Special Issue Paper

Environmetrics

Received: 11 January 2011, Accepted: 21 November 2011, Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/env.1147

Bayesian hierarchical spatio-temporal smoothing for very large datasets†

Matthias Katzfuss[a]* and Noel Cressie[b]

Spatio-temporal statistics is prone to the curse of dimensionality: one manifestation of this is inversion of the data–covariance matrix, which is not in general feasible for very-large-to-massive datasets, such as those observed by satellite instruments. This becomes even more of a problem in fully Bayesian statistical models, where the inversion typically has to be carried out many times in Markov chain Monte Carlo samplers. Here, we propose a Bayesian hierarchical spatio-temporal random effects (STRE) model that offers fast computation: Dimension reduction is achieved by projecting the process onto a basis-function space of low, fixed dimension, and the temporal evolution is modeled using a dynamical autoregressive model in time. We develop a multiresolutional prior for the propagator matrix that allows for unknown (random) sparsity and shrinkage, and we describe how sampling from the posterior distribution can be achieved in a feasible way, even if this matrix is very large. Finally, we compare inference based on our fully Bayesian STRE model with that based on an empirical-Bayesian STRE-model approach, where parameters are estimated via an expectation-maximization algorithm. The comparison is carried out in a simulation study and on a real-world dataset of global satellite CO_2 measurements. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: Bayesian hierarchical modelling; massive datasets; dimension reduction; varying model dimension; global CO_2; remote sensing

1. INTRODUCTION

This article is concerned with spatio-temporal smoothing of very-large-to-massive datasets. The increase in the amount of data being collected has provided statisticians in all disciplines with the challenge of how to cope with the new wealth of data, but it is particularly challenging for spatial and spatio-temporal statistics. Spatial statistical analyses, such as Kriging and maximum likelihood estimation, typically require solving systems of linear equations involving the covariance matrix of the data vector, which quickly becomes infeasible as the size of the dataset increases. For simplicity, we refer to solving these equations as 'inversion' of the data vector's covariance matrix because the computational complexity of the two operations is essentially the same. Computational infeasibility becomes more of a problem when data are also collected over time, which makes for even larger datasets. Because of these special computational challenges for spatial and spatio-temporal statistical analyses, in this article, we take the term 'very large dataset' to mean a dataset size that is on the order of 10^4 to 10^6 observations. Massive datasets are orders of magnitude beyond this.

Here, we develop statistical inference that does not break down as the size of the (spatio-temporal) dataset increases, by working on a reduced-dimensional space. We model the process as a random linear combination of spatial basis functions plus a spatially heterogeneous fine-scale-variation term. Instead of describing the dependence in the data using a spatio-temporal covariance function, we make use of a vector-autoregressive dynamical model for the coefficients of the basis functions. Thus, the temporal dependence in the data is explained by specifying the temporal evolution of a reduced-dimensional spatial process given its past state. This so-called spatio-temporal random effects (STRE) model contains unknown parameters.

We operate in a fully Bayesian (FB) framework, and thus we specify prior distributions for all parameters, where some are calibrated according to the variability in the data. Bayesian inference has a number of advantages: not only does it allow for correct assessment of prediction variability but it also results in a phenomenon called shrinkage, which can be very helpful in situations with high-dimensional parameter spaces. However, computational feasibility is crucial in FB inference, as inferences often rely on computationally intensive Markov chain Monte Carlo (MCMC) simulations from the posterior distribution. In this article, we propose a prior that induces sparsity and shrinkage on the first-order autoregressive parameters describing the temporal evolution of the basis-function coefficients, and we describe how MCMC sampling can be achieved in a computationally feasible way.

Examples of large spatio-temporal datasets are readily available from measurements made by earth-observing satellites. These datasets provide a distinct set of statistical challenges: Despite the large size of daily datasets, the observations can still be sparse relative to the spatial

* *Correspondence to: Matthias Katzfuss, Institut für Angewandte Mathematik, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany. E-mail: katzfuss@gmail.com*

a *Institut für Angewandte Mathematik, Universität Heidelberg, Heidelberg, Germany*

b *Department of Statistics, The Ohio State University, Columbus, OH, U.S.A.*

† *This article is published in Environmetrics as a special issue on Spatio-Temporal Stochastic Modelling (METMAV), edited by Wenceslao González-Manteiga and Rosa M. Crujeiras, University of Santiago de Compostela, Spain.*

domain of interest (often the entire globe). This requires the statistical analyst to take full advantage of spatial and temporal correlations in the true process, to fill spatial gaps. However, no process on the globe will satisfy the stationarity assumptions that are typically made in traditional spatial statistics. The STRE model proposed in this article accounts for these issues, and it can also easily handle remote-sensing measurements with areal footprints and the resulting change-of-support issues. We apply the STRE model to a dataset of global satellite CO_2 measurements: from a fixed window of daily measurements obtained from the Atmospheric InfraRed Sounder (AIRS) on the Aqua satellite, we give a sequence of complete daily maps of global CO_2 fields during this window, together with maps of their associated prediction standard errors.

An extensive review of the literature on dynamical spatio-temporal models in a hierarchical statistical framework can be found in Cresie and Wikle (2011, Chap. 7). We highlight portions of the literature that are especially relevant to our work. First, the large literature on state-space modelling can be seen as part of the hierarchical-statistical-modelling literature by noting that the measurement equation can be viewed as the data model, and the state equation can be viewed as the process model. Consequently, the STRE model referred to earlier is the state equation in a state-space model for time series (Hamilton, 1994, Chap. 13). Shumway and Stoffer (2006, Chap. 6) gave an overview of various types of state-space implementations from a general time-series perspective.

An integral part of any state-space model is the observation matrix (e.g., Shumway and Stoffer, 2006, p. 325), which maps the state variables on the reduced dimension to the process or observations at the original or physical dimension. If a state-space model is applied in a spatio-temporal context, the observation matrix typically consists of known spatial basis functions (e.g., Smith *et al.*, 1996; Kaplan *et al.*, 1998). Here, we propose methodology that allows for the use of any type of (orthogonal or non-orthogonal) spatial basis functions, and we choose bisquare functions for illustration in Sections 3 and 4. Other possible choices for the basis functions include empirical orthogonal functions (e.g., Aubry *et al.*, 1993) and wavelets (e.g., Nychka *et al.*, 2002), but both of these basis-function types are most useful for gridded data. Overviews of possible sets of basis functions in spatial and spatio-temporal applications are given in Antoulas (2005) and Wikle (2010). Alternatively, the observation matrix can be obtained by discretizing a process-convolution model (Higdon, 1998). For the basis-function approach using positive integrable functions (e.g., bisquare functions found in Cressie and Johannesson, 2008), the two approaches are similar because the basis functions can be interpreted as smoothing kernels. Instead of assuming a known observation matrix, Lopes *et al.* (2008) placed a (strong) prior on it.

Another important component of a state-space model is the form of the temporal evolution of the state variables, for which many parameterizations are possible. We assume here that the evolution is linear and first-order Markov, and therefore, the evolution is determined by a single propagator matrix. (For a more general science-based approach to the specification of the temporal evolution, see Wikle and Hooten, 2010.) This allows for Kalman-filter-type inference on the state variables (Kalman, 1960). Sparse parameterizations can be achieved by assuming that the propagator matrix is the identity (which corresponds to a random walk; see, e.g., Stroud *et al.*, 2001) or diagonal (which corresponds to separable autoregressive models; see, e.g., Lopes *et al.*, 2008). Less restrictive parameterizations that still depend on only a small number of parameters can be achieved by deriving the propagator matrix from a discretization of partial differential equations (e.g., Wikle, 2003; Cangelosi and Hooten, 2009; Stroud *et al.*, 2010) or integro-difference equations (e.g., Kot *et al.*, 1996; Wikle and Cressie, 1999; Xu *et al.*, 2005; Dewar *et al.*, 2009). If the dimension of the state space is sufficiently low, it is also possible to include more general lagged-nearest-neighbor models (Wikle *et al.*, 1998), or it might even be possible to leave the propagator matrix completely general. In this article, we take the latter approach, albeit on the reduced-dimensional space. Our formulation is based on Kalman smoothing, not filtering, and it is feasible even if the number of basis functions is moderately large. This is achieved by inducing random shrinkage and sparsity through a multiresolutional prior (which results in a 'soft' lagged-nearest-neighbor approach) inspired by the Minnesota prior in the time-series literature (Litterman, 1986; George *et al.*, 2008). The Minnesota prior shrinks the autoregressive coefficients toward a random-walk model, a feature that is also present in our prior model. We also develop a fast posterior-sampling scheme based on conditional simulation used in spatial statistics (e.g., Cressie, 1993, Sec. 3.6.2).

