Atmos. Chem. Phys., 11, 5365–5382, 2011
www.atmos-chem-phys.net/11/5365/2011/
doi:10.5194/acp-11-5365-2011

Atmospheric Chemistry and Physics

Bayesian statistical modeling of spatially correlated error structure in atmospheric tracer inverse analysis

C. Mukherjee[1], **P. S. Kasibhatla**[2], **and M. West**[1]

[1]Department of Statistical Science, Duke University, Durham, NC, USA
[2]Nicholas School of the Environment, Duke University, Durham, NC, USA

Received: 1 January 2011 – Published in Atmos. Chem. Phys. Discuss.: 19 January 2011
Revised: 22 May 2011 – Accepted: 26 May 2011 – Published: 9 June 2011

Abstract. We present and discuss the use of Bayesian modeling and computational methods for atmospheric chemistry inverse analyses that incorporate evaluation of spatial structure in model-data residuals. Motivated by problems of refining bottom-up estimates of source/sink fluxes of trace gas and aerosols based on satellite retrievals of atmospheric chemical concentrations, we address the need for formal modeling of spatial residual error structure in global scale inversion models. We do this using analytically and computationally tractable conditional autoregressive (CAR) spatial models as components of a global inversion framework. We develop Markov chain Monte Carlo methods to explore and fit these spatial structures in an overall statistical framework that simultaneously estimates source fluxes. Additional aspects of the study extend the statistical framework to utilize priors on source fluxes in a physically realistic manner, and to formally address and deal with missing data in satellite retrievals. We demonstrate the analysis in the context of inferring carbon monoxide (CO) sources constrained by satellite retrievals of column CO from the Measurement of Pollution in the Troposphere (MOPITT) instrument on the TERRA satellite, paying special attention to evaluating performance of the inverse approach using various statistical diagnostic metrics. This is developed using synthetic data generated to resemble MOPITT data to define a proof-of-concept and model assessment, and then in analysis of real MOPITT data. These studies demonstrate the ability of these simple spatial models to substantially improve over standard non-spatial models in terms of statistical fit, ability to recover sources in synthetic examples, and predictive match with real data.

Correspondence to: C. Mukherjee
(chiranjit@stat.duke.edu)

1 Introduction

1.1 Model and inference setting

Bayesian statistical techniques are increasingly being used in atmospheric chemistry inverse modeling studies to refine bottom-up trace gas and aerosol source/sink flux estimates. Past inverse studies have generally focused on analysis of surface and airborne in situ measurements from geographically distributed sites for species such as CO_2, CO, CH_4, (e.g., Enting et al., 1995; Hein et al., 1997; Houweling et al., 1999; Bergamaschi et al., 2000; Bousquet et al., 2000, 2006; Gurney et al., 2002, 2003; Kasibhatla et al., 2002; Petron et al., 2002; Peylin et al., 2002; Gerbig et al., 2003; Palmer et al., 2003; Rodenbeck et al., 2003; Fletcher et al., 2004; Michalak et al., 2004; Patra et al., 2005; Rayner et al., 2005; Baker et al., 2006; Mueller et al., 2008; Gourdji et al., 2010). More recently, inverse studies based on synthetic and real satellite retrievals of tropospheric trace gas concentration fields have identified the potential for these new measurements to further improve our understanding of trace gas fluxes at regional and sub-regional scales (e.g., Rayner and O'Brien, 2001; Jones et al., 2003; Arellano et al., 2004, 2006; Heald et al., 2004; Houweling et al., 2004; Petron et al., 2004; Chevallier et al., 2005a,b, 2007, 2009a,b; Stavrakou and Mueller, 2006; Meirink et al., 2008; Feng et al., 2009; Kopacz et al., 2009, 2010). To fully exploit the information content in these high-dimensional, spatially-dense satellite data sets, we must address questions about the nature of spatial dependencies among observations that are not predicted by existing models, and how to appropriately integrate spatial dependencies to ensure robust and unbiased inverse analyses. We show here how we can address both modeling and computational issues via Bayesian analysis of conditional autoregressive (CAR) spatial models to characterize spatial observation error fields.

The general aims of the paper are to introduce and illustrate CAR models in this context, discussing computational implementations and giving examples in analyses of synthetic and real data. Our model and analysis have additional practically relevant features, including the use of constrained priors on source fluxes, and the overall Bayesian analysis defines inference on spatial dependencies and variance parameters as well as fluxes. We regard this work as a proof-of-concept in illustrating the ability to fit, assess and compare initial classes of spatial models in this context, and – using formal statistical methods – to highlight the nature and extent of improvements over non-spatial models using synthetic and real data. While the class of CAR spatial models used here is simple and relatively limited in scope for modeling very diverse spatial patterns, it represents a first step towards more elaborate and adaptive models that may become more relevant with inverse analysis at higher resolutions and in the context of access to increasingly rich satellite data.

We begin with the canonical model

$$\boldsymbol{y} = \mathbf{K}\boldsymbol{x} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{y}$ is a $m \times 1$ vector of atmospheric concentration measurements for a particular species, $\boldsymbol{x}$ is a $n \times 1$ vector of corresponding fluxes with individual elements x_i representing source/sink categories (e.g. fossil-fuel combustion, biomass burning, etc.) and/or geographical regions, and $\mathbf{K}$ is a $m \times n$ Jacobian matrix derived from an atmospheric chemistry transport model (CTM) and describing the relationships between discretized atmospheric concentrations and fluxes corresponding to source/sink categories. The random $m \times 1$ vector $\boldsymbol{\epsilon}$ accounts for errors associated with the measurement technique, the chemical transport model, as well as representativeness errors arising from differences in resolution between the measurements and model calculated concentration fields. The vast majority of inverse modeling applications in atmospheric chemistry are based on this formulation under the assumption of linearity of atmospheric transport for unreactive species such as CO_2, and with additional linearizing assumptions with regards to chemistry for reactive species such as CO and CH_4. We also note that while the atmospheric concentration measurements are spatially and temporally resolved, the vector $\boldsymbol{y}$ is constructed by stacking measurements indexed by CTM grid cells and time. Global-scale atmospheric chemistry inverse modeling studies involving real or synthetic satellite retrievals have generally focused on analyzing monthly or weekly mean measurements that are spatially aggregated to the CTM grid resolution (typically 200–500 km in the horizontal). Satellite atmospheric concentration retrievals typically consist of vertically averaged information which can be accounted for in Eq. (1) by appropriately modifying $\mathbf{K}$ based on the specific instrument characteristics. Past studies have generally focused on estimating regionally-and monthly-aggregated fluxes, though there is increasing interest in estimating fluxes at higher temporal and spatial resolution. As a result, $m \approx 10^5$–10^6 and $n \approx 10^2$–10^5 for applications involving a year's worth of data.

Bayesian analysis generates the posterior density function $p(\boldsymbol{x} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{x})\, p(\boldsymbol{x})$ using the likelihood function $p(\boldsymbol{y}|\boldsymbol{x})$ induced by the model (Eq. 1) and a specified prior p.d.f. $p(\boldsymbol{x})$. In the context of atmospheric tracer inverse modeling, $p(\boldsymbol{x})$ represents prior knowledge of fluxes from independent, bottom-up estimates. In atmospheric chemistry inverse modeling applications it has typically been assumed that $p(\boldsymbol{x})$ and $p(\boldsymbol{\epsilon})$ are multivariate normal distributions, defined by $\boldsymbol{x} \sim N(\boldsymbol{x}_\mathrm{a}, \mathbf{S}_\mathrm{a})$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{S}_\epsilon)$ where $\boldsymbol{x}_\mathrm{a}$ and $\mathbf{S}_\mathrm{a}$ are the prior mean vector and covariance matrix for $\boldsymbol{x}$, and $\mathbf{S}_\epsilon$ is the observation error covariance matrix. Then, for known $\boldsymbol{x}_\mathrm{a}$, $\mathbf{S}_\mathrm{a}$ and $\mathbf{S}_\epsilon$, inferences are defined by the resulting posterior $(\boldsymbol{x}|\boldsymbol{y}) \sim N(\boldsymbol{x}_\mathrm{p}, \mathbf{S}_\mathrm{p})$ where

$$\begin{aligned}
&\boldsymbol{x}_\mathrm{p} = \boldsymbol{x}_\mathrm{a} + \left(\mathbf{K}'{\mathbf{S}_\epsilon}^{-1}\mathbf{K} + {\mathbf{S}_\mathrm{a}}^{-1}\right)^{-1}\mathbf{K}'{\mathbf{S}_\epsilon}^{-1}(\boldsymbol{y} - \mathbf{K}\boldsymbol{x}_\mathrm{a}) \text{ or} \\
&\boldsymbol{x}_\mathrm{p} = \boldsymbol{x}_\mathrm{a} + \mathbf{G}(\boldsymbol{y} - \mathbf{K}\boldsymbol{x}_\mathrm{a}),\ {\mathbf{S}_\mathrm{p}}^{-1} = \mathbf{K}'{\mathbf{S}_\epsilon}^{-1}\mathbf{K} + {\mathbf{S}_\mathrm{a}}^{-1} \text{ or} \\
&\mathbf{S}_\mathrm{p} = (\mathbf{I} - \mathbf{G}\mathbf{K})\mathbf{S}_\mathrm{a},
\end{aligned} \tag{2}$$

where $\mathbf{G} = \mathbf{S}_\mathrm{a}\mathbf{K}'(\mathbf{K}\mathbf{S}_\mathrm{a}\mathbf{K}' + \mathbf{S}_\epsilon)^{-1}$. Here $\boldsymbol{x}_\mathrm{p}$ and $\mathbf{S}_\mathrm{p}$ are the posterior mean and covariance, respectively. The Gaussian, linear assumptions underlie analytic tractability in defining the closed form posterior here, as well as extensions to time series of data (e.g., Prado and West, 2010), so continue to be important in enabling applications.

With some exceptions noted in Sect. 1.2, previous applications in atmospheric chemistry have generally assumed a diagonal structure for $\mathbf{S}_\epsilon$ due to the lack of effective and computationally efficient approaches to identifying and integrating relevant spatial structures. This eliminates the computational burden associated with the calculation of the matrix inverse of $\mathbf{S}_\epsilon$ in Eq. (2). The equations clearly show, however, that if spatial dependencies in the model errors exist and can be captured by a relevant non-diagonal and structured covariance matrix $\mathbf{S}_\epsilon$, this will impact on the posterior estimates $\boldsymbol{x}_\mathrm{p}$ of fluxes as well as the associated measures of uncertainties in $\mathbf{S}_\mathrm{p}$. The impact can be substantial as demonstrated by some earlier studies (e.g., Chevallier, 2007) and our examples below.

1.2 Application context and previous approaches

While the assumption of uncorrelated observational errors may be reasonable for inverse studies based on surface measurements from a limited number of geographically scattered locations, it is increasingly untenable for geographically dense satellite measurements. This is especially true when mid-and upper tropospheric tracer concentrations (where transport is relatively fast) contribute disproportionately, relative to surface and lower tropospheric concentrations, to satellite weighted column-average retrievals.

A few global scale inverse modeling studies have considered observation error correlations, and their impact on posterior flux estimates. Chevallier (2007) considered the

problem of estimating CO_2 fluxes from synthetic OCO measurements with known error correlations, and demonstrated that posterior source estimates and the corresponding uncertainties are sensitive to the treatment of observation error correlations in the inverse analysis. This study also showed that the effect of neglecting observation error correlations in the inverse analysis can be partially compensated for by techniques such as observation thinning and error variance inflation, but at the expense of not fully utilizing the information content of the measurements.

To avoid this loss of information, one can attempt to explicitly account for observation error correlations in the inverse analysis. For example, it has been proposed that spatial correlations associated with the CTM-component of the observation error can be approximated by the spatial error covariance structure from pairs of short-term chemical forecast simulations with different forecast starting times (Jones et al., 2003). This approach to characterizing the CTM forecast error is intuitively appealing, but suffers from the drawback that running multiple forward chemical model simulations over seasonal and inter-annual time scales is computationally expensive. It is also often impractical to perform simulations with independent CTMs on a routine basis to more fully characterize chemical model transport errors.

