,j=1,\dots,d \\ t,t'=1,\dots,T}} = \sigma_\omega^2 C_\theta(\mathbf{h}, \mathbf{l})$$

³Here, and in the sequel, braces are used for column stacking of the vectors involved.

where $C_\theta(\|s_i - s_j\|, |t - t'|)_{\substack{i,j=1,\dots,d \\ t,t'=1,\dots,T}} = C_\theta(\mathbf{h}, \mathbf{l})$ with elements evaluated through the separable spatio-temporal correlation function given in Eq.(6).

Thanks to the separability property, the correlation matrix $C_\theta(\mathbf{h}, \mathbf{l})$ can be written as $C_\theta(\mathbf{h}, \mathbf{l}) = C_{\theta_1}(\mathbf{l}) \otimes C_{\theta_2}(\mathbf{h})$, where $C_{\theta_1}(\mathbf{l})$ and $C_{\theta_2}(\mathbf{h})$ denote respectively the temporal and spatial correlation matrix. Thus, the parameter vector for Model A2 is $\Psi_{A2} = (\beta, \sigma_\varepsilon^2, \sigma_\omega^2, \theta_1, \theta_2)$ and the joint posterior distribution is given by the following equation:

$$\begin{aligned} f(\Psi_{A2}, U \mid Z) &\propto (\sigma_\varepsilon^2)^{-\frac{Td}{2}} \exp \left[-\frac{1}{2\sigma_\varepsilon^2} (Z - U)' (Z - U) \right] \\ &\times (\sigma_\omega^2)^{-\frac{Td}{2}} |C_\theta(\mathbf{h}, \mathbf{l})|^{-\frac{1}{2}} \exp \left[-\frac{1}{2\sigma_\omega^2} (U - X\beta)' C_\theta(\mathbf{h}, \mathbf{l})^{-1} (U - X\beta) \right] \\ &\times \prod_{i=1}^{\dim(\Psi_{A2})} f(\Psi_{A2i}) \end{aligned}$$

where, thanks to the Kronecker product properties, $|C_\theta(\mathbf{h}, \mathbf{l})| = |C_{\theta_1}(\mathbf{l})|^d |C_{\theta_2}(\mathbf{h})|^T$ and $C_\theta(\mathbf{h}, \mathbf{l})^{-1} = C_{\theta_1}(\mathbf{l})^{-1} \otimes C_{\theta_2}(\mathbf{h})^{-1}$.

Taking a Normal prior distribution for β , that is $\beta \sim N_k(0, \Sigma_0)$, from Eq.(16) it follows that the full conditional distribution for β is Gaussian with mean AB' and variance A where $A = \left(\Sigma_0^{-1} + \frac{1}{\sigma_\omega^2} X' C_\theta(\mathbf{h}, \mathbf{l})^{-1} X \right)^{-1}$ and $B = \frac{1}{\sigma_\omega^2} U' C_\theta(\mathbf{h}, \mathbf{l})^{-1} X$. Moreover, for the variance parameter we have the following conditional distributions:

$$\sigma_\omega^2 \sim IG \left(a + \frac{Td}{2}, b + \frac{1}{2} (U - X\beta)' C_\theta(\mathbf{h}, \mathbf{l})^{-1} (U - X\beta) \right),$$

$$\sigma_\varepsilon^2 \sim IG \left(a + \frac{Td}{2}, b + \frac{1}{2} (Z - U)' (Z - U) \right)$$

where a and b denote the hyperparameters of the corresponding prior Inverse-

Gamma distributions $IG(a, b)$. For the estimation of θ_1 and θ_2 we adopt the MH algorithm. Finally, let us recall that the spatio-temporal process U is sampled using the *en bloc* procedure described in [Sahu *et al.* \(2006\)](#).