The methodology proposed in this paper is specifically designed to scale up to very large or massive datasets. Early examples of dimension reduction using basis functions in state-space models applied to large spatio-temporal datasets can be found in Mardia *et al.* (1998), Wikle and Cressie (1999), and Wikle *et al.* (2001). Other approaches to statistical analysis of very large spatio-temporal datasets include multi-resolutional tree-structured models (Johannesson *et al.*, 2007) and predictive-process models (discussed briefly in the spatio-temporal setting by Banerjee *et al.*, 2008); in the latter case, the kriging equations are approximated by replacing the data locations with a smaller number of knots. In the process-convolution framework, the temporal evolution can either be modeled using a spatio-temporal smoothing kernel (e.g., Higdon, 2002) or a dynamical model for the state variables (e.g., Calder *et al.*, 2002).

The specific dynamical spatio-temporal state-space model used in this article is a reduced-rank model called the STRE model (referred to earlier). This approach was proposed by Cressie *et al.* (2010), who were motivated by the spatial-only fixed-rank model of Cressie and Johannesson (2006, 2008). Aside from a strong focus on computational scalability and no requirement for orthogonality of the basis functions, this framework has the added feature of incorporating a fine-scale-variation component (Wikle and Cressie, 1999; Wikle *et al.*, 2001; Cressie and Johannesson, 2008; Jun and Stein, 2008; Kang *et al.*, 2009; Cressie and Kang, 2010). In this article, we generalize the distributional assumptions on this component to allow for spatially heterogeneous variances using a suggestion made by Katzfuss and Cressie (2011). In recent articles, estimation of the STRE-model parameters has relied on method-of-moments estimation (Kang *et al.*, 2010) and expectation-maximization (EM) estimation (Katzfuss and Cressie, 2011); the models are hierarchical, but not FB. In the spatial-only setting, Kang and Cressie (2011) give FB inference for the spatial-random-effects model and its parameters. In the spatio-temporal setting of this article, we propose a multiresolutional sparsity-inducing and shrinkage-inducing prior for the propagator matrix of the basis-function coefficients. Together with priors on the other model parameters, this allows us to carry out FB inference for the STRE model and its parameters, in the context of spatio-temporal smoothing.

The rest of the article is organized as follows. Section 2 describes the methodology: We describe the STRE model, explain the prior distributions assumed for the parameters, and give an overview on how to sample from the posterior distribution in a computationally efficient

way. We then compare our methodology with an empirical-Bayesian, STRE-model approach that uses the EM algorithm for estimating parameters. We make the comparison in a simulation study (Section 3) and in an application to a dataset of global CO_2 measurements (Section 4). Discussion and conclusions are given in Section 5. The Appendix[‡] (not part of the article but included as supplementary material) contains details on the posterior inference and the MCMC algorithm upon which it is based.

2. BAYESIAN SPATIO-TEMPORAL SMOOTHING

2.1. The spatio-temporal random-effects model

We are interested in a spatio-temporal process $\{Y_t(\mathbf{s}):\ \mathbf{s} \in D_s,\ t \in 1, 2, \ldots\}$ on a continuous spatial domain D_s and at discrete time points $\{1, 2, \ldots\}$. As is often done in spatial statistics, we assume that the process $Y_t(\cdot)$ can be decomposed as follows,

$$Y_t(\mathbf{s}) = \mu_t(\mathbf{s}) + v_t(\mathbf{s}),\ \mathbf{s} \in D_s,\ t = 1, 2, \ldots, \tag{1}$$

where $\mu_t(\cdot)$ is large-scale trend, and $v_t(\cdot)$ accounts for spatial (and here, temporal) correlation. In what follows, we assume that $\mu_t(\cdot) := \mathbf{x}_t(\cdot)'\boldsymbol{\beta}_t$, which is a linear combination of p known covariates, $x_{t,1}(\cdot), \ldots, x_{t,p}(\cdot)$

Although our interest is in inference on $Y_t(\cdot)$, we cannot observe it perfectly. Our measurements are affected by (additive) measurement error and cannot be taken at every $(\mathbf{s}, t) \in D_s \times \{1, 2, \ldots\}$. Our focus in this article is on smoothing, namely after collecting the $n_1 + \ldots + n_T$ measurements,

$$Z_t(\mathbf{s}_{i,t}) = Y_t(\mathbf{s}_{i,t}) + \epsilon_t(\mathbf{s}_{i,t}),\ i = 1, \ldots, n_t,\ t = 1, \ldots, T,$$

we are interested in predicting the unknown quantity $Y_t(\mathbf{s})$ at every $\mathbf{s} \in D_s$ and for all time points $t = 1, \ldots, T$.

We assume that the measurement-error process, $\epsilon_t(\cdot)$, is distributed as,

$$\epsilon_t(\cdot) \sim N(0, \sigma^2_{\epsilon,t} v_{\epsilon,t}(\cdot)),\ t = 1, 2, \ldots,$$

independent of $Y_t(\cdot)$, and independent in time and space. For identifiability reasons, both the measurement-error variance $\sigma^2_{\epsilon,t}$ and the function $v_{\epsilon,t}(\cdot) > 0$ will be assumed known for the remainder of this article. Whereas it is common that $v_{\epsilon,t}(\cdot)$ is known (e.g., $v_{\epsilon,t}(\cdot) \equiv 1$), there may be no information on $\sigma^2_{\epsilon,t}$; in this case, $\sigma^2_{\epsilon,t}$ can be estimated from the data via an estimation technique based on extrapolating the variogram (Kang *et al.*, 2009). If prior instrument-calibration experiments have been carried out, $\sigma^2_{\epsilon,t}$ may in fact be known as well.

To exploit the spatio-temporal correlation in $Y_t(\cdot)$, we now specify a covariance function for the measurements and use this to form the covariance matrix, Σ, of the vector of all measurements, $\mathbf{Z}_{1:T} := [\mathbf{Z}_1', \ldots, \mathbf{Z}_T']'$, where $\mathbf{Z}_t := [Z_t(\mathbf{s}_{1,t}), \ldots, Z_t(\mathbf{s}_{n_t,t})]'$. Let $n_+ := \sum_{t=1}^T n_t$ denote the total number of observations taken at all time points combined. Now, statistical inference typically requires inversion of the $n_+ \times n_+$ matrix Σ, possibly repeatedly so at successive iterations of an estimation procedure. Because the inversion of a general $n_+ \times n_+$ matrix requires on the order of n_+^3 computations, this is infeasible for the very large spatio-temporal datasets of interest here, where $\{n_t\}$ (and possibly also T) are very large.

To achieve both computational feasibility and a flexible nonstationary model, we assume an STRE model (Cressie *et al.*, 2010) for the component $v_t(\cdot)$ in (1):

$$v_t(\mathbf{s}) = \mathbf{b}_t(\mathbf{s})'\boldsymbol{\eta}_t + \delta_t(\mathbf{s}),\ \mathbf{s} \in D_s,\ t = 1, 2, \ldots, \tag{2}$$

where $\mathbf{b}_t(\cdot) := [b_{t,1}(\cdot), \ldots, b_{t,r_t}(\cdot)]'$ is an r_t-dimensional vector of known spatial basis functions; $\boldsymbol{\eta}_t$ is a random coefficient vector of length r_t; and the fine-scale variation,

$$\delta_t(\cdot) \sim N(0, \sigma^2_{\delta,t} v_{\delta,t}(\cdot)), \tag{3}$$

is *a priori* independent of $\{\boldsymbol{\eta}_t\}$ and independent in time and space. The basis functions in $\mathbf{b}_t(\cdot)$ do *not* have to be orthogonal. However, it is recommended that they be of different spatial resolutions $1, \ldots, C$, which can capture different scales of spatial variation (Cressie and Johannesson, 2008). The fine-scale variation, $\{\delta_t(\cdot):\ t = 1, 2, \ldots\}$, can be viewed as an attempt to account for the error introduced by the dimension reduction. The temporal evolution of $\{Y_t(\cdot)\}$ is determined by a vector-autoregressive (VAR) model for $\{\boldsymbol{\eta}_t : t = 0, 1, \ldots, T\}$:

$$\boldsymbol{\eta}_t | \boldsymbol{\eta}_{t-1} \sim N_{r_t}(H_t \boldsymbol{\eta}_{t-1}, U_t),\ t = 1, \ldots, T, \tag{4}$$

with initial state $\boldsymbol{\eta}_0 \sim N_{r_0}(\mathbf{0}, K_0)$. The $r_t \times r_{t-1}$ matrix H_t and the $r_t \times r_t$ matrix U_t will be referred to as the propagator matrix and the innovation covariance matrix, respectively.