An alternative approach involves statistically modeling spatial observation error structures, and determining the associated parameters of the statistical error model as part of the inverse analysis. A key challenge in this context is computation. Traditional spatial modeling utilizing standard Gaussian processes based on a spatial distance-based correlation function (e.g., Rue and Held, 2005) have been explored to a degree (e.g., Michalak et al., 2004; Mueller et al., 2008; Gourdji et al., 2010) in terms of characterizing spatial error structure in the prior. In the context of modeling observation error structures in a fully Bayesian framework, this approach is computationally severely limited due to the resulting needs to perform multiple matrix inversions on covariance matrices of order m. Our interest in exploring alternatives that do not involve approximation short-cuts addresses the scale-up issues head-on while evaluating the flexibility of the class of conditional autoregressive spatial models.

2 Statistical modeling developments

2.1 Overview

We discuss an approach that uses alternative spatial structures that (a) recognize and exploit the fact that the data is inherently grid-based, and (b) provide access to effective statistical computation using Bayesian simulation methods, specifically Markov chain Monte Carlo (MCMC) analysis (e.g., Gelman et al., 2004; Prado and West, 2010, chapter 1). We show how this allows direct and appropriate modeling of spatial dependencies in observation errors in analysis that integrates evaluation of the spatial field structure together with inference on source fluxes. Our analysis involves additional modeling advances for the inverse problem that include use of non-normal priors for fluxes to properly reflect the fact that the sources of interest in this study are positive, and the integration of missing data analysis to account for and infer missing retrievals. We further note that computer code (in Matlab) for all our analyses reported is available for others to explore and use.

Models of spatial structure in $\mathbf{S}_\epsilon$ involve additional unknown parameters that define the spatial dependencies; denote these by $\boldsymbol{\theta}$. Further, since satellite retrievals are subject to substantial missing data we explicitly recognize this; we denote by M the set of indices of missing retrievals, $M \subset \{1:m\}$, while H is the set of indices for observed retrievals. Thus the observed data is $\boldsymbol{y}_H$ and the missing data $\boldsymbol{y}_M$. Then, for a given prior $p(\boldsymbol{x})$, the formal Bayesian inference problem is to compute and summarize the posterior $p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y}_M | \boldsymbol{y}_H)$. We do this using custom development of standard Bayesian statistical simulation methods based on MCMC; some summary aspects are mentioned here and in the Appendix, with full technical details provided in the supplementary material on statistical computation.

2.2 Spatial error structure: conditional autoregressive (CAR) model formulation

An approach based on Gaussian conditional autoregressive (CAR) spatial models is able, as we show, to define realistic and appropriate spatial structures for geographically dense satellite retrieval data on a lattice, while leading to a computationally tractable methodology for atmospheric tracer inverse modeling. The approach takes advantage of the fact that, under certain conditions, it is possible to statistically model the precision matrix $\mathbf{S}_\epsilon^{-1}$ as a very sparse matrix defined by a very small number of parameters, and that these parameters can be efficiently inferred using MCMC algorithms.

In the basic model of Eq. (1), $\boldsymbol{y}$ represents the vectorized set of retrievals from the original global rectangular lattice, or grid. Suppose that $\boldsymbol{y}$ represents the vector of retrievals for a single month, and that $\boldsymbol{\epsilon}^{\mathrm{CAR}}$ (the notation change is to explicitly reflect the assumption of a CAR spatial structure) refers to the corresponding errors. Specification of a CAR model starts with a $m \times m$ proximity matrix, $\mathbf{W}$, that designates weights to the neighbors for each grid cell. In this application, we define the elements of $\mathbf{W}$ as

$$w_{ij} = \begin{cases} \exp(-\delta_{ij}) \text{ if cell } i \text{ and } j \text{ are neighbors}, \\ 0 \text{ if } i = j, \\ 0 \text{ otherwise}, \end{cases} \quad (3)$$

where $\delta_{ij} \geq 0$ measures distance between the centroids of grid cells i, j; with $(\mathrm{lat}_i, \mathrm{long}_i)$ representing centroid of cell i, this is given by $\delta_{ij}^2 = (\mathrm{lat}_i - \mathrm{lat}_j)^2 + (\mathrm{long}_i - \mathrm{long}_j)^2$.

The CAR model introduces spatial dependencies through the complete conditional distributions for all elements of $\boldsymbol{\epsilon}^{\mathrm{CAR}}$; the error in cell i depends on its neighbors as in

$$\epsilon_i^{\mathrm{CAR}} \big| \{\epsilon_j^{\mathrm{CAR}},\ j \neq i\} \sim N\Big[\sum_{j \neq i} (\rho\, w_{ij}/w_{i+})\, \epsilon_j^{\mathrm{CAR}}, \tau_c^2/w_{i+} \Big], \quad \text{for } i = 1,2,\ldots,m, \tag{4}$$

and where $w_{i+} = \sum_{j=1}^{m} w_{ij}$. It is clear from the forms of these conditional distributions that the spatial dependence parameter ρ defines association via the implied linear regression of each cell on its neighbors. In tandem, the scale parameter τ_c controls levels of variation in these conditional distributions. Global correlations between separated grid cells are induced as a result; cell i depends on a more distant cell k transitively through its neighbors that link to neighbors of cell k, for example. Both parameters ρ, τ_c are to be estimated.

In the examples in this paper, we adopt a first-order neighborhood approach in which the 8 cells that are physical neighbors (N, NE, E, SE, S, SW, W, NW) of grid cell i are neighbors in the model, and only those. It turns out that this first-order dependence structure is capable of capturing spatial patterns sufficient to reflect much of the residual dependency we see in MOPITT data, though more elaborate neighborhoods could be examined using the same Bayesian approach; the changes would simply use a different proximity matrix.

Define the $m \times m$ matrix $\mathbf{D}_w = \mathrm{diag}(w_{1+}, w_{2+}, \ldots, w_{m+})$ and the spatial precision matrix $\mathbf{U} = \tau_c^{-2}(\mathbf{D}_w - \rho\mathbf{W})$. It follows from the specification of the CAR model that the joint distribution of all m error values is

$$\boldsymbol{\epsilon}^{\mathrm{CAR}} \sim N(\mathbf{0}, \mathbf{U}^{-1}). \tag{5}$$

That is, the error covariance matrix $\mathbf{S}_\epsilon$ is replaced by the spatially structured CAR covariance matrix $\mathbf{U}^{-1}$ that has non-zero pairwise correlations between cells over larger distances induced by the local neighborhood dependencies even though $\mathbf{U}$ itself has zero entries between cells that are not neighbors. Model fitting and inference relies on posterior estimation of $\mathbf{U}$ through estimation of $\boldsymbol{\theta} = (\rho, \tau_c)$ as uncertain parameters. Some of the computational tractability in dealing with the spatial structure as an ingredient of the inverse analysis comes through the fact that the precision matrix $\mathbf{U}$ is sparse; i.e., the elements U_{ij} are non-zero only when cells i, j are neighbors.

CAR models are capable of representing spatial structure that has traditionally been modeled via spatial distance-based correlation functions, referred to in the statistical literatures as Gaussian processes (GP) (e.g., Rue and Held, 2005). One of several often used forms is the exponential kernel in which the correlation between grid cells i, j is $\exp(-d_{ij}/L)$ where d_{ij} is the great circle distance between the centroids of the cells and L is the range parameter. The appropriate way to compare with CAR models is to look at the conditional regression coefficients implied by such a GP model in comparison to those that are used to define the CAR model, simply the $\rho w_{ij}/w_{i+}$ term of Eq. (4). Figure 1 shows such a comparison for several relevant values of L, each plotted side-by-side with a corresponding CAR model, on a 5×5 grid beyond which the GP coefficients are negligible. This ability to adequately match local spatial patterns is a general feature of CAR models while its computational accessibility makes it a clearly dominant choice over GP models for all but very small problems.

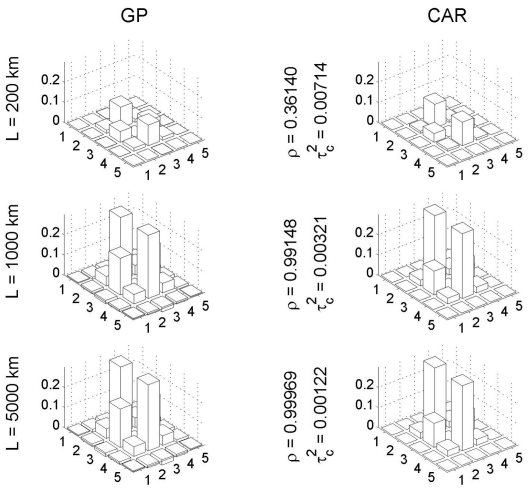

Fig. 1. Left: conditional regression coefficients from Gaussian process models (GP) with exponential decay correlation kernel $\exp(-d/L)$ for several values of the range length L. Here d is the centroid-centroid distance between cells and the regression coefficients plotted are those for the regression of the central cell (3,3) on the rest. The mesh size is not constant over the entire lattice; assuming the mean radius of the Earth to be 6371 km, the average size of a 4° latitude × 5° longitude cell is 445 km × 556 km. Right: conditional regression coefficients from CAR model using ρ values fitted to match the regressions in the corresponding GP model.

Finally, note that we may aggregate measurements over a series of time epochs (e.g. from multiple months). In this case, the variance matrix of the extended $\boldsymbol{\epsilon}^{\mathrm{CAR}}$ will be block diagonal, with the number of diagonal blocks $\mathbf{U}^{-1}$ equal to the number of time epochs considered in the analysis.

2.3 Prior specification for fluxes

To date, inverse applications in atmospheric chemistry have relied heavily on the linear-Gaussian theory associated with multivariate normal priors over $\boldsymbol{x}$ and the resulting analytic closed-form posterior summaries in Eq. (2). However, the fluxes of interest are often strictly non-negative, as for example, when CO sources are being estimated. In exploring posterior inferences using the traditional normal priors, we routinely encounter posterior densities that give appreciable

probability to negative flux values; this is a purely technical issue as, in such cases, the data:prior synthesis is surely consistent with very low or zero flux, whereas the mathematical assumption of normal priors leads to unconstrained posteriors that support practically meaningless negative values.

Modern Bayesian computational methods allow us to do away with such physically unrealistic assumptions that have historically been made for purely mathematical tractability reasons. Here we simply adapt the usual assumptions and utilize priors for strictly positive fluxes that are the usual normal distributions but now truncated at some small chosen lower bound to ensure scientifically relevant posterior inferences that disallow negative values. The traditional normal prior specification based on bottom-up fluxes $\boldsymbol{x}_a = (x_{\mathrm{a},i})$ has been to adopt a multivariate normal prior with $\mathbf{S}_\mathrm{a}$ diagonal and having ith diagonal variance element $S_{\mathrm{a},i}$ for each source $i = 1,\ldots,n$. In particular, we define $S_{\mathrm{a},i} = c_\mathrm{a}^2 x_{\mathrm{a},i}^2$ based on a specified coefficient of variation c_a. The direct modification to ensure non-negative fluxes above a lower value is to take the prior as a product of independent priors over sources $p(x_i)$ where each is defined by

$$x_i \sim N(m_{\mathrm{a},i}, v_{\mathrm{a},i})\ I(x_i > t_i), \quad i = 1,\ldots,n. \tag{6}$$

Here $I(\cdot)$ is the indicator function,

$$I(x_i > t_i) = \begin{cases} 1 & \text{if } x_i > t_i, \\ 0 & \text{if } x_i \le t_i, \end{cases}$$

and t_i is a pre-specified small lower bound on realistic flux levels.

Given the prior flux estimates $x_{\mathrm{a},i}$, and prior flux variance $S_{\mathrm{a},i}$ we can numerically match the expectation and variance of the prior distribution given by Eq. (6) with $x_{\mathrm{a},i}$ and $S_{\mathrm{a},i}$ to determine the values of the required prior parameters $m_{\mathrm{a},i}$ and $v_{\mathrm{a},i}$. In the examples below, we use this specification with the lower bound on fluxes defined as $t_i = x_{\mathrm{a},i}/4$.

2.4 Accounting for missing retrieval data

Satellite retrievals are inherently subject to missing data. This translates into an index set M for the CO retrievals that are missing, while those indexed in the set H are recorded. Writing the observed data sub-vector as $\boldsymbol{y}_H$ and the missing data sub-vector as $\boldsymbol{y}_M$, we include the information that $\boldsymbol{y}_M$ is missing in the analysis. This is done in standard Bayesian fashion: $\boldsymbol{y}_M$ is included as part of the inference problem in an extended analysis that computes and summarizes aspects of the posterior $p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y}_M \mid \boldsymbol{y}_H)$.