B.3 Model A3

Using the notation introduced in the previous section for Model A2, the posterior distribution for Model A3 is

$$f(\Psi_{A3} | Z) \propto |\Sigma_{\omega+\varepsilon}|^{-\frac{1}{2}} \exp \left[-\frac{1}{2} (Z - X\beta)' \Sigma_{\omega+\varepsilon}^{-1} (Z - X\beta) \right] \times \prod_{i=1}^{dim(\Psi_{A3})} f(\Psi_{A3i}) \quad (16)$$

where $\Psi_{A3-1} = (\beta, \sigma_\varepsilon^2, \sigma_\omega^2, \theta_{A3-1})$ for Model A3-1 and $\Psi_{A3-2} = (\beta, \sigma_\varepsilon^2, \sigma_\omega^2, \theta_{A3-2})$ for Model A3-2. The variance-covariance matrix in (16) is given by

$$\Sigma_{\omega+\varepsilon} = \sigma_\omega^2 C_\theta(\mathbf{h}, \mathbf{l}) + \sigma_\varepsilon^2 I_{dT}$$

where the elements of $C_\theta(\mathbf{h}, \mathbf{l})$ are evaluated through the nonseparable correlation functions $C_\theta(h, l)$ given in Eq. (2.7) of (7).

From Eq.(16) we can easily obtain that the full conditional distribution for β is Gaussian with mean AB' and variance A where $A = (\Sigma_0^{-1} + X' \Sigma_{\omega+\varepsilon}^{-1} X)^{-1}$ and $B = Z' \Sigma_{\omega+\varepsilon}^{-1} X$, assuming a Normal prior distribution for β , that is $\beta \sim N_k(0, \Sigma_0)$. All the remaining parameters are estimated through the Metropolis-Hastings algorithm.

B.4 Model B

Given the hierarchical structure of Model B (see Equations (8) and (9)), Eq.(14) can be rewritten as

$$f(\Psi_B | Z) = \prod_{t=1}^T f(Z_t | Y, \Psi_B) f(Y | \Psi_B) f(\Psi_B) \quad (17)$$

where $\Psi_B = (\beta, \sigma_\varepsilon^2, \sigma_\omega^2, \sigma_\eta^2, \theta, \rho)$ and $Y = \{Y_1, \dots, Y_T\}$, with Y_t denoting the constant in space unidimensional latent process at time t , with $t = 1, \dots, T$. The Markovian structure defined in Eq.(9) leads us to rewrite the term $f(Y | \Psi_B)$ in (17) as follows

$$f(Y | \Psi_B) = f(Y_0 | \Psi_B) \prod_{t=1}^T f(Y_t | Y_{t-1}, \Psi_B). \quad (18)$$

Moreover, the conditional probability distributions that occur in the previous equations are the following:

$$\begin{aligned} f(Z_t | Y, \Psi_B) &\sim N_d(X_t \beta + K Y_t, \Sigma_{\omega+\varepsilon}), \\ f(Y_t | Y_{t-1}, \Psi_B) &\sim N_1(\rho Y_{t-1}, \sigma_\eta^2), \\ f(Y_0 | \Psi_B) &\sim N_1(0, \sigma_B^2), \end{aligned}$$

where $\Sigma_{\omega+\varepsilon}$ is defined by Eq.(15), as for Model A1, and K is a $(d \times 1)$ -dimensional ones vector. Taking the logarithm of Eq. (17), the posterior

distribution for Model B is given by

$$\begin{aligned}
f(\Psi_B | Z) &\propto |\Sigma_{\omega+\varepsilon}|^{-\frac{1}{2}} \exp \left[-\frac{1}{2} \sum_{t=1}^T (Z_t - X_t \beta - KY_t)' \Sigma_{\omega+\varepsilon}^{-1} (Z_t - X_t \beta - KY_t) \right] \\
&\times (\sigma_B^2)^{-\frac{1}{2}} \exp \left(-\frac{1}{2} Y_0^2 \right) \\
&\times (\sigma_\eta^2)^{-\frac{T}{2}} \exp \left[-\frac{1}{2\sigma_\eta^2} \sum_{t=1}^T (Y_t - \rho Y_{t-1})^2 \right] \prod_{i=1}^{\dim(\Psi_B)} f(\Psi_{Bi}).
\end{aligned}$$