Although models (2) and (4) have been introduced with much generality, in this article we assume henceforth that $\mathbf{b}_t(\cdot) \equiv \mathbf{b}(\cdot)$, $r_t \equiv r$, $\sigma^2_{\delta,t} \equiv \sigma^2_\delta$, $v_{\delta,t}(\cdot) \equiv v_\delta(\cdot)$, $H_t \equiv H$, and $U_t \equiv U$, during the period $\{0, \ldots, T\}$. Strictly speaking, this is not needed in an FB framework; however, assumptions of this sort allow practical identifiability and result in well mixed MCMC samples from the posterior distribution.

As the number of basis functions r is much smaller than the sample sizes $\{n_t\}$, assumption (2) results in dimension reduction because the computational complexity for processing the measurements taken at time point t is reduced to $\mathcal{O}(n_t r^3)$ from $\mathcal{O}(n_t^3)$; see Cressie *et al.*

[‡]Supplementary material may be found in the online version of this article.

(2010). Additionally, the VAR model (4) is a state-space model that allows for sequential (in time) processing of data observed at subsequent time points via Kalman-filter-type and Kalman-smoother-type algorithms. This ensures that the computational cost of inference on the process components $\{\boldsymbol{\eta}_t\}$ and $\{\delta_t(\mathbf{s})\}$ (given the unknown parameters) for all observed time points $t = 1, \ldots, T$ is still linear in n_+; that is, inference for our model can scale up to very-large-to-massive datasets.

In summary, we have introduced the *data model*,

$$\mathbf{Z}_t = \mathbf{Y}_t + \boldsymbol{\epsilon}_t,\ t = 1, \ldots, T, \tag{5}$$

where, based on the STRE model, we can write,

$$\mathbf{Y}_t = X_t\boldsymbol{\beta}_t + B_t\boldsymbol{\eta}_t + \boldsymbol{\delta}_t,\ t = 1, \ldots, T. \tag{6}$$

Together, (4) and (6) specify a *process model* for $\{\mathbf{Y}_t\}$. In (5) and (6), we have stacked scalars into vectors and row vectors into matrices such that, for example, the ith row of the $n_t \times r$ matrix B_t is given by $\mathbf{b}(\mathbf{s}_{i,t})'$. The corresponding covariance matrices are $K_t := \text{var}(\boldsymbol{\eta}_t)$, $t = 1, \ldots, T$, and

$$D_t := \text{var}(\boldsymbol{\delta}_t + \boldsymbol{\epsilon}_t) = \sigma_\delta^2 V_{\delta,t} + \sigma_{\epsilon,t}^2 V_{\epsilon,t},\ t = 1, \ldots, T,$$

which is a diagonal matrix with $V_{\delta,t} := \text{diag}(v_\delta(\mathbf{s}_{1,t}), \ldots, v_\delta(\mathbf{s}_{n_t,t}))$ and $V_{\epsilon,t} := \text{diag}(v_\epsilon(\mathbf{s}_{1,t}), \ldots, v_\epsilon(\mathbf{s}_{n_t,t}))$.

The process model is not fully specified yet. We often want to predict $Y_t(\mathbf{s})$ at a set of spatial locations that is different from the measurement locations. Here, we assume (without loss of generality) that the set of prediction locations at time t is a *superset* of the measurement locations observed at time t, so that we can write $\mathbf{Y}_t = M_t\mathbf{Y}_t^P$, $t = 1, \ldots, T$, where $\mathbf{Y}_t^P$ is the process vector of all $m_t > n_t$ prediction locations for time t. This is achieved by allowing the observation locations to be included in the set of all prediction locations. We assume that there are no duplicate measurements, and hence M_t is an $n_t \times m_t$ incidence matrix of 0s except for one 1 in each row. In practice, the set of prediction locations has often been a fine grid over the spatial domain of interest, resulting in $m_t \equiv m$, where the data locations are moved to their nearest respective grid cells.

We use a superscript P (as in 'prediction') when a vector or matrix has been obtained by evaluating appropriate processes at all prediction locations, so that the process model is,

$$\mathbf{Y}_t^P = X_t^P\boldsymbol{\beta}_t + B_t^P\boldsymbol{\eta}_t + \boldsymbol{\delta}_t^P,\ t = 1, \ldots, T, \tag{7}$$

where $\{\boldsymbol{\eta}_t\}$ satisfies (4). This implies that $X_t = M_tX_t^P$, $B_t = M_tB_t^P$, and $\boldsymbol{\delta}_t = M_t\boldsymbol{\delta}_t^P$. The diagonal matrix $\text{var}(\boldsymbol{\delta}_t^P) =: \sigma_\delta^2 V_{\delta,t}^P$ is an $m_t \times m_t$ matrix, where $V_{\delta,t}^P$ will be modeled succeedingly.

2.2. Prior distributions

Until now, we have implicitly assumed a *known* vector of process-model parameters $\boldsymbol{\theta_P}$, which contains the trend coefficients $\{\boldsymbol{\beta}_t : t = 1, \ldots, T\}$, the fine-scale-variation variance σ_δ^2, (the parameters in) the function $v_\delta(\cdot)$, and the elements defining the matrices that describe the VAR process, K_0, H, and U. Of course, $\boldsymbol{\theta_P}$ will rarely be known in practice. We could take an empirical-Bayesian approach to inference, in which we estimate the parameters either via a method-of-moments technique (Wikle and Cressie, 1999; Kang *et al.*, 2010) or via the EM algorithm (Xu and Wikle, 2007; Fassò and Cameletti, 2009; Katzfuss and Cressie, 2011). Instead, in this article, we take a Bayesian approach and specify prior distributions for all unknown parameters (e.g., Wikle *et al.*, 1998). This results in a *parameter model*, which, together with the data model (5) and the process model (7) given earlier, leads to posterior inference via Bayes' theorem. Recall that our goal is smoothing; inference is implemented using MCMC simulations described in Section 2.4 and the Appendix.

All parameters in $\boldsymbol{\theta_P}$ are assumed to be *a priori* independent, unless specifically stated otherwise. For the parameters $\{\boldsymbol{\beta}_t\}$ and σ_δ^2, we assume virtually noninformative priors (see Appendix for details). The prior distributions for the covariance matrices K_0 and U are each taken to be a multiresolutional Givens-angle prior (Kang and Cressie, 2011). As this prior distribution has been considered in detail in previous work, we only give a brief review in the Appendix.

The function $v_\delta(\cdot)$ determines the form of the heterogeneity of the fine-scale-variation variance in (3), namely $\text{var}(\delta(\cdot)|\sigma_\delta^2, v_\delta(\cdot)) = \sigma_\delta^2 v_\delta(\cdot)$. Following a suggestion made by Katzfuss and Cressie (2011), we assume a stochastic volatility model of the form,

$$v_\delta(\cdot) := \exp\{\mathbf{b}_\delta(\cdot)'\boldsymbol{\eta}_\delta\}, \tag{8}$$