The missing rows of the transport matrix $\mathbf{K}_{(i,*)}$ for each $i \in M$ are linearly interpolated using the neighboring grid cells (since instrument characteristics required to calculate the corresponding elements of $\mathbf{K}$ are unavailable) and the corresponding unknowns y_i, $i \in M$ are assigned values that are repeatedly updated via simulations from the relevant conditional posterior predictive distributions in the Bayesian MCMC analysis, noted in the next section and detailed further in the Appendix.

2.5 Bayesian computation

Iterative posterior simulation using MCMC has been the de facto standard in the statistical community for Bayesian statistical computation for some years (Gelman et al., 2004). Recalling that $\boldsymbol{\theta}$ stands for unknown parameters in the error variance $\mathbf{S}_\epsilon$, the full joint posterior distribution for all unknowns $\{\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y}_M\}$ conditional on $\boldsymbol{y}_H$ is evaluated by simulating a large Monte Carlo sample, and basing inferences on numerical summaries of that sample. MCMC analysis performs this simulation iteratively, successively updating each of the unknowns by simulation from a relevant conditional distribution that may involve some of the most recently simulated, or imputed, values of other unknowns. Our MCMC strategy for the current context is outlined in the Appendix, with further technical details provided in the supplemental documentation.

3 Synthetic data studies

We first demonstrate the approach with an extensive set of synthetic data analyses that parallel the problem of estimating CO sources from MOPITT data considered by Arellano et al. (2004). We utilize synthetic data in order to provide a context where the true sources $\boldsymbol{x}$ are known, and to compare the CAR spatial model with a non-spatial (NS) statistical model to evaluate the effect of neglecting "true" spatial error correlations on the inverse source estimates. The NS model is a special case of the CAR model obtained when $\rho = 0$; in that special case, we denote the scale parameter by τ_n rather than τ_c. We pay special attention to evaluating performance of the inverse approach using various statistical diagnostic metrics.

3.1 Generation of synthetic data with spatially correlated errors

The inverse problem (Arellano et al., 2004) consists of using Level 2 V3 MOPITT daytime column CO retrievals from April–December 2000 to estimate annual CO emissions for $n = 15$ source categories consisting of: (a) fossil fuel/biofuel (FFBF) combustion in 7 geographical regions, (b) biomass burning (BIOM) in 7 geographical regions, and (c) and oxidation of biogenic isoprene and monoterpenes on a global-scale (BIOG). The geographical extent of each of the FFBF and BIOM regions is shown in Fig. 2. CO production from methane oxidation is not estimated as part of the inversion, but is taken into account by pre-subtracting its contribution to the MOPITT retrievals. The Jacobian matrix $\mathbf{K}$ is constructed by applying MOPITT averaging kernels to gridded CO fields calculated using an offline,

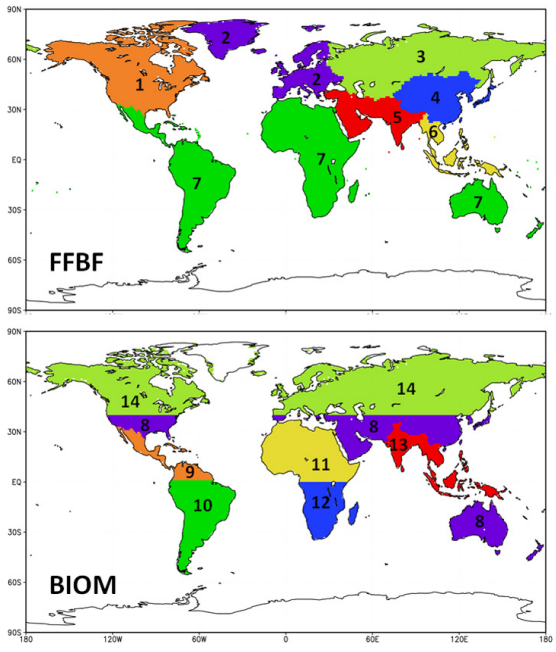

Fig. 2. Color-coded definition of fossil-fuel (FFBF) and biomass burning (BIOM) CO source regions (after Arellano et al., 2004). Numbers represent the source category number used in the text: 1 = FFBF North America (FFBF-NAM); 2 = FFBF Europe (FFBF-EUR); 3 = FFBF Russia (FFBF-RUS); 4 = FFBF East Asia (FFBF-EAS); 5 = FFBF South Asia (FFBF-SAS); 6 = FFBF Southeast Asia (FFBF-SEA); 7 = FFBF Rest of the World (FFBF-ROW); 8 = BIOM Other (BIOM-OTH); 9 = BIOM Northern Latin America (BIOM-NLA); 10 = BIOM Southern Latin America (BIOM-SLA); 11 = BIOM Northern Africa (BIOM-NAF); 12 = BIOM Southern Africa (BIOM-SAF); 13 = BIOM South and Southeast Asia (BIOM-SSA); 14 = BIOM Boreal (BIOM-BOR). Source category 15 (not shown here) represents the global source of CO from biogenic hydrocarbon oxidation (BIOG).

tagged tracer version of the GEOS-Chem CTM at a resolution of 4×5 degrees. The gridded, quality controlled MOPITT dataset used in the original analysis spans 50° N–50° S, yielding a lattice of 26×72 grid cells that vectorizes to 1872 observations on a monthly basis. Combining data from April–December 2000 yields a complete retrievals vector $\boldsymbol{y}$ of length $m = 16\,848$ and $\mathbf{K}$ of dimension $16\,848 \times 15$. The original analysis of Arellano et al. (2004) used only those MOPITT retrievals satisfying certain quality control metrics. We modified these to require at least 5 days of observations each month for a site to be considered valid; sites not meeting this are those treated as having missing data, yielding 139 missing values. Further details on MOPITT data processing and construction of $\mathbf{K}$ are given in Arellano et al. (2004); they also give the prior source vector $\boldsymbol{x}_\mathrm{a}$ and the diagonal matrix $\mathbf{S}_\mathrm{a}$ with ith diagonal element $S_{\mathrm{a},i} = c_\mathrm{a}^2 x_{\mathrm{a},i}^2$ for a constant coefficient of variation $c_\mathrm{a} = 0.5$. We use these values as the basis for positively constrained priors $x_i \sim N(m_{\mathrm{a},i}, v_{\mathrm{a},i}) I(x_i > t_i)$ taking $t_i = x_{\mathrm{a},i}/4$ such that the prior mean of each x_i is the specified bottom-up value $x_{\mathrm{a},i}$ and the prior variance is $S_{\mathrm{a},i}$.

We generate a single synthetic MOPITT CO retrieval data set with spatially-correlated errors by first generating a "true" source vector, $\tilde{\boldsymbol{x}}$, simply sampling from the truncated normal priors; we use $\tilde{}$ to denote the synthetic quantities throughout. We next construct a "true" error covariance matrix, $\tilde{\mathbf{S}}_\epsilon$, that includes off-diagonal terms representing spatial error correlations for cells within the same month; we assume no between-month dependencies. Individual elements of this matrix are specified using an exponentially decaying correlation kernel; thus, for grid cells i, j the covariance element in $\tilde{\mathbf{S}}_\epsilon$ within each month is $\tilde{S}_{\epsilon,ij} = \sigma^2 \exp(-d_{ij}/L)$ where d_{ij} is the great circle distance between the cells and L the range parameter. We further take the constant observation error σ as 20 % of the global, annual-mean MOPITT-equivalent CO columns from the CTM with prior CO source estimates. Corresponding error terms are then simulated from $\tilde{\boldsymbol{\epsilon}} \sim N(\mathbf{0}, \tilde{\mathbf{S}}_\epsilon)$ and the synthetic CO observations vector is calculated directly from the model, Eq. (1), i.e., $\tilde{\boldsymbol{y}} = \mathbf{K}\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{\epsilon}}$. Finally, the grid cells for which the real MOPITT data is missing are then masked as missing, defining the index sets M and H. We explore 6 repeat versions of this synthetic data using $L = 100$, 200, 500, 1000, 2000 and 5000 km, respectively; this generates 6 synthetic data sets reflecting varying degrees of spatial error correlation.

We repeat this exercise to generate 1000 replicate simulations in order to quantify Monte Carlo variability and resulting accuracy of reconstructions of the true source fluxes. Figures 3 and 4 show spatial plots of elements of the simulated $\tilde{\boldsymbol{\epsilon}}$ and $\tilde{\boldsymbol{y}}$ for the month of December 2000 for one particular realization of $\tilde{\boldsymbol{x}}$ randomly drawn from the set of 1000 replicates.

3.2 Results: model adequacy and model comparisons

For each value of L, we compare the posterior mean estimates of CO fluxes with the known "true" fluxes for each of 1,000 synthetic data sets. Figures 5 and 6 show the results via scatter plots of estimated CO flux versus the "truth" for one FFBF and one BIOM source category, respectively. Similar comparisons for the remaining source categories are shown in the supplementary material. For one randomly selected synthetic data set, Figs. 7 and 8 display the estimated 95 % posterior credible intervals for each of the 15 source categories in both the CAR and NS model analyses; also shown are the corresponding 95 % prior credible intervals and the values of the "true" CO fluxes. It is readily evident from these figures that the performance of the NS model is comparable to that of the CAR model when the degree of spatial error correlation is very low (i.e., L is rather small). The CAR model is clearly superior at higher values of L over the range of "true" fluxes considered in this analysis. Results

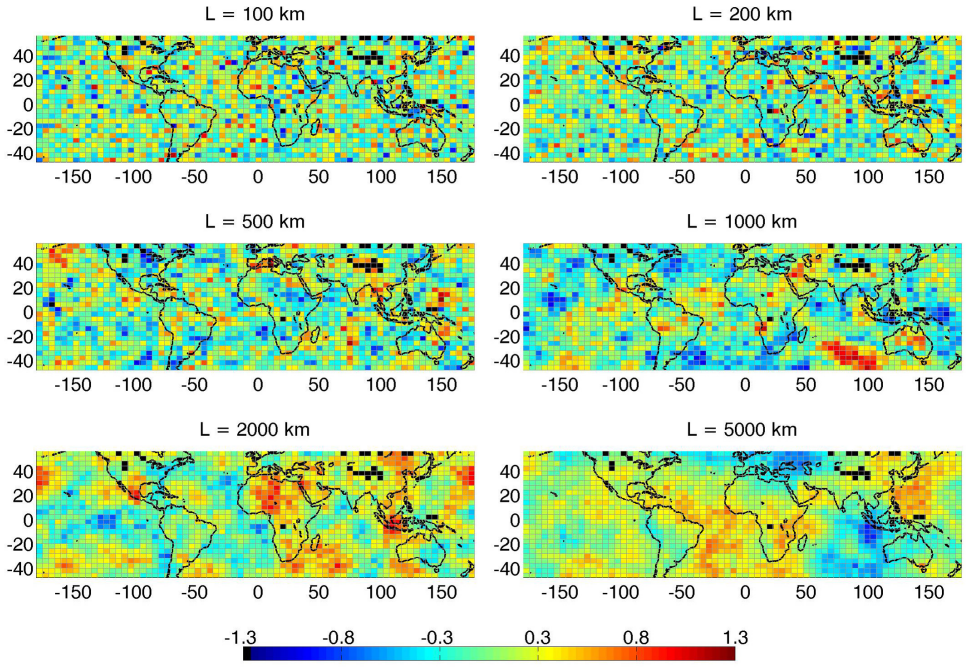

Fig. 3. Spatial images of randomly selected realizations of the synthetic MOPITT CO column observation errors $\tilde{\epsilon}$ (in units of 10^{18} molecules CO cm^{-2}) for December 2000 for different values of L. The black cells represent missing observations.