Straightforward calculation yields a Gaussian full conditional distribution for β with mean AB' and covariance A where $A = \left(\Sigma_0^{-1} + \sum_{t=1}^T X_t' \Sigma_{\omega+\varepsilon}^{-1} X_t \right)^{-1}$ and $B = \sum_{t=1}^T (Z_t' \Sigma_{\omega+\varepsilon}^{-1} X_t - (KY_t)' \Sigma_{\omega+\varepsilon}^{-1} X_t)$, assuming a Normal prior distribution for β , that is $\beta \sim N_k(0, \Sigma_0)$. For the estimation of σ_η^2 we consider an Inverse Gamma prior $IG(a, b)$ and sample from the following full conditional distribution

$$\sigma_\eta^2 \sim IG \left(\frac{T}{2} + a, b + \frac{1}{2} \sum_{t=1}^T (Y_t - \rho Y_{t-1})^2 \right).$$

For the other parameters, $\sigma_\varepsilon^2, \sigma_\omega^2, \theta$ and ρ , we adopt the MH algorithm, while the latent process Y is sampled using the Forward Filtering Backward Sampling (FFBS) as described in [West and Harrison \(1997\)](#).

B.5 Model C

With regard to Model C, Eq.(17) and Eq.(18) still hold substituting Ψ_B with Ψ_C and using the following conditional distributions:

$$\begin{aligned} f(Z_t | Y, \Psi_C) &\sim N_d(X_t\beta + Y_t, \Sigma_\varepsilon), \\ f(Y_t | Y_{t-1}, \Psi_C) &\sim N_d(PY_{t-1}, \Sigma_\omega), \\ f(Y_0 | \Psi_C) &\sim N_d(0, \Sigma_0), \end{aligned}$$

where $Y_t = (Y(s_1, t), \dots, Y(s_d, t))'$ and $P = \rho I_d$ is the $AR(1)$ transition matrix, $\Sigma_\varepsilon = \sigma_\varepsilon^2 I_d$ and $\Sigma_\omega = \sigma_\omega^2 C_\theta(\|s_i - s_j\|)_{i,j=1,\dots,d} = \sigma_\omega^2 C_\theta(\mathbf{h})$, with elements evaluated through the spatial correlation function appearing in (5). It follows that the posterior distribution is

$$\begin{aligned} f(\Psi_C | Z) &\propto (\sigma_\varepsilon^2)^{-\frac{dT}{2}} \exp \left[-\frac{1}{2\sigma_\varepsilon^2} \sum_{t=1}^T (Z_t - X_t\beta - Y_t)' (Z_t - X_t\beta - Y_t) \right] \\ &\times |\Sigma_0|^{-\frac{1}{2}} \exp \left[-\frac{1}{2} Y_0' \Sigma_0^{-1} Y_0 \right] \\ &\times (\sigma_\omega^2)^{-\frac{dT}{2}} |C_\theta(\mathbf{h})|^{-\frac{T}{2}} \exp \left[-\frac{1}{2\sigma_\omega^2} \sum_{t=1}^T (Y_t - PY_{t-1})' C_\theta(\mathbf{h})^{-1} (Y_t - PY_{t-1}) \right] \\ &\times \prod_{i=1}^{\dim(\Psi_C)} f(\Psi_{Ci}). \end{aligned}$$

Straightforward calculation yields a Gaussian full conditional distribution for β with mean AB' and covariance A where $A = \left(\Sigma_0^{-1} + \frac{1}{\sigma_\varepsilon^2} \sum_{t=1}^T X_t' X_t \right)^{-1}$ and $B = \frac{1}{\sigma_\varepsilon^2} \sum_{t=1}^T (Z_t' X_t - Y_t' X_t)$, assuming a Normal prior distribution for β , that is $\beta \sim N_k(0, \Sigma_0)$. Moreover, for the variance parameters we have the

following conditional posterior distributions:

$$\begin{aligned}\sigma_\omega^2 &\sim IG\left(\frac{dT}{2} + a, b + \frac{1}{2} \sum_{t=1}^T (Y_t - PY_{t-1})' C_\theta(\mathbf{h})^{-1} (Y_t - PY_{t-1})\right), \\ \sigma_\varepsilon^2 &\sim IG\left(\frac{dT}{2} + a, b + \frac{1}{2} \sum_{t=1}^T (Z_t - X_t\beta - Y_t)' (Z_t - X_t\beta - Y_t)\right),\end{aligned}$$

where a and b denote the hyperparameters of the corresponding Inverse-Gamma priors $IG(a, b)$. The parameters θ and ρ are estimated using the MH algorithm while latent process Y is sampled using the FFBS algorithm.