where $\mathbf{b}_\delta(\cdot)$ is a known vector of r_δ basis functions and, for example, could be a subvector of $\mathbf{b}(\cdot)$. The prior distribution on $v_\delta(\cdot)$ is induced by $\boldsymbol{\eta}_\delta \sim N_{r_\delta}(\mathbf{0}, \sigma_{\eta_\delta}^2 I_{r_\delta})$, where $\sigma_{\eta_\delta}^2$ is a *known* hyperparameter. This model allows for flexible estimation of the heterogeneity (in space), in that $v_\delta(\mathbf{s})$ can multiplicatively modify the overall fine-scale-variation variance, σ_δ^2, at any location $\mathbf{s} \in D_s$. The exponential function in (8) ensures that the resulting variance of $\delta(\cdot)$ is positive, and the prior mean, $E(\boldsymbol{\eta}_\delta) = \mathbf{0}$, allows shrinkage of the resulting variance of $\delta(\cdot)$ toward the overall variance parameter, σ_δ^2, at any point in space. By modelling the function (on the log-scale) as a linear combination of basis functions, we ensure fast computation even when the function has to be evaluated at a large number of observed or prediction locations. The hyperparameter $\sigma_{\eta_\delta}^2$ can be chosen in accordance with prior beliefs on how different the fine-scale variation is expected to be in different parts of the spatial domain of interest. Consider the variance ratio $R := \text{var}(\delta(\mathbf{s}_1))/\text{var}(\delta(\mathbf{s}_2)) = v_\delta(\mathbf{s}_1)/v_\delta(\mathbf{s}_2)$, where $\mathbf{s}_1$ and $\mathbf{s}_2$ are chosen to be locations at the centres of two distant (normalized) basis functions, so that $(\mathbf{b}_\delta(\mathbf{s}_1) - \mathbf{b}_\delta(\mathbf{s}_2))'(\mathbf{b}_\delta(\mathbf{s}_1) - \mathbf{b}_\delta(\mathbf{s}_2)) \approx 2$. This implies that, approximately, $\exp\{R\} \sim N(0, 2\sigma_{\eta_\delta}^2)$. When $1/2$ and 2 are chosen as the lower and upper endpoints, respectively, of a 95% credible interval for R, this results in a value of $\sigma_{\eta_\delta}^2 \approx 0.25^2$ for the hyperparameter.

2.3. The prior on the propagator matrix H

Let us now turn to the prior assumptions for the propagator matrix H. We first develop a two-stage prior that ensures that the full conditional distribution of $\mathbf{h} := \text{vec}(H')$ is available in closed form (see Appendix), and then we give some insight into the ramifications of this prior specification. To motivate our prior on H, note that the impact of H on the temporal evolution of the process is plainly obvious from,

$$E(\eta_{t,i}|\mathbf{h}_i, \boldsymbol{\eta}_{t-1}) = \mathbf{h}_i'\boldsymbol{\eta}_{t-1} = \textstyle\sum_{j=1}^{r} h_{ij}\eta_{t-1,j},\ t = 1, 2, \ldots,$$

where $\mathbf{h}_i'$ denotes the ith row of H. Thus, h_{ij} describes the autoregressive effect of $\eta_{t-1,j}$ on $\eta_{t,i}$; intuitively, we want this effect to diminish as the (physical) distance between the jth basis function and the ith basis function increases. This intuition is complicated by the fact that we want the basis functions in $\mathbf{b}(\cdot)$ to be made up of C (say) resolutions. We can write H (after appropriate ordering) as a block matrix,

$$H =: \left[\begin{array}{c|c|c} H_{11} & \cdots & H_{1C} \\ \hline \vdots & \ddots & \vdots \\ \hline H_{C1} & \cdots & H_{CC} \end{array}\right], \tag{9}$$

where the block H_{kl} contains the elements of H that describe how the basis-function coefficients of resolution k at time point t are affected by the basis-function coefficients of resolution l at the previous time point $t-1$.

In light of this role that the elements of H play on the temporal evolution of the process, we assume that the (i, j)th element of H has the (conditional) prior distribution,

$$h_{ij}|\boldsymbol{\theta}_H \overset{\text{ind}}{\sim} N(\mu_{c_i} I(i=j), \tau^2_{c_i,c_j}\, g(d_{ij}; \alpha_{c_i,c_j}, \gamma_{c_i,c_j})^2),\ i = 1, \ldots, r,\ j = 1, \ldots, r, \tag{10}$$

where c_i denotes the resolution to which the ith basis function belongs; the quantity $d_{ij} \in [0, 1]$, with $\max\{d_{ij}\} = 1$, is the normalized distance between the centres of the ith and the jth basis functions; $\boldsymbol{\theta}_H$ is a vector of parameters describing the prior distribution of H that consists of $\{\mu_k\colon k = 1, \ldots, C\}$, $\{\tau^2_{kl}\colon k, l = 1, \ldots, C\}$, $\{\alpha_{kl}\colon k, l = 1, \ldots, C\}$, and $\{\gamma_{kl}\colon k, l = 1, \ldots, C\}$; and

$$g(d; \alpha, \gamma) := \begin{cases} 1 - (d/\alpha)^{\exp(\gamma)}, & d \leqslant \alpha \\ 0, & d > \alpha \end{cases} \tag{11}$$

is a function of normalized distance with (random) range parameter $\alpha \in [0, 1]$ and (random) shape parameter $\gamma \in \mathbb{R}$ (see the left panel of Figure 1; more details are given later in this subsection). Note that, to include the case $\alpha = 0$, we define $0/0 = 0$ in (11).

At the second level of the prior distribution on H, we assume that all parameters in $\boldsymbol{\theta}_H$ are independently distributed according to,

$$\begin{aligned} \mu_k &\overset{\text{ind}}{\sim} N(1, \sigma^2_{\mu,k}),\ k = 1, \ldots, C, \\ \tau^2_{kl} &\overset{\text{ind}}{\sim} IG(a_{\tau,kl}, b_{\tau,kl}),\ k, l = 1, \ldots, C, \\ \alpha_{kl} &\overset{\text{iid}}{\sim} U(0, 1),\ k, l = 1, \ldots, C, \\ \gamma_{kl} &\overset{\text{iid}}{\sim} N(\mu_\gamma, \sigma^2_\gamma),\ k, l = 1, \ldots, C, \end{aligned} \tag{12}$$

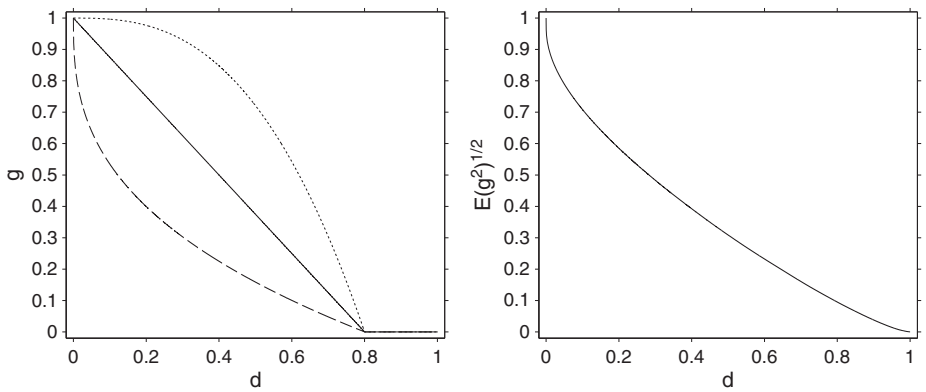

Figure 1. Left panel: The function $g(\cdot; \alpha, \gamma)$ for $\alpha = 0.8$ and $\gamma = 0$ (solid line), $\gamma = -1$ (dashed), and $\gamma = 1$ (dotted). Right Panel: The function $\sqrt{E(g(\cdot; \alpha_{kl}, \gamma_{kl})^2)}$ describes the shrinkage (on the standard-deviation scale) induced by the prior on H as a function of the basis-function distance; see (14).

where are all parameters specifying these distributions are fixed, as follows: First, the choice of $E(\mu_k) = 1$ is based on our desire to centre the noninformative prior of H at the identity matrix, which is a random-walk model. Second, the parameters $\{\gamma_{kl}\}$ determine the shape of (11) on the interval $(0, \alpha_{kl})$. A natural centering for these parameters is $\mu_\gamma = 0$, as $g(\cdot\,; \alpha, \gamma = 0)$ is a straight line from the point $(0, 1)$ to $(\alpha, 0)$. To find a good value for σ^2_γ, consider that square-root distances, absolute distances, and squared distances are often used in practice. To ensure that these values are contained in an *a priori* 95% credible interval for the exponent of the distance d in the function (11), we set $\sigma^2_\gamma = 0.5^2$, so that the endpoints of the credible interval are given by $1/e \approx 0.37$ and $e \approx 2.72$ (see the dashed and dotted lines in the left panel of Figure 1). Finally, the remaining parameters $\{\sigma^2_{\mu,k}\}$, $\{a_{\tau,kl}\}$, and $\{b_{\tau,kl}\}$ are calibrated to the data through an initial point estimate of H, such as the EM estimate (see Appendix for details).