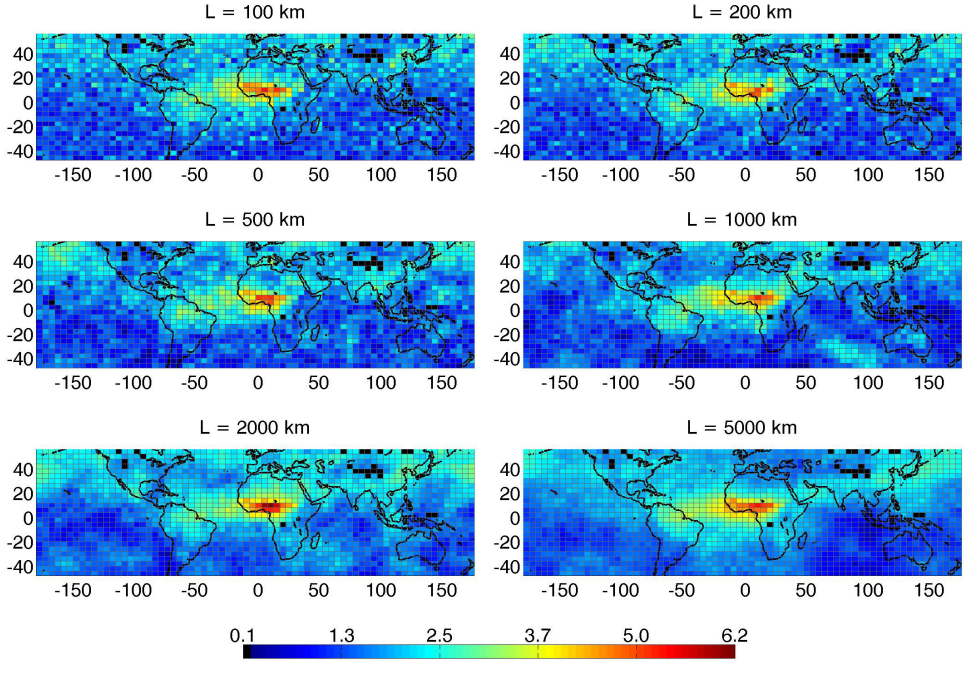

Fig. 4. Spatial images of synthetic MOPITT CO column measurements $\tilde{y}$ (in units of 10^{18} molecules CO cm^{-2}) corresponding to the errors in Fig. 3. The black cells represent missing observations.

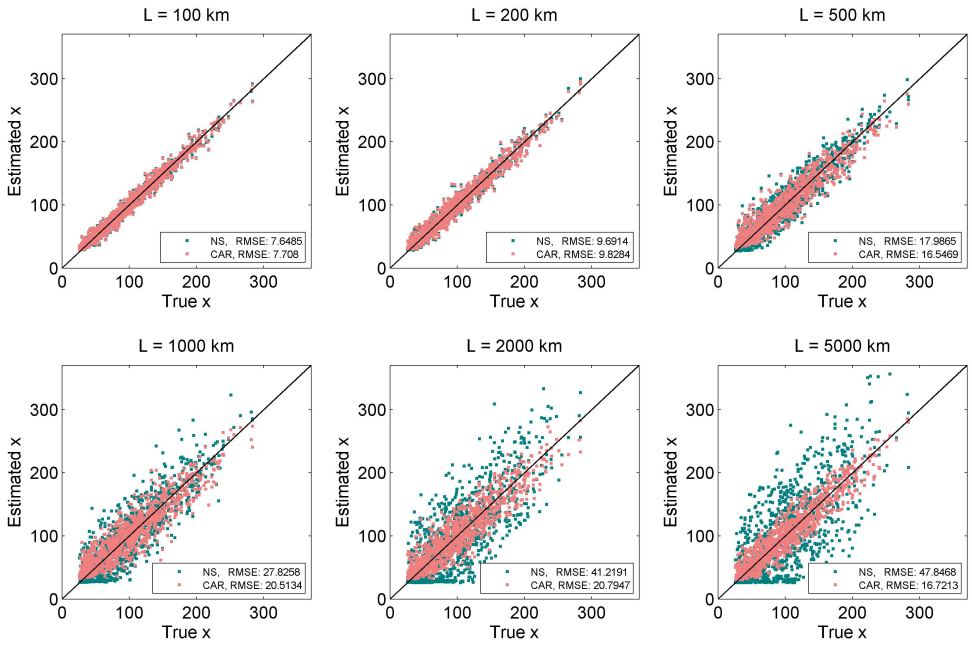

Fig. 5. Scatter plots of "true" x_i versus estimated x_i (in units of Tg CO yr^{-1}) for the FFBF North America (FFBF-NAM) source category computed from 1000 synthetic data sets. RMSE is the root mean square error metric of the estimated x_i from the "true" x_i.

(see supplementary material) are similar across the range of synthetic data sets considered here.

To summarize performance across the 1000 synthetic datasets, and to provide further insight into the relative performance of the CAR and NS approaches in reconstructing source fluxes, we calculate two metrics for each CO source category using the MCMC-sampled posterior distributions:

$$\text{Success Rate} = (\#\text{times the "true" CO flux falls within posterior 95\% interval})/1000, \quad (7)$$

$$\text{Learning Ratio} = \text{average of } (\text{Prior 95\% interval length}/\text{Posterior 95\% interval length}).$$

Figure 9 shows a combined plot of these two metrics for each CO category for different values of L. Notice that the Learning Ratio>1 in each source category under both models, indicating both the models learn significant information from the data; these ratios differ for different sources, reflecting the fact that the measurements provide varying degrees of information on the magnitude of different sources. The Success Rates demonstrate precision of the analysis in the standard statistical coverage sense; Fig. 9 supports the point already noted that at low spatial dependencies the NS and the CAR models have comparable accuracy, whereas the CAR model very substantially outperforms the NS model when spatial structure becomes practically meaningful.

In this synthetic context and with the spatial model based on a single spatial dependence parameter, posterior distributions and resulting credible intervals tend to be quite precise due to the large amount of data. This is perfectly appropriate and a consequence of model assumptions and data-model match. In this specific model the results serve as a first, proof-of-concept and example of the approach for accounting for spatial error correlation structures only; the high level of posterior precision about fluxes may be reduced in more elaborate spatial models for higher resolution data, and further work will aim to explore this in new case studies.

Table 1. Posterior means and 95 % credible intervals for log(BF(CAR:NS)) from the analyses of 1000 synthetic data sets.

L	log(BF(CAR:NS))	95 % CI
100	3.69	[−36.26, 42.01]
200	154.86	[102.99, 208.23]
500	2331.80	[2136.52, 2535.91]
1000	6040.01	[5643.34, 6473.56]
2000	10 450.46	[9644.83, 11 346.37]
5000	16 059.21	[14 180.68, 17 959.34]

Further substantiation in favor of spatial modeling with the CAR approach comes from formal statistical summaries for model comparison. A key, standard measure of relative fit of two models is the Bayes factor (an integrated variant of a likelihood ratio); to compare the CAR with the NS approach, this is

$$\text{BF(CAR : NS)} = p_{\text{CAR}}(\boldsymbol{y}_H)/p_{\text{NS}}(\boldsymbol{y}_H)$$

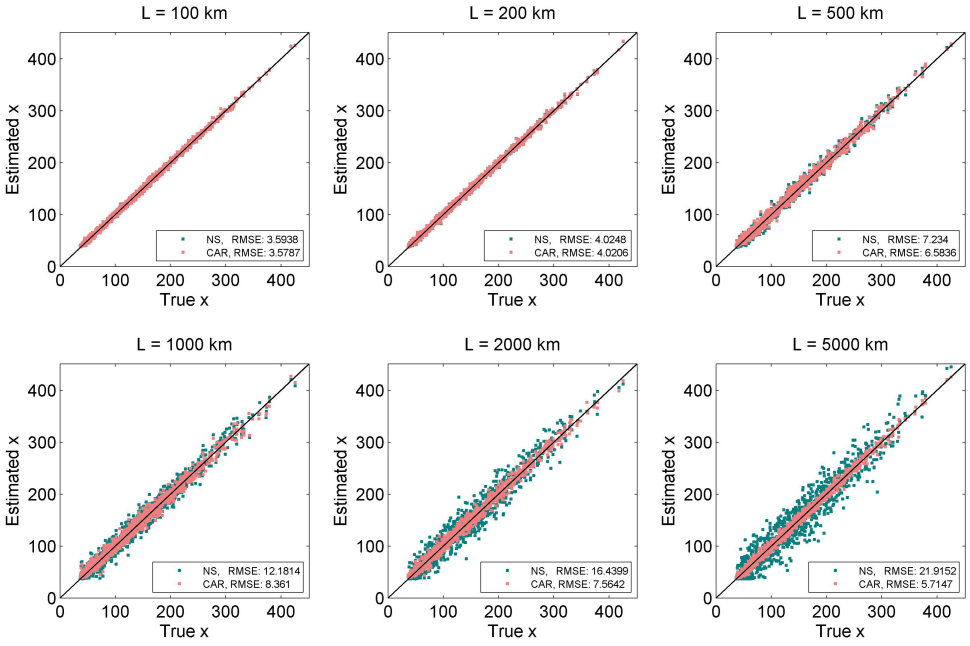

Fig. 6. Scatter plots as in Fig. 5 now for the BIOM Southern Africa (BIOM-SAF) source category.

where $p_*(\boldsymbol{y}_H)$ is the marginal probability density function of the observed data $\boldsymbol{y}_H$ under the assumptions of the model $* = \text{CAR}$ or $* = \text{NS}$, respectively; $p_*(\boldsymbol{y}_H)$ is otherwise referred to as the marginal likelihood or evidence for model $*$, and the ratio form in the Bayes factor measures relative evidence on a likelihood scale; see Bernardo and Smith (1994, chapter 6) and West and Harrison (1997, chapters 11 and 12), for example. A Bayes factor $\text{BF}(\text{CAR}:\text{NS})>1$ indicates that the data favors the CAR over NS model, with values of 100 or more indicating very substantial evidence indeed. From the MCMC analysis of each synthetic data set we can estimate the values of $p_{\text{CAR}}(\boldsymbol{y}_H)$ and $p_{\text{NS}}(\boldsymbol{y}_H)$ and hence estimate the Bayes factor for that particular data set. Table 1 reports the averages of the values over the 1000 simulated data sets, together with the associated 95 % intervals. We see significantly positive values of log(BF(CAR:NS)) for larger L, indicating extremely strong evidence in favor of the spatial model over the non-spatial model.

We illustrate the effectiveness of CAR in modeling these synthetic (non-CAR) spatial dependencies by comparing the synthetic data with samples from the posterior predictive distribution. Exploring posterior predictions is a traditional statistical method for both informal and formal evaluation of model fit; here we simply present graphical summaries of prediction from the model. For any of the posterior MCMC draws of $\{x, \boldsymbol{\theta}, \boldsymbol{y}_M\}$ we can, using these values, directly simulate additional synthetic data $\boldsymbol{y}$ from the model; such simulations generate random draws from the posterior predictive distribution – i.e., synthetic representations of what data will look like if the model is true. Often, simply exploring graphical and numerical summaries of posterior predictive simulated data sets can highlight ways in which the model is inadequate when compared to the real data (West and Harrison, 1997; Gelman et al., 2004). Figures 10 and 11 show representative posterior predictive samples from the NS and CAR models for $L = 100$ km and $L = 5000$ km, respectively. It is clear that the spatial patterns of posterior predictive samples from the CAR model are visually similar to the synthetic data, while they are clearly noisier for the NS model for higher spatial dependences.

4 Analysis of real MOPITT retrievals

We now consider the analysis applied to MOPITT retrievals of CO columns, paralleling the study of Arellano et al. (2004) but now including spatial CAR structure, modified priors, and formal treatment of missing data. Figure 12 displays 95 % posterior credible intervals from CAR and NS models for each source category, and Table 2 presents detailed posterior summaries. We observe significant differences between the two analyses for several of the CO source categories considered here. In particular, the CAR analysis suggests that, for several of the FFBF and BIOM source categories the top-down estimates are not as inconsistent with the bottom-up estimates as is suggested by the NS analysis. Again, as with the synthetic data study above, we note relatively precise posterior intervals based on the MCMC approach; these are accurate summaries of uncertainties conditional on the assumed form of the model.