C Spatial prediction

Spatial prediction at a new location s_0 and time t_0 (with $1 \leq t_0 \leq T$) is based on the posterior predictive distribution of $z(s_0, t_0)$ which is given by

$$f(z(s_0, t_0) \mid Z) = \int f(z(s_0, t_0), u(s_0, t_0), \Psi \mid Z) du(s_0, t_0) d\Psi. \quad (19)$$

Note that for all the models described in Section 2 except Model A2, we can marginalize over U and (19) reduces to the following equation

$$f(z(s_0, t_0) \mid Z) = \int f(z(s_0, t_0), \Psi \mid Z) d\Psi = \int f(z(s_0, t_0) \mid Z, \Psi) f(\Psi \mid Z) d\Psi. \quad (20)$$

In practice, in order to obtain a prediction using MCMC methods, the posterior predictive distribution (20) is sampled by composition (Tanner, 1996). This means that we first draw from the posterior $f(\Psi \mid Z)$ which then allows us to draw from $f(z(s_0, t_0) \mid Z, \Psi)$. Therefore, to define $f(z(s_0, t_0) \mid Z, \Psi)$

we refer to the standard multivariate Normal theory. In fact, from the joint distribution of the data Z and $z(s_0, t_0)$, that is

$$\begin{pmatrix} Z \\ z(s_0, t_0) \end{pmatrix} \Big| \Psi \sim N_{dT+1} \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma'_{12} & \Sigma_{22} \end{pmatrix} \right],$$

it derives that

$$z(s_0, t_0) \mid Z, \Psi \sim N_1 \left(\mu_2 + \Sigma'_{12} \Sigma_{11}^{-1} (Z - \mu_1), \Sigma_{22} - \Sigma'_{12} \Sigma_{11}^{-1} \Sigma_{12} \right). \quad (21)$$

Note that Σ_{11} , the variance-covariance matrix of the data Z , has dimension $(dT \times dT)$ while Σ_{22} is a scalar. Moreover, the covariance vector Σ_{12} has dimension $(dT \times 1)$ and its generic element is $Cov(z(s_i, t), z(s_0, t_0))$ ($i = 1, \dots, d; t = 1, \dots, T$). In the following subsections we report schematically for each model the elements μ_1 , μ_2 , Σ_{11} , Σ_{22} and Σ_{12} of (21) referring to the notation introduced in Section B. Moreover, let $\tilde{\mathbf{h}}$ denote the vector of distances $\|s_i - s_0\|$ with $i = 1, \dots, d$, and $\tilde{\mathbf{l}}$ the vector of temporal lags $\|t - t_0\|$ with $t = 1, \dots, T$. Thus, for example, the term $C_\theta(\tilde{\mathbf{h}})$ is the correlation vector whose generic element is $C_\theta(\|s_i - s_0\|)$, with $i = 1, \dots, d$.

C.1 Model A1

For Model A1 it holds that $\mu_1 = X\beta$, where $X = \{X_1, \dots, X_T\}$ is the $(dT \times k)$ array of covariates, and Σ_{11} is a block diagonal matrix with blocks given by $\Sigma_{\omega+\varepsilon}$ and defined by (15). Moreover, $\mu_2 = x(s_0, t_0)\beta$, where $x(s_0, t_0)$ is the k -dimensional covariate vector observed at time t_0 at site s_0 , and $\Sigma_{22} = \sigma_\omega^2 + \sigma_\varepsilon^2$.

Finally, let us denote with $\mathbf{0}_{d \times 1}$ a d -dimensional row vector of zeros and with $C_\theta(\tilde{\mathbf{h}})$ the correlation vector. Hence the covariance vector Σ_{12} is defined by

$$\Sigma_{12} = \left(\underbrace{\mathbf{0}_{d \times 1}}_{t=1}, \dots, \underbrace{\sigma_\omega^2 C_\theta(\tilde{\mathbf{h}})}_{t=t_0}, \dots, \underbrace{\mathbf{0}_{d \times 1}}_{t=T} \right)'$$

where it can be seen that the d non-zero elements occur at the time point t_0 .