We shall now interpret the prior assumptions on H made earlier, and discuss their ramifications. As noted earlier, the prior distribution for H implies that the temporal evolution of $\{\boldsymbol{\eta}_t\}$ is *a priori* centred on the random walk, $E(H) = E(E(H|\{\mu_k\})) = I_r$, because of the assumption that $E(\mu_k) = 1$, $k = 1, \ldots, C$, in (12). In light of this, it should be noted that our prior for H can be considered a noninformative prior. It only uses information on where the basis functions are located in the spatial domain, D_s, and to which resolutions the basis functions belong. The prior is ideally suited for applications in which no prior information about the temporal evolution of the process is available, or in situations where the goal is to validate or check existing scientific models about the temporal evolution of the process from a complete map of predictions derived from data. Our prior mean meets the constant-mean and mass-balance requirements for propagator matrices formulated in Gelpke and Künsch (2001).

Integrating over the parameter $\{\tau_{kl}\}$, we have,

$$h_{ij}|\mu_{c_i}, \alpha_{c_i,c_j}, \gamma_{c_i,c_j} \sim t_{2a_{\tau,c_ic_j}}\left(\mu_{c_i} I(i=j), \frac{b_{\tau,c_ic_j}}{a_{\tau,c_ic_j}-1} g(d_{ij}; \alpha_{c_i,c_j}\gamma_{c_i,c_j})^2\right) , \tag{13}$$

where $t_\nu(\mu, \sigma^2)$ denotes a generalized t-distribution with ν degrees of freedom, location parameter μ, and scale parameter σ (Bishop, 2006, Sec. 2.3.7).

The off-diagonal elements of H are shrunk towards zero (or even set equal to zero), depending on the distance and the resolutions of the corresponding basis functions. The marginal variance of h_{ij} is,

$$\mathrm{var}(h_{ij}) = \mathrm{var}E(h_{ij}|\boldsymbol{\theta}_H) + E\,\mathrm{var}(h_{ij}|\boldsymbol{\theta}_H) = \sigma^2_{\mu,c_i} I(i=j) + \frac{b_{\tau,c_ic_j}}{a_{\tau,c_ic_j}-1} E\left(g(d_{ij}; \alpha_{c_i,c_j}, \gamma_{c_i,c_j})^2\right) , \tag{14}$$

where the square root of the expectation on the right-hand side (i.e., on the standard-deviation scale) is shown in the right panel of Figure 1. Thus, the prior variance of h_{ij} is monotone decreasing as a function of d_{ij}.

The prior distributions on the parameters $\{\alpha_{kl}\}$ and $\{\gamma_{kl}\}$ can be interpreted as controlling the sparsity and the shrinkage on the elements of H, respectively. The range parameters $\{\alpha_{kl}\}$ induce sparsity in H in that, assuming a uniform prior on α_{c_i,c_j},

$$P(h_{ij} = 0) = P(g(d_{ij}; \alpha_{c_i,c_j}, \gamma_{c_i,c_j}) = 0) = P(\alpha_{c_i,c_j} \leq d_{ij}) = d_{ij} \ .$$

Hence, it becomes more and more likely that h_{ij} is zero with increasing distance between the centres of basis functions i and j. That is, we are essentially specifying the (random) dimension of the parameter space by generating $\{\alpha_{kl}\}$. Other choices than a uniform distribution for the priors of $\{\alpha_{kl}\}$ are possible and can result in interesting dynamical structure for the model. For example, if the priors on $\{\alpha_{kl}\}$ are all point masses at zero, then $g(d; 0, \gamma) = I(d = 0)$, which results in a diagonal H with parameter μ_k down the diagonal of H_{kk}, $k = 1, \ldots, C$. Even this simple multiresolutional H induces complex spatio-temporal dependence, since $\mathrm{cov}(\boldsymbol{\eta}_{t+1}, \boldsymbol{\eta}_t) = HK_t$ has non-trivial, nonstationary cross-dependence.

Given $\{\tau_{kl}\}$ and $\{\alpha_{kl}\}$, the parameters $\{\gamma_{kl}\}$ control the amount of shrinkage of the nonzero elements of H as a function of the basis-function distances. If the parameter γ_{kl} is nonnegative, then conditional on α_{kl}, the function $g_{kl}(\cdot)$ is concave on the interval $(0, \gamma_{kl})$; for nonpositive γ_{kl}, the function is convex on the interval (see the left panel of Figure 1). Therefore, very large values for $\{\gamma_{kl}\}$ make for little shrinkage (up to distances smaller than $\{\alpha_{kl}\}$).

Lastly, the marginal covariance between two elements, h_{i_1,j_1} and h_{i_2,j_2}, of H, after integrating out the prior distributions on $\boldsymbol{\theta}_H$ given by (12), is,

$$\mathrm{cov}(h_{i_1,j_1}, h_{i_2,j_2}) = \{\mathrm{var}(h_{i_1,j_1} + h_{i_2,j_2}) - \mathrm{var}(h_{i_1,j_1}) - \mathrm{var}(h_{i_2,j_2})\}/2 = \sigma^2_{\mu,k} I(i_1 = j_1) I(i_2 = j_2) I(c_{i_1} = c_{i_2} = k).$$

This means that the only *a priori* nonzero correlations between elements of H are those where both are diagonal elements within the same resolution. However, as mentioned earlier, all elements within each block H_{kl} are *a priori* statistically dependent, for $k, l = 1, \ldots, C$.

Note that our prior for H makes use of distances between basis functions. This distance is quite intuitive if the basis functions have a clear 'centre' (e.g., bisquare functions). For other basis functions, one could use the 'centre of energy' (e.g., Wickerhauser, 1994, p. 164), defined as $\int \mathbf{s}\, b(\mathbf{s})^2 \mathrm{d}\mathbf{s} / \int b(\mathbf{s})^2 \mathrm{d}\mathbf{s}$ for a continuous basis function $b(\cdot)$. However, this centre of energy might not be easily interpretable for basis functions with non-compact support (e.g., Fourier functions or empirical orthogonal functions), and so our prior for H might not be generally applicable for those functions (unless $\alpha_{kl} \equiv 0$ for all $k, l = 1, \ldots, C$, in which case H is diagonal).

Because H refers to a reduced-dimensional space, it might seem unnecessary to look for sparsity in the $r \times r$ matrix H. However, the number of basis functions r, can be moderately large, and it is usually larger than T (e.g., in Section 4, we have $r = 380$ and $T = 16$). This may result in practical nonidentifiability, for which regularization (here, sparsity and shrinkage) would be needed (see Appendix for more details).

2.4. Markov chain Monte Carlo inference

For a set of generic vectors $\{\mathbf{x}_t\}$, define $\mathbf{x}_{t_1:t_2} := [\mathbf{x}'_{t_1}, \ldots, \mathbf{x}'_{t_2}]'$. Recall that our goal in this spatio-temporal context is smoothing, not filtering. After having observed data $\mathbf{Z}_{1:T} = \mathbf{z}_{1:T}$, FB inference is based on the posterior distribution of $\mathbf{Y}^P_{1:T}$ (i.e., that of $\boldsymbol{\eta}_{1:T}$ and $\boldsymbol{\delta}^P_{1:T}$) and the unknown parameters, given the data. Unfortunately, this posterior distribution is not available in closed form. Instead, we sample from the posterior distribution via MCMC simulation. As the methodology developed in this article is intended to be used on very large (or even massive) datasets, computational feasibility and speed are of great concern. We employ a Gibbs sampler (Geman and Geman, 1984) with some Metropolis–Hastings (Metropolis *et al.*, 1953; Hastings, 1970) updates where necessary. In this section, we give an overview of the techniques used to sample the unknowns in the MCMC; details are given in the Appendix.

First, consider the basis-function coefficients $\boldsymbol{\eta}_{0:T}$. Owing to their strong temporal dependence, it is not advised to update each $\boldsymbol{\eta}_t$ individually, which would result in slow convergence of the MCMC. Instead, we update the entire vector $\boldsymbol{\eta}_{0:T}$ at once, jointly with $\boldsymbol{\beta}_{1:T}$ and $\boldsymbol{\delta}^P_{1:T}$, by modifying a technique called the forward-filtering, backward-sampling algorithm (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994). The number of operations required for this algorithm is linear in each n_t for each iteration of the MCMC, which is essential to scalability of the algorithm as a whole.