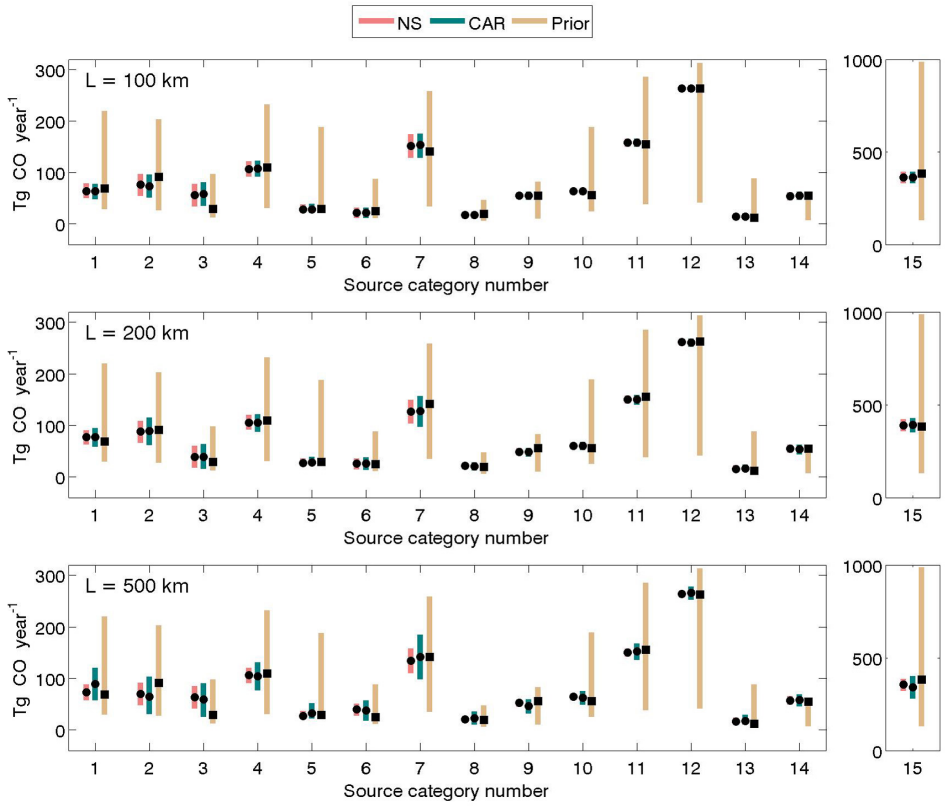

Fig. 7. Plots of 95 % posterior credible intervals for NS and CAR model analyses for $L = 100$, 200 and 500 km, showing summary inferences of source magnitude for all 15 CO source categories for one synthetic dataset (see Fig. 2 for definition of source category numbers). Posterior means for both are marked with dots inside the corresponding intervals. Alongside we plot 95 % prior credible intervals for the corresponding source and indicate the "true" CO source with a square.

Model comparison using Bayes factors yields $\log(\mathrm{BF}(\mathrm{CAR}:\mathrm{NS})) > 10073$, simply overwhelming statistical evidence that the CAR model substantially improves model:data match due to the presence of significant spatial residual structure. To further indicate the ability of the spatial model to reflect realistic spatial structure, Fig. 13 displays two randomly selected posterior predictive samples from the NS and CAR model analyses, together with the actual data for December 2000. The spatial dependence patterns in the CAR samples are visually similar to that in the real data, while the NS samples are again noisier. These preliminary comparisons suggest that accounting for spatial error structures in the real data is important in the context of constraining CO sources using spatially-dense satellite measurements. Further investigations at higher spatial and temporal resolution with the latest version of the MOPITT dataset, as well with retrievals from other satellite instruments, are required to more fully characterize and account for these spatial error patterns and to refine top-down CO source estimates.

5 Concluding remarks

The fast-expanding ability to access increasingly high-resolution atmospheric data using satellite imagery raises exciting opportunities for substantial advances in data synthesis in inverse modeling. Capitalizing on this opportunity will involve increased attention to core challenges that are inherently statistical in nature. The work presented here reflects this view and exemplifies the potential to address rather basic yet challenging problems of very large-scale spatial modeling, coupled with refined prior specifications and treatment of missing data, in inverse studies of atmospheric trace gas source/sink flux estimation. To date, although the broader field of atmospheric chemistry inverse modeling has become heavily invested in statistical methods, there has been limited development of what are standard statistical approaches utilizing Bayesian simulation methods, including MCMC. The work here demonstrates the utility of the Bayesian perspective and the enabling computational methodology provides for extending inverse modeling frameworks to incorporate relevant spatial stochastic structure.

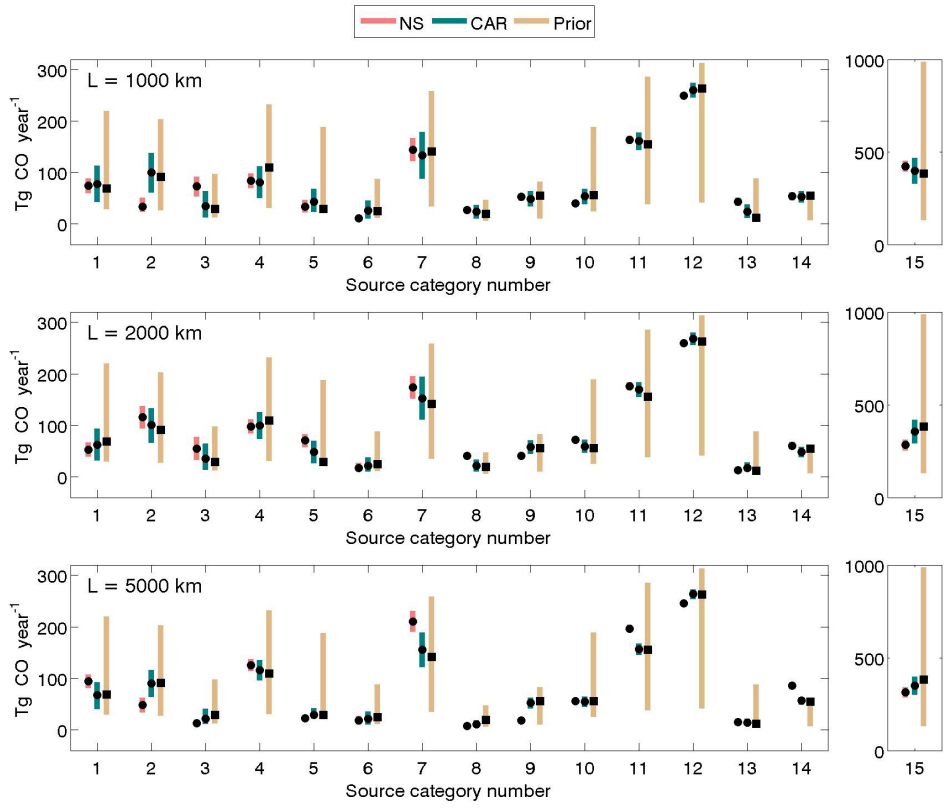

Fig. 8. Plots as in Fig. 7, now based on L =1000, 2000, and 5000 km.

Table 2. Summary of posterior inferences for the CO fluxes (in units of Tg CO yr^{-1}) and model parameters from analysis of actual MOPITT retrieval data. See Fig. 2 for definition of source category numbers.

	Prior		NS model		NS model	CAR model		CAR model
	mean	(SD)	mean	(SD)	95 % CI	mean	(SD)	95 % CI
x_1	102.99	(51.49)	83.13	(3.46)	[76.26, 89.85]	73.24	(10.84)	[51.97, 94.11]
x_2	95.20	(47.60)	37.18	(5.42)	[26.90, 48.40]	42.71	(10.87)	[25.22, 66.09]
x_3	45.72	(22.86)	95.43	(5.98)	[83.56, 106.92]	54.17	(11.41)	[31.45, 76.10]
x_4	108.72	(54.36)	195.06	(3.58)	[188.09, 202.19]	158.55	(8.61)	[141.74, 175.70]
x_5	88.14	(44.07)	147.75	(3.14)	[141.61, 153.93]	116.96	(7.48)	[102.37, 131.79]
x_6	41.03	(20.51)	70.29	(2.84)	[64.72, 75.93]	67.90	(7.38)	[53.43, 82.71]
x_7	120.98	(60.49)	265.42	(5.63)	[254.48, 276.36]	108.83	(13.95)	[81.20, 136.02]
x_8	21.99	(10.99)	64.52	(1.56)	[61.49, 67.59]	38.94	(4.24)	[30.63, 47.23]
x_9	38.58	(19.29)	96.56	(1.68)	[93.24, 99.78]	34.23	(4.31)	[25.80, 42.77]
x_{10}	88.28	(44.14)	100.03	(1.47)	[97.11, 102.93]	59.41	(4.07)	[51.39, 67.26]
x_{11}	133.91	(66.95)	92.25	(1.83)	[88.70, 95.86]	52.56	(4.53)	[43.72, 61.44]
x_{12}	146.51	(73.25)	92.53	(1.50)	[89.62, 95.48]	90.99	(3.77)	[83.71, 98.61]
x_{13}	41.51	(20.75)	104.48	(2.13)	[100.31, 108.66]	44.27	(4.74)	[34.78, 53.47]
x_{14}	28.05	(14.02)	13.09	(1.88)	[9.45, 16.81]	10.32	(2.20)	[7.19, 15.34]
x_{15}	462.12	(231.06)	152.00	(7.73)	[136.47, 166.94]	224.74	(20.73)	[183.57, 265.80]
τ_n^2	0.1333	(0.0666)	0.0277	(3.0e-4)				
τ_c^2	1.0662	(0.5331)				0.0008	(8.9e-6)	[0.0008, 0.0008]
ρ	0.5000	(0.2890)				0.99992	(3.6e-5)	[0.99983, 0.99997]

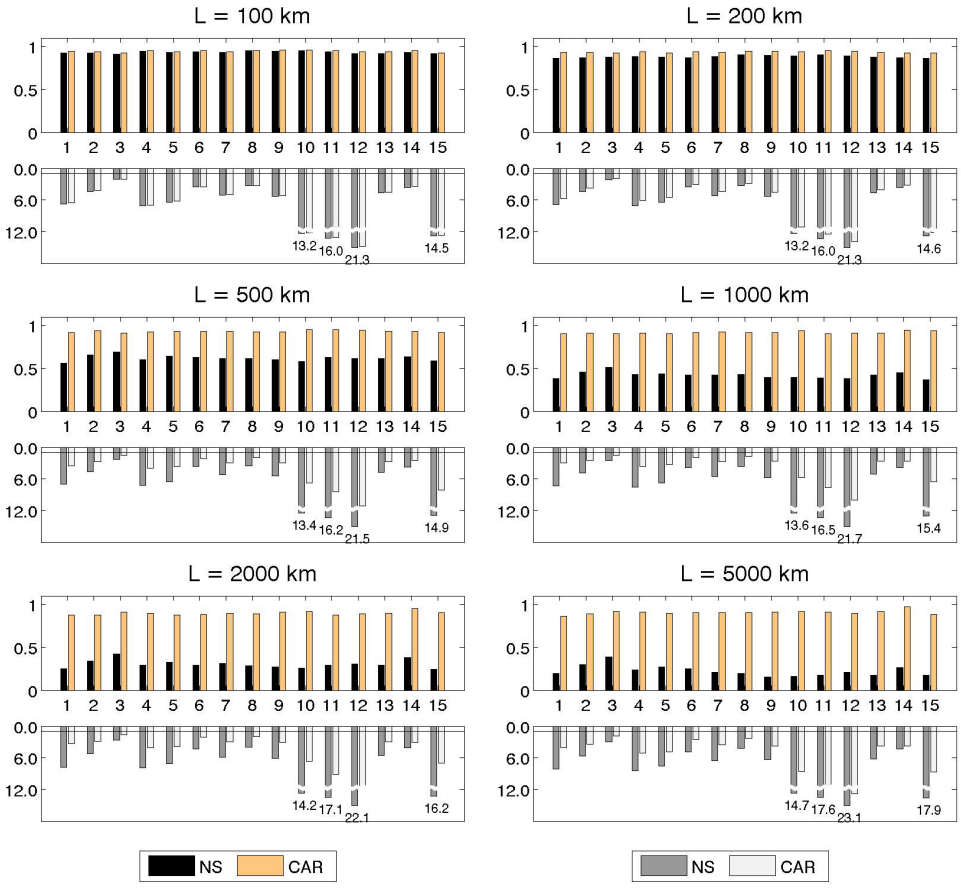

Fig. 9. Success Rate (top panels) and Learning Ratio (bottom panels) for all CO source categories for different values of L. See Fig. 2 for definition of source category numbers.