C.2 Model A2

For Model A2 it is not possible to marginalize over U and thus for defining the posterior predictive distribution we refer to (19) and rewrite it as follows:

$$f(z(s_0, t_0) | Z) = \int f(z(s_0, t_0) | u(s_0, t_0), \Psi) f(u(s_0, t_0) | U, \Psi) f(U, \Psi | Z) du(s_0, t_0) dU d\Psi$$

where $f(z(s_0, t_0) | u(s_0, t_0), \Psi) \sim N_1(u(s_0, t_0), \sigma_\varepsilon^2)$. To define $f(u(s_0, t_0) | U, \Psi)$ we consider the following joint distribution

$$\left(\begin{array}{c} U \\ u(s_0, t_0) \end{array} \middle| \Psi \right) \sim N_{dT+1} \left[\left(\begin{array}{c} X\beta \\ x(s_0, t_0)\beta \end{array} \right), \left(\begin{array}{cc} \Sigma_U = \sigma_\omega^2 C_\theta(\mathbf{h}, \mathbf{1}) & \Sigma_{12} \\ \Sigma'_{12} & \sigma_\omega^2 \end{array} \right) \right]$$

with the covariance vector Σ_{12} given by

$$\Sigma_{12} = \sigma_\omega^2 \left(C_{\theta_1}(|t - t_0|)_{t=1, \dots, T} \otimes C_{\theta_2}(\|s_i - s_0\|)_{i=1, \dots, d} \right) = \sigma_\omega^2 \left(C_{\theta_1}(\tilde{\mathbf{1}}) \otimes C_{\theta_2}(\tilde{\mathbf{h}}) \right).$$

Now it can be easily derived that

$$u(s_0, t_0) | U, \Psi \sim N_1 \left(x(s_0, t_0)\beta + \frac{\Sigma'_{12} C_\theta(\mathbf{h}, \mathbf{l})^{-1} (U - X\beta)}{\sigma_\omega^2}, \sigma_\omega^2 - \frac{\Sigma'_{12} C_\theta(\mathbf{h}, \mathbf{l})^{-1} \Sigma_{12}}{\sigma_\omega^2} \right). \quad (22)$$

Using composition sampling, once a value from (22) is drawn, it is possible to obtain a prediction generating a value from $f(z(s_0, t_0) | u(s_0, t_0), \Psi)$.

C.3 Model A3

For Model A3 we have that $\mu_1 = X\beta$, where $X = \{X_1, \dots, X_T\}$ is the $(dT \times k)$ array of covariates, and $\Sigma_{11} = \Sigma_{\omega+\varepsilon}$ is the variance-covariance matrix defined in Section B.3. Moreover, $\mu_2 = x(s_0, t_0)\beta$ and $\Sigma_{22} = \sigma_\omega^2 + \sigma_\varepsilon^2$. Finally, let us denote with $C_\theta(\tilde{\mathbf{h}}, |t - t_0|) = C_\theta(\|s_i - s_0\|, |t - t_0|)_{i=1, \dots, d}$ the $(d \times 1)$ -dimensional correlation vector at time t evaluated through the nonseparable correlation function $C_\theta(h, l)$ given in Eq.(7). It follows that the covariance vector Σ_{12} is given by

$$\Sigma_{12} = \sigma_\omega^2 \left(C_\theta(\tilde{\mathbf{h}}, |1 - t_0|), \dots, C_\theta(\tilde{\mathbf{h}}, |t - t_0|), \dots, C_\theta(\tilde{\mathbf{h}}, |T - t_0|) \right)'.$$

C.4 Model B

For Model B Eq.(19) needs to be rewritten as follows in order to consider Y :

$$f(z(s_0, t_0) | Z) = \int f(z(s_0, t_0) | y(s_0, t_0), \Psi) f(y(s_0, t_0) | Y, \Psi) f(Y, \Psi | Z) dy(s_0, t_0) d\Psi. \quad (23)$$