Updating the propagator matrix H (or, equivalently, $\mathbf{h} := \text{vec}(H')$) and its random hyperparameters poses its own challenges. As $\mathbf{h}$ is an r^2-dimensional vector, direct sampling becomes impossible when r, the number of basis functions, is moderately large. To avoid this, we employ a technique similar to conditional simulation used in geostatistics (for details on spatial conditional simulation, see, e.g., Cressie, 1993, Sec. 3.6.2). Considering the hyperparameters $\{\alpha_{kl}\}$, which control the sparsity of H, we notice that there is actually a change of dimension in the parameter space of $\mathbf{h}$. Depending on the value of $\{\alpha_{kl}\}$ sampled in the MCMC algorithm, a number of elements of $\mathbf{h}$ will each have a variance of zero, and their full conditional distributions will be point masses at zero. If we marginalize over $\mathbf{h}$ when updating $\{\alpha_{kl}\}$ and use the conditional-simulation technique for sampling $\mathbf{h}$ mentioned earlier, we avoid having to use an explicit reversible-jump MCMC; see the Appendix.

We recommend updating those parameters with analytically intractable conditional distributions (i.e., $\boldsymbol{\eta}_\delta$, $\{\alpha_{kl}\}$, and the parameters in K_0 and U) using the adaptive Metropolis–Hastings algorithm of Haario *et al.* (2001); see the Appendix for details. The full conditional distributions of $\{\boldsymbol{\beta}_t\}$ and σ^2_δ are also given in the Appendix.

Let $\boldsymbol{\theta}$ be a vector containing all unknowns, $\boldsymbol{\eta}_{0:T}$, $\boldsymbol{\delta}^P_{1:T}$, $\boldsymbol{\theta}_P$, and $\boldsymbol{\theta}_H$. Samples from the posterior distribution of $\boldsymbol{\theta}$ given the data are obtained as follows: We begin the MCMC sampler with some starting value $\boldsymbol{\theta}^{[0]}$, and then we obtain $\boldsymbol{\theta}^{[l]}$, $l = 1, 2, \ldots$, by updating each component of $\boldsymbol{\theta}$ given the most recent value of all other components, as described in the Appendix. After a 'burn-in' of L_b iterations, the algorithm should be sampling from the target (joint posterior) distribution. From a total number of L_a iterations, the first L_b are discarded, and we consider the set $\{\boldsymbol{\theta}^{[L_b+1]}, \ldots, \boldsymbol{\theta}^{[L_a]}\}$ to be a sample from the joint posterior distribution of all unknowns given the data.

To return to the issue of scalability of the algorithm for very large datasets, we note that the number of computations required at each iteration of the MCMC is linear in each n_t. However, each update of H requires inversion of a sparse $rT \times rT$ matrix with at most $rT(r+T-1)$ nonzero elements (see the Appendix). This implies that if r and T are both very large, the algorithm can become fairly slow.

3. SIMULATION STUDY: FULLY BAYESIAN FIXED RANK SMOOTHING VERSUS EXPECTATION-MAXIMIZATION FIXED RANK SMOOTHING

Instead of specifying prior distributions for all parameters in the model described in Section 2.1, we could pursue empirical-Bayesian inference via fixed rank smoothing (FRS), as described in Cressie *et al.* (2010). To do this, we must first estimate the parameters, and then we obtain the posterior distribution of $\mathbf{Y}^P_{1:T}$ given the data, by assuming that all parameters are known and fixed at their estimated values. We can estimate the parameters in the STRE model using maximum-likelihood estimation via the EM algorithm, which is shown by Katzfuss and Cressie (2011) to be preferable to the binned-method-of-moments estimation of Cressie *et al.* (2010) when the data are Gaussian. This EM-FRS procedure is therefore a natural candidate for comparison to the fully Bayesian inference (FB-FRS) proposed in Section 2. In this section, we carry out a simulation study to assess parameter estimation, accuracy of predictions, and the accuracy of inferred prediction uncertainties.

3.1. Simulation setup

The simulated data are meant to be a simplistic version of satellite data. The spatial domain is one-dimensional, $D_s := \{1, \ldots, 256\}$, and there are $T = 16$ time points. The 'satellite' has a repeat cycle of two time units. The two tracks of the satellite have a width of 64: For t odd, the tracks are $\{1, \ldots, 64\}$ and $\{129, \ldots, 192\}$; for t even, the tracks are $\{65, \ldots, 128\}$ and $\{193, \ldots, 256\}$. To simulate non-retrievals because of cloud cover and other problems, 50% of the values within each track at each time point are declared missing at random. This results in $n_t = 64$ observations at each time point.

The basis functions we use are bisquare functions,

$$f_{\text{bi}}(\mathbf{s}) := \{1 - (\|\mathbf{s} - \mathbf{c}\|/w)^2\}^2 I(\|\mathbf{s} - \mathbf{c}\| < w) \tag{15}$$

where $\mathbf{c}$ is the centre point, $w > 0$ is the specified range, and $I(\cdot)$ is an indicator function. In this simulation study, we have $r = 5$ bisquare basis functions from $C = 2$ resolutions, as depicted in Figure 2. The one basis function of the first resolution has a range of $w = 144$ and is centred at 128. The four basis functions of the second resolution have a range of $w = 38$ and are centred at 32.5, 96.5, 160.5, and 224.5, respectively.

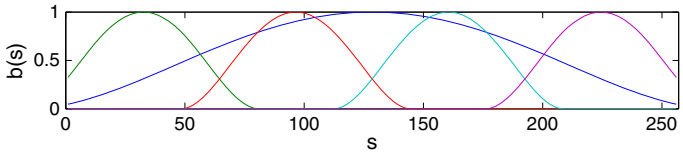

Figure 2. The five basis functions ($r = 5$) of two resolutions ($C = 2$) used in the simulation study.

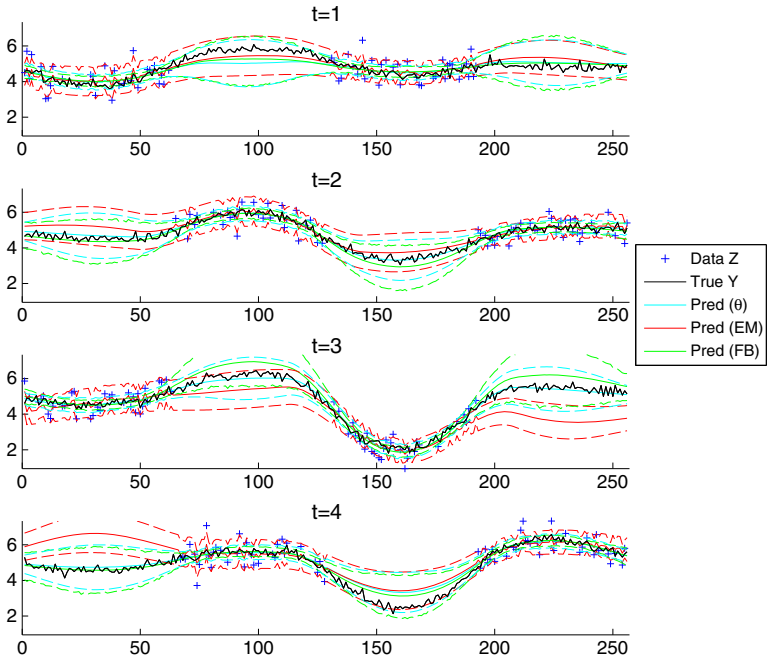

Figure 3. One realization of the data (blue crosses) observed at the first four time points in the simulation study for SNR=2. Also shown are FRS predictions using the true parameter values (light blue), FRS predictions using the EM parameter estimates (red), and Bayesian posterior means (green); dashed lines are the respective 95% credible/prediction intervals. The true process values are shown in black.

With the exception of the ranges and centre points of the basis functions, this setup is exactly the same as the one used in the simulation study given in Katzfuss and Cressie (2011). The parameters used in the simulation are also calibrated in the same way as in that article. The true matrix parameters are calibrated to match as closely as possible (as measured by the Frobenius norm) a stationary exponential spatial covariance of the form $\mathrm{cov}(v_t(i), v_t(j)) = \exp(-|i-j|/25)$ and a lag-1 temporal correlation of 0.8 (see Cressie *et al.*, 2010, Sec. 3, for more details). The fine-scale-variation proportion is set to 0.05 (where this calibration is based on $\boldsymbol{\eta}_\delta \equiv \mathbf{0}$), which results in $\sigma^2_\delta = 0.0297$, held constant over time. The function $v_\delta(\cdot) = \exp\{\mathbf{b}_\delta(\cdot)'\boldsymbol{\eta}_\delta\}$ is determined by two bisquare functions of range $w = 96$, centred at 64 and 192, and we set $\boldsymbol{\eta}_\delta = [-0.25, 0.25]'$. For the EM algorithm, $\boldsymbol{\eta}_\delta$ was estimated as described in Katzfuss and Cressie (2011, Sec. 3.4), but here we consider it to be a fixed parameter (i.e., we do not assume a prior distribution for $\boldsymbol{\eta}_\delta$ in the EM algorithm).