Our analysis framework explicitly paralleled the earlier work of Arellano et al. (2004) in order to demonstrate the extensions with spatial CAR modeling and other statistical innovations. Based on this proof-of-concept in studies of both synthetic and real MOPITT retrieval data, some next steps include expanding the model framework to explicitly represent time dependencies in data and models. The extension of spatial modeling into a temporal framework is standard, and this, together with extension to model potential time-variations in source fluxes, will rely on additional Bayesian computational methods that are well-developed in other areas of multivariate time series analysis (West and Harrison, 1997; Prado and West, 2010). Additional important directions include the application of these techniques to carbon dioxide and methane as high quality satellite measurements of these climatically important gases become available, especially in the context of source/sink estimation with high spatial resolution.

With regard to computational demands for higher resolution problems, we note that the running time for the MCMC algorithm increases roughly linearly with n, the size of the source vector. This is based on evaluations of computational time using our current Matlab code and is approximately linear in FLOP counts, stopwatch time (*tic toc*) and elapsed CPU time (*cputime*). This is the best one can expect and indicates that indeed the analysis and computational approach is scalable to much larger and higher resolution problems. Moreover, the matrix multiplications that constitute much of the computational burden with increasing m can be trivially parallelized on multi-core machines or clusters, or via exploitation of GPU parallelization, suggesting the opportunity for very substantial gains with larger m.

Finally, we note again that the analysis presented is intended as a first, proof-of-principle analysis and example of the opportunity to integrate spatial structure into inversions. The single-dependence parameter CAR model is likely overly simplistic as a representation of spatial errors that combine model misfit and natural local dependencies in satellite retrieval data. To represent additional complexity in spatial structure, especially with regard to the potential opportunities to fit higher resolution models that can capture more refined structure, extensions or alternatives to CAR models will be needed. Indeed, one of our current research

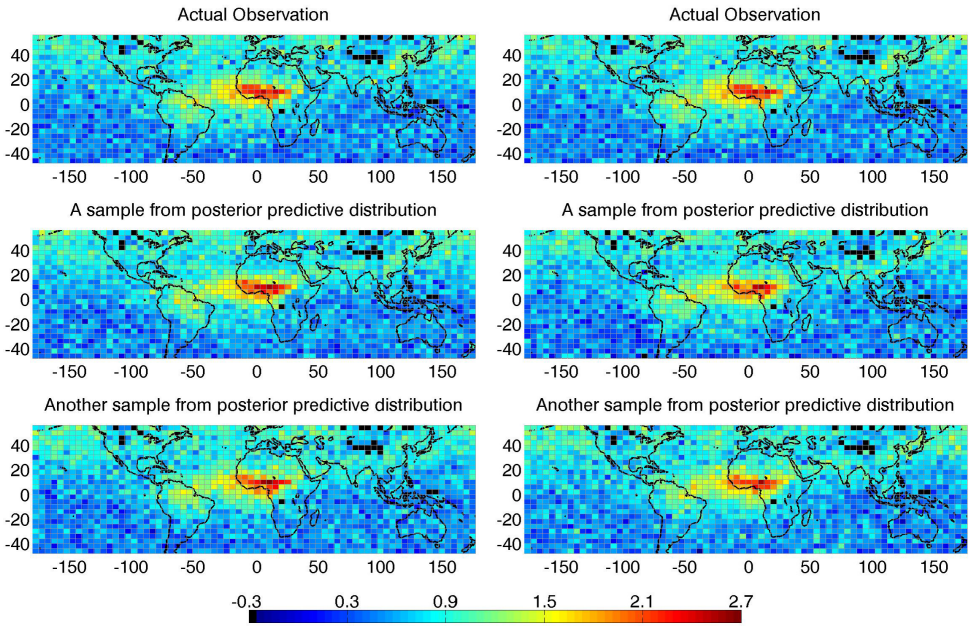

Fig. 10. Two samples of column CO fields (in units of 10^{18} molecules CO cm^{-2}) from the posterior predictive distributions of NS (lower 2 left panels) and CAR (lower 2 right panels) models for December 2000 for $L = 100$ km; top panels show the corresponding synthetic data.

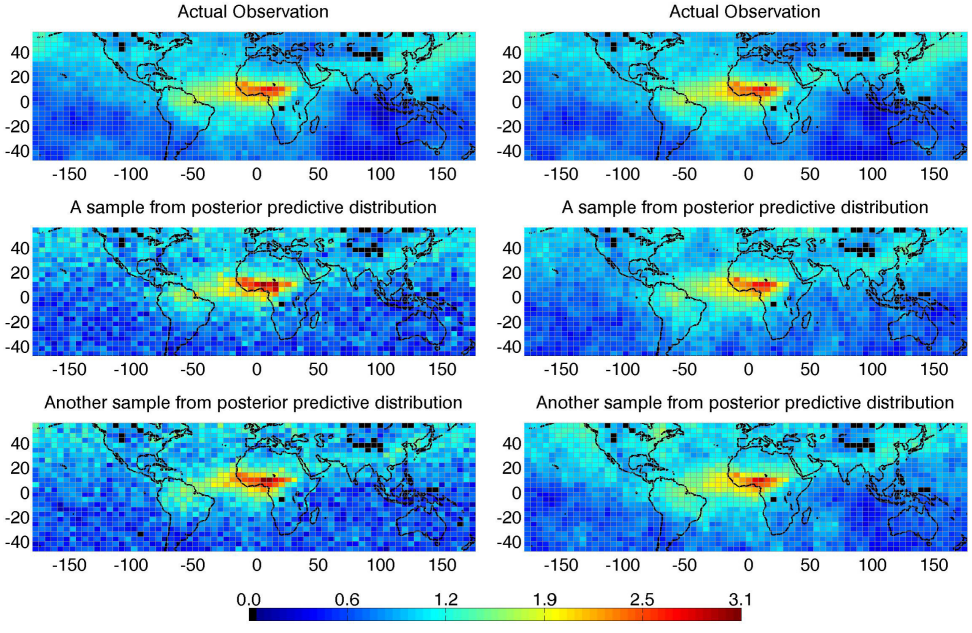

Fig. 11. Posterior predictive plots as in Fig. 10, now with $L = 5000$ km.

projects is to extend the general strategy of the paper to models that permit changes in the local dependency parameter across the spatial region; importantly, such extensions will also need to come through precision matrix models, rather than covariance models, for both the feasibility of computations and, we believe, for practical realism. Our work in this paper demonstrates the ability of simple CAR models to substantially improve over the standard non-spatial models in terms of statistical fit, source flux recovery in synthetic examples, and predictive match with the real data; this gives us a basis to move ahead with development of more general, flexible and likely realistic spatial structures in future work.

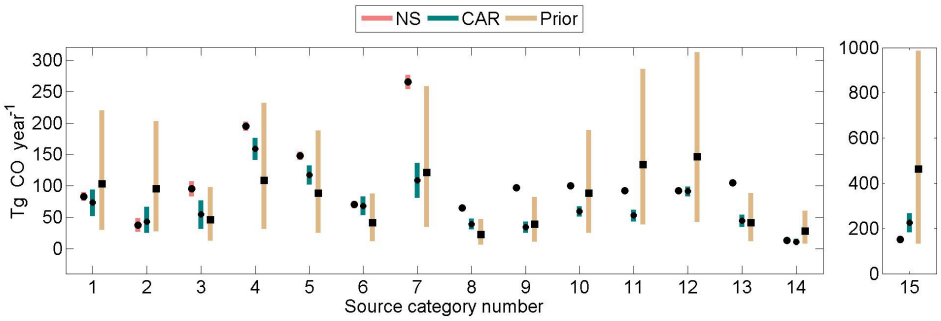

Fig. 12. Plots of 95 % prior and posterior credible intervals for the real MOPITT data inversion. Posterior means for both models are marked with dots inside the corresponding intervals. The squares on the prior credible intervals represent the corresponding prior means. See Fig. 2 for definition of source category numbers.

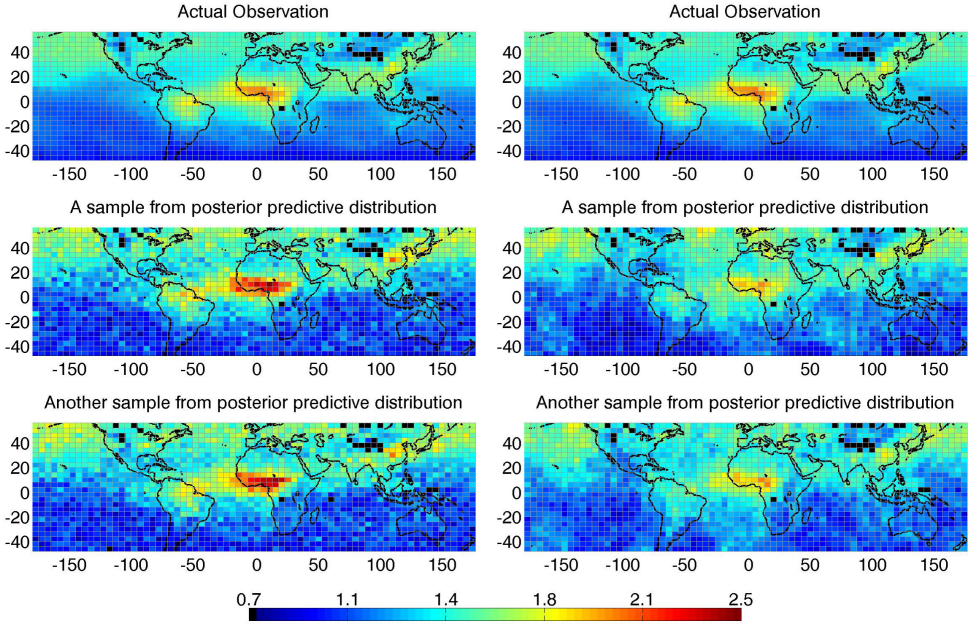

Fig. 13. Two samples of column CO fields (in units of 10^{18} molecules CO cm^{-2}) from the posterior predictive distributions of NS (lower 2 plots, left panels) and CAR (lower 2 plots, right panels) models for December 2000; the top panels show the corresponding MOPITT retrieval data.

Appendix A

Posterior computation

The Markov chain Monte Carlo posterior simulator successively re-simulates values of all of the unknowns $\{\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y}_M\}$ to draw a large Monte Carlo sample from the full joint posterior $p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y}_M \mid \boldsymbol{y}_H)$. Initializing at (essentially arbitrary) starting values $\boldsymbol{\theta}, \boldsymbol{y}_M$, the MCMC proceeds through many iterations to revise the full set of unknowns, at each iterate stepping through the stages below to stochastically update $\boldsymbol{x}$ conditional on the last values of $\{\boldsymbol{\theta}, \boldsymbol{y}_M\}$, then $\boldsymbol{\theta}$ conditional on the latest values of $\{\boldsymbol{x}, \boldsymbol{y}_M\}$, and then $\boldsymbol{y}_M$ conditional on the latest values of $\{\boldsymbol{x}, \boldsymbol{\theta}\}$. The specific distributions used for each of these three stages are summarized here with more technical details in the supplemental documentation.

We use the following notation:

i. $\boldsymbol{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)'$

ii. $\mathbf{K}_{(*,i)}$ is the ith column of $\mathbf{K}$

iii. $\mathbf{K}_{(*,-i)}$ is the submatrix of $\mathbf{K}$ obtained by deleting the ith column

iv. $\boldsymbol{y}_A = (y_i)_{i \in A}$, a subvector of $\boldsymbol{y}$

v. $\mathbf{K}_{(A,*)} = (\mathbf{K}_{(i,*)})_{i \in A}$, a submatrix of $\mathbf{K}$

vi. $\mathbf{U}_{A,B} = (U_{ij})_{i\in A, j\in B}$, a submatrix of $\mathbf{U}$

vii. M is the set of indices for missing retrievals, $M \subset \{1:m\}$, while $H = \{1:m\}\backslash M$

viii. IG(a,b) stands for an inverse gamma distribution

We give summary details for MCMC in both the non-spatial and CAR model contexts. As described in Sect. 2.3, analysis is based on the use of the truncated normal priors for sources with that for the i-th source being $x_i \sim N(m_{\mathrm{a},i}, v_{\mathrm{a},i})I(x_i > t_i)$ where $m_{\mathrm{a},i}$ and $v_{\mathrm{a},i}$ are numerically specified so that the prior has mean $x_{\mathrm{a},i}$ and variance $S_{\mathrm{a},i}$ with $t_i = x_{\mathrm{a},i}/4$.