As described in Section B.4, the process Y for a given time point is constant over space so that $y(s_0, t_0) = y(t_0)$ for any s_0 , and $y(t_0)$ is esti-

mated using the FFBS algorithm. To draw a value from the distribution $f(z(s_0, t_0) | y(s_0, t_0), \Psi)$, we refer to Eq.([21](#)) with $\mu_1 = X\beta + Y$ (where $X = \{X_1, \dots, X_T\}$ and $Y = \{Y_1, \dots, Y_T\}$), $\mu_2 = x(s_0, t_0)\beta + y(t_0)$, $\Sigma_{22} = \sigma_\omega^2 + \sigma_\varepsilon^2$ and Σ_{11} is a block diagonal matrix with blocks given by $\Sigma_{\omega+\varepsilon}$ and defined by ([15](#)). Finally, the covariance vector Σ_{12} is given by

$$\Sigma_{12} = \left(\underbrace{\mathbf{0}_{d \times 1}}_{t=1}, \dots, \underbrace{\sigma_\omega^2 C_\theta(\tilde{\mathbf{h}})}_{t=t_0}, \dots, \underbrace{\mathbf{0}_{d \times 1}}_{t=T} \right)'$$

as in the case of Model A1 of Section [C.1](#).

C.5 Model C

With reference to Model C, Eq.([23](#)) still holds but the term $f(y(s_0, t_0) | Y, \Psi)$ has to be derived appropriately. To this purpose, consider the following joint distribution:

$$\left(\begin{array}{c} Y \\ y(s_0, t_0) \end{array} \middle| \Psi \right) \sim N_{dT+1} \left[\left(\begin{array}{c} \rho \tilde{Y} \\ \rho y(s_0, t_0 - 1) \end{array} \right), \left(\begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma'_{12} & \sigma_\omega^2 \end{array} \right) \right]$$

where $\tilde{Y} = \{Y_0, \dots, Y_{T-1}\}$, the variance-covariance matrix of Y is a block diagonal matrix with blocks given by $\Sigma_\omega = \sigma_\omega^2 C_\theta(\tilde{\mathbf{h}})$, and the covariance vector Σ_{12} is

$$\Sigma_{12} = \left(\underbrace{\mathbf{0}_{d \times 1}}_{t=1}, \dots, \underbrace{\sigma_\omega^2 C_\theta(\tilde{\mathbf{h}})}_{t=t_0}, \dots, \underbrace{\mathbf{0}_{d \times 1}}_{t=T} \right)',$$

as in the case of Model A1 and of Model B. Thus, it can be easily derived that

$$y(s_0, t_0) \mid Y, \Psi \sim N_1 \left(\rho y(s_0, t_0 - 1) + \Sigma'_{12} \Sigma_{11}^{-1} \left(Y - \rho \tilde{Y} \right), \sigma_\omega^2 - \Sigma'_{12} \Sigma_{11}^{-1} \Sigma_{12} \right). \quad (24)$$

Using composition sampling for the posterior predictive distribution (23), once a value from (24) is drawn, we can obtain a prediction generating a value from $f(z(s_0, t_0) \mid y(s_0, t_0), \Psi)$.

	ψ	φ
Model A3-1	$\psi(x) = (ax^\alpha + b)/(b(ax^\alpha + 1))$	$\varphi(x) = \exp(-cx^\gamma)$
Model A3-2	$\psi(x) = (ax^\alpha + 1)^\tau$	$\varphi(x) = (1 + cx^\gamma)^{-\nu}$

Table 1: Function $\psi(x)$ and $\varphi(x)$ that define the nonseparable spatio-temporal correlation function of Model A3-1 and A3-2.