The measurement-error variance is also constant over time. It is determined by the signal-to-noise ratio (SNR; defined as in Section 3.1 of Cressie *et al.*, 2010, a calibration that is based on $\boldsymbol{\eta}_\delta \equiv \mathbf{0}$), for which we have chosen two levels: SNR=2, resulting in $\sigma^2_\epsilon = 0.2968$, and SNR=5, resulting in $\sigma^2_\epsilon = 0.1187$. The variance σ^2_ϵ is assumed known for both the FB-FRS and the EM-FRS procedures. Finally, a constant mean of $\mu = 5$ is chosen (i.e., $x_t(\mathbf{s}) \equiv 1$ and $\beta_t \equiv 5$). An example of the data simulated from the STRE model (Section 2.1) is shown in Figure 3. (We only show the first four time points; the setup for the remaining time points is analogous.)

3.2. Simulation results

Using this setup, we generate 1000 datasets for both levels of the SNR. For each dataset, we obtain posterior samples from our MCMC algorithm, and we calculate the posterior means and the posterior 2.5-percentiles and 97.5-percentiles (based on the prior distributions given in Sections 2.2, 2.3, and the Appendix). In addition, we obtain FRS predictions and standard errors based on EM parameter estimation and, as a reference, we also obtain FRS predictions and prediction standard errors using the true parameters $\boldsymbol{\theta}$. We use the true parameter values to initialize the EM algorithm and to calibrate the priors for the FB procedure. We save both Bayesian posterior samples of all parameters

and EM parameter estimates for each dataset. Figure 3 shows the predictions and credible/prediction intervals for all three procedures for (the first four time points of) one simulated dataset, to illustrate inference on the process $\{Y_t(\cdot)\colon\ t = 1, \ldots, T\}$.

We summarize the results in Table 1. Generally, all summaries of the results are computed over all 1000 simulated datasets. However, the EM algorithm failed to converge for 17 of the datasets for SNR= 5 and for 39 of the datasets for SNR= 2 (see 'Success rate'), and so, the results from these datasets were excluded from the analysis. The first mean squared prediction error (MSPE) is taken over all 256 spatial locations at all T=16 time points. The summaries denoted 'on track' and 'off track' are only taken over the spatial locations for each time point that were considered on or off track, respectively, as described in the previous subsection. The interval score (IS) is defined as (Gneiting and Raftery, 2007, Sec. 6.2),

$$\mathrm{IS}_\alpha(l, u; y) = (u - l) + 2\{(l - y)_+ + (y - u)_+\}/\alpha,$$

where l and u are, respectively, the lower and upper endpoints of a $(1-\alpha)$ confidence interval (we use $\alpha = 0.05$), y is the true value, and $(x)_+ := xI(x > 0)$. This scoring rule combines the width of the confidence interval with a penalty for not containing the true value.

We can see from Table 1 that the posterior mean from our FB-FRS procedure is a considerably better predictor than that from the EM-FRS procedure, both on and off track. At least for SNR=5, the MSPE of the FB-FRS posterior mean is fairly close to the MSPE of the FRS procedure using the true parameter values (i.e., 'perfect' parameter estimation). The difference between FB-FRS and EM-FRS is even greater when we consider the prediction-uncertainty assessment. Possibly due to an overestimation of σ^2_δ, the confidence intervals for EM-FRS are too wide on track, but too narrow off track (see also Figure 3). The IS for FB-FRS is much closer to the IS for FRS using the true parameters than it is to the IS for EM-FRS. From Table 2, the FB posterior means of the *parameters* also result in (mostly) smaller mean squared estimation errors than the EM estimates.

Finally, we show the inference on the propagator matrix H in Figure 4 by taking elementwise medians of the estimates and posterior summaries based on each of the 1000 simulated datasets for SNR=5. Clearly, the lack of regularization in the EM estimates leads to a complete misestimation of the first row of the matrix H. Because the parameter estimates are simply plugged into the FRS equations, this misestimation is not accounted for in the EM-FRS prediction uncertainties. Although it seems from the FB posterior mean of H that the shrinkage induced by the prior might be too strong, we can see that the estimated posterior standard deviations actually reflect the magnitude

Table 1. For the simulation experiment, prediction results for FB-FRS, EM-FRS, and FRS using the true parameters

	SNR = 5				SNR = 2			
	True θ	EM	FB	EM/FB	True θ	EM	FB	EM/FB
Success rate	—	0.983	1.000	—	—	0.961	1.000	—
MSPE	0.115	0.198	0.146	1.361	0.140	0.239	0.195	1.229
MSPE - on track	0.034	0.045	0.036	1.239	0.044	0.060	0.049	1.228
MSPE - off track	0.196	0.352	0.255	1.379	0.237	0.418	0.340	1.229
IS - on track	0.862	1.145	0.913	1.254	0.975	1.400	1.092	1.282
IS - off track	2.011	3.920	2.360	1.661	2.256	5.011	2.769	1.810
CIW - on track	0.718	1.077	0.737	1.462	0.817	1.287	0.853	1.510
CIW - off track	1.704	1.550	1.956	0.793	1.856	1.616	2.281	0.709
CIC - on track (t=8, s=96)	0.908	0.990	0.949	—	0.958	0.990	0.958	—
CIC - off track (t=2,s=32)	0.949	0.745	0.949	—	0.927	0.708	0.938	—

FB, fully Bayesian; FRS, fixed rank smoothing; EM, expectation-maximization; SNR, signal-to-noise ratio; MSPE, (Empirical) mean squared prediction error; IS, interval score; CIW, credible/prediction interval width (nominal 95% intervals); CIC, credible/prediction interval coverage (target is 95%).

Table 2. Mean squared estimation errors for the scalar parameters (Bayes estimates are posterior means)

	SNR = 5			SNR = 2		
	EM	FB	EM/FB	EM	FB	EM/FB
Success rate	0.983	1.000	—	0.961	1.000	—
μ_t	0.237	0.065	3.665	0.202	0.109	1.859
σ^2_δ ($\times 100$)	0.504	0.007	74.482	1.505	0.025	60.719
$\eta_{\delta,1}$	0.047	0.051	0.919	0.082	0.066	1.233
$\eta_{\delta,2}$	0.099	0.057	1.724	0.104	0.079	1.318

SNR, signal-to-noise ratio; EM, expectation-maximization; FB, fully Bayesian.

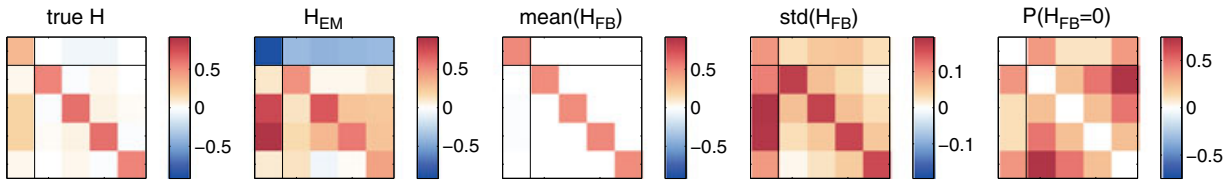

Figure 4. Propagator matrices. The left-hand panel shows the true H. All values in the other panels are elementwise medians over the 1000 simulations from the simulation study (SNR=5). Shown are the EM estimates, the posterior means, the posterior standard deviations, and the posterior probabilities of the elements being zero. The black lines partition H as in (9).