A1 Single epoch data

We first give summaries for the analysis of a single epoch of data – a single monthly retrieval $\boldsymbol{y}$ in the case of MOPITT data. The extension to the context analyzed in our examples, paralleling Arellano et al. (2004) with multi-month retrievals and time-invariant fluxes, is then also summarized in the following subsection.

A1.1 Posterior computation for the non-spatial model

The prior distribution for τ_n^2 is an inverse Gamma distribution, $\tau_n^2 \sim$IG(α_n, λ_n) where we set the prior mean, $\mathbb{E}(\tau_n^2) = \sigma^2$ (known), and the coefficient of variation to 0.5 so that $\alpha_n = 6$ and $\lambda_n = 5\sigma^2$. The MCMC algorithm alternatively samples from the following conditional distributions:

– For $i = 1,\ldots,n$, resample the ith source element from

$$\left(x_i \mid \boldsymbol{x}_{-i}, \tau_n^2, \boldsymbol{y}, \mathbf{K}\right) \sim N(E_{n,i}, V_{n,i})\ I(x_i > t_i)$$

where

$$V_{n,i} = \left(v_{\mathrm{a},i}^{-1} + \tau_n^{-2}\mathbf{K}'_{(*,i)}\mathbf{K}_{(*,i)}\right)^{-1}$$

and

$$E_{n,i} = V_{n,i}\left\{v_{\mathrm{a},i}^{-1}m_{\mathrm{a},i} + \tau_n^{-2}\mathbf{K}'_{(*,i)}(\boldsymbol{y} - \mathbf{K}_{(*,-i)}\boldsymbol{x}_{-i})\right\}.$$

– Resample $(\tau_n^2 \mid \boldsymbol{x}, \boldsymbol{y}, \mathbf{K}) \sim$ IG$(\alpha_n + m/2, \lambda_n + q/2)$ where $q = (\boldsymbol{y} - \mathbf{K}\boldsymbol{x})'(\boldsymbol{y} - \mathbf{K}\boldsymbol{x})$.

– Resample values of the missing data vector from the conditional posterior predictive distribution $(\boldsymbol{y}_M \mid \boldsymbol{x}, \tau_n^2, \mathbf{K}) \sim N(\mathbf{K}_{(M,*)}\boldsymbol{x}, \tau_n^2\mathbf{I})$.

A1.2 Posterior computation for the CAR model

The prior for $\tau_c^2 \sim$IG(α_c, λ_c) with prior mean, $\mathbb{E}(\tau_c^2) = 8\sigma^2$ (known, sets unbiased prior when there is no spatial dependence, i.e. $\rho = 0$), and coefficient of variation set at 0.5; this implies $\alpha_c = 6$ and $\lambda_c = 40\sigma^2$. Further, we adopt the uniform prior for ρ on $0 < \rho < 1$. The MCMC algorithm alternatively samples from the following conditional distributions:

– For $i = 1,\ldots,n$, resample the ith source element from

$$\left(x_i \mid \boldsymbol{x}_{-i}, \tau_c^2, \rho, \boldsymbol{y}, \mathbf{K}\right) \sim N(E_{c,i}, V_{c,i})I(x_i > t_i)$$

where

$$V_{c,i} = \left(v_{\mathrm{a},i}^{-1} + \mathbf{K}'_{(*,i)}\mathbf{U}\mathbf{K}_{(*,i)}\right)^{-1}$$

and

$$E_{c,i} = V_{c,i}\left\{v_{\mathrm{a},i}^{-1}m_{\mathrm{a},i} + \mathbf{K}'_{(*,i)}\mathbf{U}(\boldsymbol{y} - \mathbf{K}_{(*,-i)}\boldsymbol{x}_{-i})\right\}$$

with $\mathbf{U} = \tau_c^{-2}(\mathbf{D}_w - \rho^*\mathbf{W})$.

– Resample $(\tau_c^2 \mid \boldsymbol{x}, \rho, \boldsymbol{y}, \mathbf{K}) \sim$IG$(\alpha_c + m/2, \lambda_c + q/2)$ where $q = (\boldsymbol{y} - \mathbf{K}\boldsymbol{x})'(\mathbf{D}_w - \rho\mathbf{W})(\boldsymbol{y} - \mathbf{K}\boldsymbol{x})$.

– Resample $(\rho \mid \boldsymbol{x}, \tau_c^2, \boldsymbol{y}, \mathbf{K})$ with a random-walk Metropolis step as follows. First, sample a candidate value $\rho^* \sim N(\rho, s^2)$ and compute $\mathbf{U}^* = \tau_c^{-2}(\mathbf{D}_w - \rho^*\mathbf{W})$. The candidate value is then accepted with probability

$$\alpha = \min\left\{1, \frac{N(\boldsymbol{y} \mid \mathbf{K}\boldsymbol{x}, \mathbf{U}^{*-1})}{N(\boldsymbol{y} \mid \mathbf{K}\boldsymbol{x}, \mathbf{U}^{-1})}\right\};$$

if accepted, set $\rho = \rho^*$ and $\mathbf{U} = \mathbf{U}^*$; otherwise retain the previous values $\rho, \mathbf{U}$. The step size s is defined adaptively during the initial burn-in phase of the MCMC.

– Resample values of the missing data vector from the conditional posterior predictive distribution

$$\left(\boldsymbol{y}_M \mid \boldsymbol{x}, \tau_c^2, \rho, \boldsymbol{y}_H, \mathbf{K}\right) \sim$$
$$N\left[\mathbf{K}_{(M,*)}\boldsymbol{x} - \mathbf{U}_{M,M}^{-1}\mathbf{U}_{M,H}(\boldsymbol{y}_H - \mathbf{K}_{(H,*)}\boldsymbol{x}),\ \mathbf{U}_{M,M}^{-1}\right].$$

Note here how the spatial covariance structure in $\mathbf{U}$ plays a key role in determining the relative weightings of cells having observed data via the current values of multiple regression coefficients (in the regression of $\boldsymbol{y}_M$ on $\boldsymbol{y}_H$). Actual sampling from this distribution does not in fact require any matrix inversions; only a Cholesky decomposition of a square matrix whose dimension is the number of missing values $|M|$ (137×137 in our MOPITT study context) (Rue, 2001; Rue and Held, 2005). As a result, these successive imputations of missing data to represent the posterior estimates and uncertainties about $\boldsymbol{y}_M$ do not add measurably to the overall computational burden of the MCMC.

A2 Multi-epoch data with time-invariant fluxes

To parallel Arellano et al. (2004), consider now the case of several epochs (e.g. months) of retrieval data. In epoch $t = 1,\ldots,T$, retrievals follow the model of Eq. (1) where we now index by t, viz. $\boldsymbol{y}_t = \mathbf{K}_t\boldsymbol{x} + \boldsymbol{\epsilon}_t$ for $t = 1,\ldots,T$. We can simply

stack the vectors of retrievals to obtain a model as in Eq. (1) for the full set of T epochs. Thus, we now understand that (a) $\boldsymbol{y}$ is the $(mT)\times 1$ vector obtained by stacking the T vectors $\boldsymbol{y}_t$, (b) $\mathbf{K}$ is the $(mT)\times n$ matrix obtained by stacking the T matrices $\mathbf{K}_t$, and (c) $\boldsymbol{\epsilon}$ is the $(mT)\times 1$ matrix obtained by stacking the T error vectors $\boldsymbol{\epsilon}_t$. The above MCMC analysis is modified in minor technical details as a result, as follows.

A2.1 Posterior computation for the non-spatial model

The one modification needed to the summary in Appendix A.1 above is in resampling τ_n; now the degrees-of-freedom of the inverse gamma distribution is mT instead of m, reflecting the T epochs of data, viz. $(\tau_n^2 \mid \boldsymbol{x},\boldsymbol{y},\mathbf{K}) \sim \mathrm{IG}(\alpha_n + mT/2, \lambda_n + q/2)$ where $q = (\boldsymbol{y}-\mathbf{K}\boldsymbol{x})'(\boldsymbol{y}-\mathbf{K}\boldsymbol{x})$.

A2.2 Posterior computation for the CAR model

There are three modifications to the single epoch summary of Appendix A.2 above: details of resampling τ_c^2, ρ and then $\boldsymbol{y}_M$, as follows.

– Resample $(\tau_c^2 \mid \boldsymbol{x},\rho,\boldsymbol{y},\mathbf{K}) \sim \mathrm{IG}(\alpha_c + mT/2, \lambda_c + q/2)$ where $q = \sum_{t=1}^{T}(\boldsymbol{y}_t - \mathbf{K}_t\boldsymbol{x})'(\mathbf{D}_w - \rho\mathbf{W})(\boldsymbol{y}_t - \mathbf{K}_t\boldsymbol{x})$.

– In the Metropolis-Hastings resampling of ρ, the acceptance probability α is modified to

$$\alpha = \min\left\{1, \prod_{t=1}^{T}\frac{N(\boldsymbol{y}_t \mid \mathbf{K}_t\boldsymbol{x}, \mathbf{U}^{*-1})}{N(\boldsymbol{y}_t \mid \mathbf{K}_t\boldsymbol{x}, \mathbf{U}^{-1})}\right\}.$$

– Resample values of the missing data vector from the conditional posterior predictive distribution

$$\left(\boldsymbol{y}_M \mid \boldsymbol{x}, \tau_c^2, \rho, \boldsymbol{y}_H, \mathbf{K}\right) \sim N\left[\mathbf{K}_{(M,*)}\boldsymbol{x} - \mathbf{V}_{M,M}^{-1}\mathbf{V}_{M,H}\left(\boldsymbol{y}_H - \mathbf{K}_{(H,*)}\boldsymbol{x}\right), \mathbf{V}_{M,M}^{-1}\right]$$

where $\mathbf{V}$ is the $(mT)\times(mT)$ block diagonal matrix comprised of T diagonal $M\times M$ blocks $\mathbf{U}$; that is, $\mathbf{V} = \mathbf{I}\otimes\mathbf{U}$ where $\mathbf{I}$ is the $T\times T$ identity matrix and $\otimes$ represents Kronecker product.

Supplementary material related to this article is available online at: http://www.atmos-chem-phys.net/11/5365/2011/acp-11-5365-2011-supplement.zip.

Acknowledgements. Aspects of the research reported here were supported by NASA grants NNX08AL03G, NNX08AQ04G, and NNX08AF64G.

Edited by: C. Gerbig

References

Arellano, A. F. J., Kasibhatla, P. S., Goglio, L., van der Werf, G. R., and Randerson, J. T.: Top-down estimates of global CO sources using MOPITT measurements, Geophys. Res. Lett., 31, L01104, doi:10.1029/2003GL018609, 2004.

Arellano, A. F., Kasibhatla, P. S., Giglio, L., van der Werf, G. R., Randerson, J. T., and Collatz, G. J.: Time-dependent inversion estimates of global biomass-burning CO emissions using measurement of pollution in the troposphere (MOPITT) measurements, J. Geophys. Res., 111, D09303, doi:10.1029/2005JD006613, 2006.

Baker, D. F., Law, R. M., Gurney, K. R., Rayner, P., Peylin, P., Denning, A. S., Bousquet, P., Bruhwiler, L., Chen, Y. H., Ciais, P., Fung, I. Y., Heimann, M., John, J., Maki, T., Maksyutov, S., Masarie, K., Prather, M., Pak, B., Taguchi, S., and Zhu, Z.: TransCom 3 inversion intercomparison: impact of transport model errors on the interannual variability of regional CO_2 fluxes, 1988–2003, Global Biogeochem. Cy., 20, GB1002, doi:10.1029/2004GB002439, 2006.

Bergamaschi, P., Hein, R., Heimann, M., and Crutzen, P. J.: Inverse modeling of the global CO cycle 1. Inversion of CO mixing ratios, J. Geophys. Res.-Atmos., 105, 1909–1927, 2000.

Bernardo, J. M. and Smith, A. F. M.: Bayesian Theory, Wiley, London, 1994.

Bousquet, P., Peylin, P., Ciais, P., Le Quere, C., Friedlingstein, P., and Tans, P. P.: Regional changes in carbon dioxide fluxes of land and oceans since 1980, Science, 290, 1342–1346, 2000.