	Model A1		Model A2		Model A3-1	
Intercept	3.881	(3.815; 3.946)	3.938	(3.868; 4.003)	3.935	(3.848; 4.021)
A	-0.155	(-0.169; -0.140)	-0.142	(-0.220; -0.062)	-0.162	(-0.195; -0.128)
UTMX	-0.102	(-0.124; -0.080)	-0.110	(-0.197; -0.027)	-0.112	(-0.155; -0.069)
UTMY	-0.079	(-0.097; -0.061)	-0.072	(-0.121; -0.015)	-0.076	(-0.108; -0.044)
WS	-0.091	(-0.109; -0.073)	-0.078	(-0.107; -0.054)	-0.086	(-0.101; -0.072)
HMIX	-0.041	(-0.047; -0.035)	-0.075	(-0.101; -0.046)	-0.050	(-0.056; -0.044)
TEMP	-0.262	(-0.309; -0.217)	-0.103	(-0.166; -0.051)	-0.138	(-0.171; -0.103)
PREC	-0.085	(-0.107; -0.062)	-0.105	(-0.127; -0.082)	-0.090	(-0.105; -0.076)
EMI	0.063	(0.053; 0.073)	0.083	(0.038; 0.133)	0.054	(0.031; 0.076)
	Model A3-2		Model B		Model C	
Intercept	3.928	(3.853; 4.001)	3.955	(3.773; 4.119)	3.956	(3.672; 4.197)
A	-0.157	(-0.187; -0.127)	-0.172	(-0.186; -0.157)	-0.152	(-0.187; -0.118)
UTMX	-0.108	(-0.145; -0.071)	-0.113	(-0.130; -0.095)	-0.092	(-0.173; 0.001)
UTMY	-0.074	(-0.102; -0.047)	-0.073	(-0.086; -0.060)	-0.129	(-0.196; -0.059)
WS	-0.090	(-0.105; -0.076)	-0.085	(-0.100; -0.071)	-0.032	(-0.052; -0.012)
HMIX	-0.050	(-0.056; -0.044)	-0.042	(-0.048; -0.036)	-0.024	(-0.041; -0.007)
TEMP	-0.128	(-0.165; -0.091)	-0.308	(-0.349; -0.267)	-0.218	(-0.296; -0.119)
PREC	-0.093	(-0.107; -0.078)	-0.090	(-0.106; -0.074)	-0.040	(-0.060; -0.019)
EMI	0.058	(0.037; 0.079)	0.061	(0.050; 0.072)	0.049	(0.030; 0.067)

Table 2: Posterior estimates (mean and 95% credible interval in brackets) of the covariate coefficient vector β .

	Model A1	Model A2	Model A3-1
σ_ω^2	0.237 (0.198; 0.283)	0.247 (0.229; 0.267)	0.077 (0.070; 0.083)
σ_ε^2	0.071 (0.066; 0.077)	0.015 (0.013; 0.018)	0.034 (0.028; 0.039)
	Model A3-2	Model B	Model C
σ_ω^2	0.077 (0.071; 0.084)	0.063 (0.053; 0.073)	0.950 (0.822; 1.091)
σ_ε^2	0.034 (0.029; 0.040)	0.051 (0.042; 0.060)	0.013 (0.011; 0.015)
σ_η^2	— —	0.065 (0.052; 0.081)	— —

Table 3: Posterior estimates (mean and 95% credible interval in brackets) of the variance parameters.

	Model A1	Model A2
θ	0.0033 (0.0025; 0.0042)	— —
θ_1	— —	0.492 (0.473; 0.511)
θ_2	— —	0.032 (0.029; 0.035)
	Model B	Model C
θ	0.050 (0.040; 0.061)	0.0022 (0.0019; 0.0024)
ρ	0.831 (0.738; 0.919)	0.654 (0.629; 0.677)

Table 4: Posterior estimates (mean and 95% credible interval in brackets) of the correlation function parameters for Model A1, A2, B and C.

	Model A3-1	Model A3-2
a	0.058 (0.045; 0.074)	4.088 (3.533; 4.708)
α	0.736 (0.650; 0.831)	0.800 (0.743; 0.859)
b	0.047 (0.036; 0.058)	— —
c	0.549 (0.448; 0.659)	0.982 (0.823; 1.158)
γ	0.110 (0.089; 0.132)	0.184 (0.156; 0.212)
ν	— —	0.801 (0.743; 0.864)
τ	— —	0.4756 (0.434; 0.517)

Table 5: Posterior estimates (mean and 95% credible interval in brackets) of the correlation parameters for Model A3-1 and A3-2.

Criteria	Model					
	A1	A2	A3-1	A3-2	B	C
N. of parameters (β 's excluded)	3	4	7	8	5	4
– estimated by MH	3	2	7	8	4	2
Size of the biggest matrix to be inverted	$(d \times d)$	$(T \times T)$	$(dT \times dT)$	$(dT \times dT)$	$(d \times d)$	$(d \times d)$
Estimation time (sec per iteration)	0.014	3.992	45.299	47.466	0.063	0.360
Prediction time (sec per iteration)	0.026	27.288	17.733	20.367	0.024	0.038
Prediction capability	**	*	**	**	**	**

Table 6: Performance and complexity of the models.