Bousquet, P., Ciais, P., Miller, J. B., Dlugokencky, E. J., Hauglustaine, D. A., Prigent, C., Van der Werf, G. R., Peylin, P., Brunke, E. G., Carouge, C., Langenfelds, R. L., Lathiere, J., Papa, F., Ramonet, M., Schmidt, M., Steele, L. P., Tyler, S. C., and White, J.: Contribution of anthropogenic and natural sources to atmospheric methane variability, Nature, 443, 439–443, 2006.

Chevallier, F.: Impact of correlated observation errors on inverted CO_2 surface fluxes from OCO measurements, Geophys. Res. Lett., 34, L24804, doi:10.1029/2007GL030463, 2007.

Chevallier, F., Engelen, R. J., and Peylin, P.: The contribution of AIRS data to the estimation of CO_2 sources and sinks, Geophys. Res. Lett., 32, L23801, doi:10.1029/2005GL024229, 2005a.

Chevallier, F., Fisher, M., Peylin, P., Serrar, S., Bousquet, P., Breon, F. M., Chedin, A., and Ciais, P.: Inferring CO_2 sources and sinks from satellite observations: method and application to TOVS data, J. Geophys. Res., 110, D24309, doi:10.1029/2005JD006390, 2005b.

Chevallier, F., Breon, F. M., and Rayner, P. J.: Contribution of the Orbiting Carbon Observatory to the estimation of CO_2 sources and sinks: theoretical study in a variational data assimilation framework, J. Geophys. Res., 112, D09307, doi:10.1029/2006JD007375, 2007.

Chevallier, F., Engelen, R. J., Carouge, C., Conway, T. J., Peylin, P., Pickett-Heaps, C., Ramonet, M., Rayner, P. J., and Xueref-Remy, I.: AIRS-based versus flask-based estimation of carbon surface fluxes, J. Geophys. Res., 114, D20303, doi:10.1029/2009JD012311, 2009a.

Chevallier, F., Maksyutov, S., Bousquet, P., Breon, F. M., Saito, R., Yoshida, Y., and Yokota, T.: On the accuracy of the CO_2 surface fluxes to be estimated from the GOSAT observations, Geophys. Res. Lett., 36, L19807, doi:10.1029/2009GL040108, 2009b.

Enting, I. G., Trudinger, C. M., and Francey, R. J.: A synthesis inversion of the concentration and delta-C-13 of atmospheric CO_2, Tellus B, 47, 35–52, 1995.

Feng, L., Palmer, P. I., Bösch, H., and Dance, S.: Estimating surface CO_2 fluxes from space-borne CO_2 dry air mole fraction observations using an ensemble Kalman Filter, Atmos. Chem. Phys., 9, 2619–2633, doi:10.5194/acp-9-2619-2009, 2009.

Fletcher, S. E. M., Tans, P. P., Bruhwiler, L. M., Miller, J. B., and Heimann, M.: CH_4 sources estimated from atmospheric observations of CH_4 and its C-13/C-12 isotopic ratios: 1. Inverse modeling of source processes, Global Biogeochem. Cy., 18, GB4004, doi:10.1029/2004GB002223, 2004.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B.: Bayesian Data Analysis, Texts in Statistical Science, 2nd edn., Chapman & Hall/CRC Press, New York, 2004.

Gerbig, C., Lin, J. C., Wofsy, S. C., Daube, B. C., Andrews, A. E., Stephens, B. B., Bakwin, P. S., and Grainger, C. A.: Toward constraining regional-scale fluxes of CO_2 with atmospheric observations over a continent: 2. Analysis of COBRA data using a receptor-oriented framework, J. Geophys. Res., 108, 4757, doi:10.1029/2003JD003770, 2003.

Gourdji, S. M., Hirsch, A. I., Mueller, K. L., Yadav, V., Andrews, A. E., and Michalak, A. M.: Regional-scale geostatistical inverse modeling of North American CO_2 fluxes: a synthetic data study, Atmos. Chem. Phys., 10, 6151–6167, doi:10.5194/acp-10-6151-2010, 2010.

Gurney, K. R., Law, R. M., Denning, A. S., Rayner, P. J., Baker, D., Bousquet, P., Bruhwiler, L., Chen, Y. H., Ciais, P., Fan, S., Fung, I. Y., Gloor, M., Heimann, M., Higuchi, K., John, J., Maki, T., Maksyutov, S., Masarie, K., Peylin, P., Prather, M., Pak, B. C., Randerson, J., Sarmiento, J., Taguchi, S., Takahashi, T., and Yuen, C. W.: Towards robust regional estimates of CO_2 sources and sinks using atmospheric transport models, Nature, 415, 626–630, 2002.

Gurney, K. R., Law, R. M., Denning, A. S., Rayner, P. J., Baker, D., Bousquet, P., Bruhwiler, L., Chen, Y. H., Ciais, P., Fan, S. M., Fung, I. Y., Gloor, M., Heimann, M., Higuchi, K., John, J., Kowalczyk, E., Maki, T., Maksyutov, S., Peylin, P., Prather, M., Pak, B. C., Sarmiento, J., Taguchi, S., Takahashi, T., and Yuen, C. W.: TransCom 3 CO_2 inversion intercomparison: 1. Annual mean control results and sensitivity to transport and prior flux information, Tellus B, 55, 555–579, 2003.

Heald, C. L., Jacob, D. J., Jones, D. B. A., Palmer, P. I., Logan, J. A., Streets, D. G., Sachse, G. W., Gille, J. C., Hoffman, R. N., and Nehrkorn, T.: Comparative inverse analysis of satellite (MOPITT) and aircraft (TRACE-P) observations to estimate Asian sources of carbon monoxide, J. Geophys. Res., 109, D23306, doi:10.1029/2004JD005185, 2004.

Hein, R., Crutzen, P. J., and Heimann, M.: An inverse modeling approach to investigate the global atmospheric methane cycle, Global Biogeochem. Cy., 11, 43–76, 1997.

Houweling, S., Kaminski, T., Dentener, F., Lelieveld, J., and Heimann, M.: Inverse modeling of methane sources and sinks using the adjoint of a global transport model, J. Geophys. Res., 104, 26137–26160, 1999.

Houweling, S., Breon, F.-M., Aben, I., Rödenbeck, C., Gloor, M., Heimann, M., and Ciais, P.: Inverse modeling of CO_2 sources and sinks using satellite data: a synthetic inter-comparison of measurement techniques and their performance as a function of space and time, Atmos. Chem. Phys., 4, 523–538, doi:10.5194/acp-4-523-2004, 2004.

Jones, D. B. A., Bowman, K. W., Palmer, P. I., Worden, J. R., Jacob, D. J., Hoffman, R. N., Bey, I., and Yantosca, R. M.: Potential of observations from the Tropospheric Emission Spectrometer to constrain continental sources of carbon monoxide, J. Geophys. Res.-Atmos., 108, 4789, doi:10.1029/2003JD003702, 2003.

Kasibhatla, P., Arellano, A., Logan, J. A., Palmer, P. I., and Novelli, P.: Top-down estimate of a large source of atmospheric carbon monoxide associated with fuel combustion in Asia, Geophys. Res. Lett., 29, 1900, doi:10.1029/2002GL015581, 2002.

Kopacz, M., Jacob, D. J., Henze, D. K., Heald, C. L., Streets, D. G., and Zhang, Q.: Comparison of adjoint and analytical Bayesian inversion methods for constraining Asian sources of carbon monoxide using satellite (MOPITT) measurements of CO columns, J. Geophys. Res., Biogeosciences (0148-0227), 114 (d4), D04305 10.1029/2007JD009264, 2009.

Kopacz, M., Jacob, D. J., Fisher, J. A., Logan, J. A., Zhang, L., Megretskaia, I. A., Yantosca, R. M., Singh, K., Henze, D. K., Burrows, J. P., Buchwitz, M., Khlystova, I., McMillan, W. W., Gille, J. C., Edwards, D. P., Eldering, A., Thouret, V., and Nedelec, P.: Global estimates of CO sources with high resolution by adjoint inversion of multiple satellite datasets (MOPITT, AIRS, SCIAMACHY, TES), Atmos. Chem. Phys., 10, 855–876, doi:10.5194/acp-10-855-2010, 2010.

Meirink, J. F., Bergamaschi, P., Frankenberg, C., d'Amelio, M. T. S., Dlugokencky, E. J., Gatti, L. V., Houweling, S., Miller, J. B., Roeckmann, T., Villani, M. G., and Krol, M. C.: Four-dimensional variational data assimilation for inverse modeling of atmospheric methane emissions: analysis of SCIAMACHY observations, J. Geophys. Res., 113, D17301, doi:10.1029/2007JD009740, 2008.

Michalak, A. M., Bruhwiler, L., and Tans, P. P.: A geostatistical approach to surface flux estimation of atmospheric trace gases, J. Geophys. Res., 109, D14109, doi:10.1029/2003JD004422, 2004.

Mueller, K. L., Gourdji, S. M., and Michalak, A. M.: Global monthly averaged CO_2 fluxes recovered using a geostatistical inverse modeling approach: 1. Results using atmospheric measurements, J. Geophys. Res.-Atmos., 113, D21114, doi:10.1029/2007JD009734, 2008.

Palmer, P. I., Jacob, D. J., Jones, D. B. A., Heald, C. L., Yantosca, R. M., Logan, J. A., Sachse, G. W., and Streets, D. G.: Inverting for emissions of carbon monoxide from Asia using aircraft observations over the Western Pacific, J. Geophys. Res.-Atmos., 108, 8828, doi:10.1029/2003JD003397, 2003.

Patra, P. K., Maksyutov, S., Ishizawa, M., Nakazawa, T., Takahashi, T., and Ukita, J.: Interannual and decadal changes in the sea-air CO_2 flux from atmospheric CO_2 inverse modeling, Global Biogeochem. Cy., 19, GB4013, doi:10.1029/2004GB002257, 2005.

Petron, G., Granier, C., Khattatov, B., Lamarque, J. F., Yudin, V., Muller, J. F., and Gille, J.: Inverse modeling of carbon monoxide surface emissions using Climate Monitoring and Diagnostics Laboratory network observations, J. Geophys. Res., 107, 4761, doi:10.1029/2001JD001305, 2002.

Petron, G., Granier, C., Khattatov, B., Yudin, V., Lamarque, J. F., Emmons, L., Gille, J., and Edwards, D. P.: Monthly CO surface sources inventory based on the 2000–

2001 MOPITT satellite data, Geophys. Res. Lett., 31, L21107, doi:10.1029/2004GL020560, 2004.

Peylin, P., Baker, D., Sarmiento, J., Ciais, P., and Bousquet, P.: Influence of transport uncertainty on annual mean and seasonal inversions of atmospheric CO_2 data, J. Geophys. Res., 107, 4385, doi:10.1029/2001JD000857, 2002.

Prado, R. and West, M.: Time Series: Modelling, Computation and Inference, Chapman & Hall/CRC Press, The Taylor Francis Group, London, 2010.

Rayner, P. J. and O'Brien, D. M.: The utility of remotely sensed CO_2 concentration data in surface source inversions, Geophys. Res. Lett., 28, 175–178, 2001.

Rayner, P. J., Scholze, M., Knorr, W., Kaminski, T., Giering, R., and Widmann, H.: Two decades of terrestrial carbon fluxes from a carbon cycle data assimilation system (CCDAS), Global Biogeochem. Cy., 19, GB2026, doi:10.1029/2004GB002254, 2005.

Rodenbeck, C., Houweling, S., Gloor, M., and Heimann, M.: CO_2 flux history 1982–2001 inferred from atmospheric data using a global inversion of atmospheric transport, Atmos. Chem. Phys., 3, 1919–1964, 2003.

Rue, H.: Fast sampling of Gaussian Markov random fields, J. Roy. Stat. Soc. B, 63, 325–338, 2001.

Rue, H. and Held, L.: Gaussian Markov Random Fields: Theory and Applications, 104, Monographs on Statistics and Applied Probability, Chapman & Hall, London, 2005.

Stavrakou, T. and Mueller, J. F.: Grid-based versus big region approach for inverting CO emissions using Measurement of Pollution in the Troposphere (MOPITT) data, J. Geophys. Res., 111, D15304, doi:10.1029/2005JD006896, 2006.

West, M. and Harrison, P.: Bayesian Forecasting and Dynamic Models, 2nd edn., Springer-Verlag, New York, 1997.