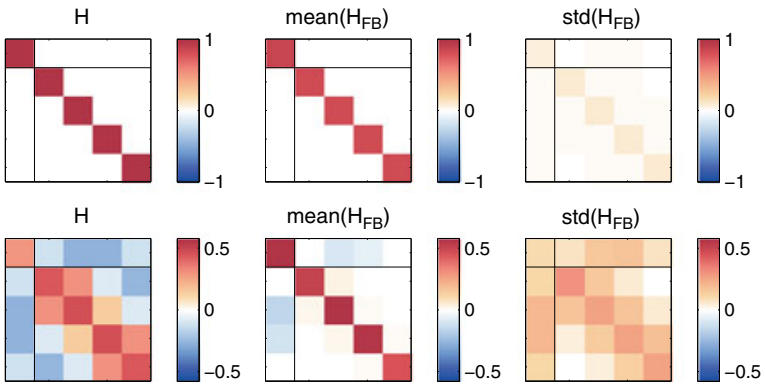

Figure 5. Top row: true $H = 0.8I_r$. Bottom row: true $H = P_K \text{diag}(.05, .08, .10, .94, .97) P_K$. Shown are the true H (first column), and the elementwise medians over the 1000 simulations from the two additional simulation studies (SNR=5): posterior means (middle column) and posterior standard deviations (right column). The black lines partition H as in (9).

of the true off-diagonal elements (left panel) quite well. The estimated probability of each element being equal to zero is also in agreement with how close to zero the true values are.

To see how well a sparse H and a full H can be recovered by our model, we carried out two more simulations. The setup was kept the same as before (with SNR=5), except that now, we specified the true H to be $H = 0.8I_r$ and $H = P_K \text{diag}(.05, .08, .10, .94, .97) P_K$, respectively, where P_K is the eigenvector matrix of the true K (which was again calibrated against an exponential covariance model). The results, shown in Figure 5, indicate that our model adapts well to very sparse or full propagator matrices.

4. ANALYSIS OF GLOBAL CO_2 DATA

This section contains an application of our proposed fully Bayesian STRE methodology to a very large real-world dataset of global CO_2 measurements. We obtain the posterior distribution of all parameters and the spatio-temporal process of interest, and we compare results with those from empirical-Bayesian STRE methodology based on the EM algorithm.

4.1. Spatio-temporal data: mid-tropospheric CO_2 measurements from the Atmospheric InfraRed Sounder

The spatio-temporal dataset under consideration consists of 16 days of measurements of global mid-tropospheric CO_2, which were recorded by the Atmospheric InfraRed Sounder (AIRS) on board NASA's Aqua satellite (Chahine *et al.*, 2006). The dataset is available from http://airs.jpl.nasa.gov/AIRS_CO2_Data/, and it is the same as the one analyzed in Katzfuss and Cressie (2011). Only CO_2 measurements between $-60°$ and $90°$ latitude are available, because corresponding data at latitudes south of $-60°$ have not been released by AIRS yet. The unit of measurement is parts per million (ppm). The measurements are taken at roughly 1:30 pm local time, and we considered here those for 1 May through 16 May 2003, which from now on are referred to as days 1 through 16, respectively.

We have gridded the data onto a very fine grid, as in Katzfuss and Cressie (2011), which allows us to compare the results described in that paper. However, we would like to emphasize that neither methodology requires gridded data. The hexagonal grid (ISEA Aperture 3 Hexagons at resolution 8) of size $m_t \equiv 61,236$ was obtained using DGGRID software (Sahr, 2003). On each day, roughly 12 000, or 20%, of the grid cells contained data; orbit geometry, cloud cover, and retrieval convergence criteria caused the remaining grid cells to contain no data. If a particular grid cell contained more than one of the original measurements on a particular day, the data value at that grid cell was taken to be the average of those measurements and the measurement-error covariance matrix was modified correspondingly, so that $v_{\epsilon,t}(S_{i,t}) = 1/N_t(S_{i,t})$, where $N_t(S_{i,t})$ is the number of measurements contained in grid cell $S_{i,t}$ at time t. For illustration, the gridded data of day 1 are shown in the top panel of Figure 6.

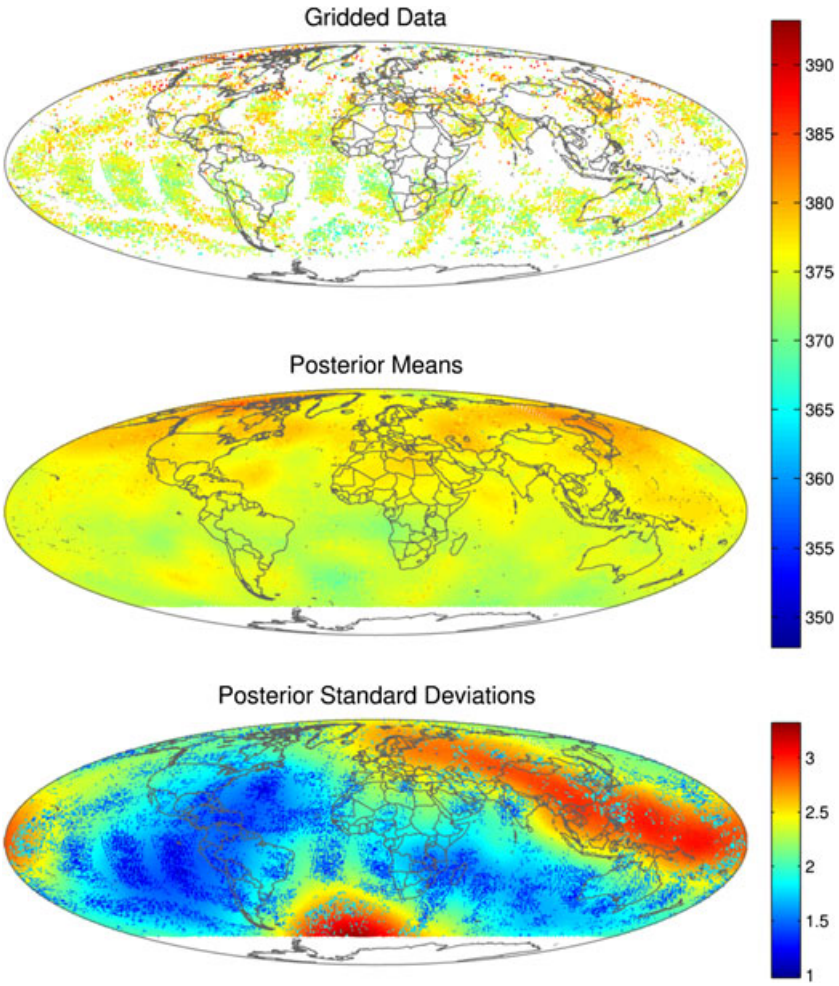

Figure 6. Gridded Atmospheric InfraRed Sounder measurements of mid-tropospheric CO_2 on 1 May 2003, $\mathbf{z}_1$ (top); posterior means (middle) and posterior standard deviations (bottom) of $\{Y_1(\mathbf{s})\colon \mathbf{s} \in D_s\}$. Units are parts per million.

The measurement-error variances, $\{\sigma^2_{\epsilon,t}\}$, are assumed known in our model (Section 2.1); in reality, these variances were estimated prior to the actual analysis using the variogram-extrapolation technique described in Kang *et al.* (2009) and adapted by Katzfuss and Cressie (2011). There, we obtained a pooled estimate, $\hat{\sigma}^2_\epsilon = 5.6062\,\text{ppm}^2$, for all days $t = 1, \ldots, 16$.

The large-scale spatial trend was assumed to be determined by an intercept and a latitudinal gradient; that is, we set $\mathbf{x}_t(\cdot) = [1 \;\; \text{lat}(\cdot)]'$, independent of t.

Our model in Section 2.1 is described for measurements made at a point level, but here in Section 4, we take into account that our data has areal support. This change-of-support problem can be handled quite easily in the STRE model, by replacing the quantities in (5)–(7) by averages over the respective grid cells; more details can be found in Katzfuss and Cressie (2011).

For the basis functions, we used $r = 380$ bisquare functions defined by (15), from three resolutions. The set of functions was identical to that used in Katzfuss and Cressie (2011); in Section 5.2 of that article, the reader can also find a brief discussion of how basis functions can be chosen.

We need to ensure that the distance measure d in (11) is normalized so that $\max\{d_{ij}\} = 1$. Here, for our basis functions located on the globe, we normalized the spherical distances between each pair of basis-function centres by dividing them by $\pi \cdot$earth's radius $\approx \pi \cdot 6371$ km, which is the maximum great-arc distance that two points on the globe can be apart.

The as-of-yet unspecified values of the hyperparameters in the priors on $\{\boldsymbol{\beta}_t\}$, σ^2_δ, $\boldsymbol{\theta}_H$, K_0, and U were calibrated using the EM estimates of the respective parameters, as described in the Appendix.

4.2. Posterior results

We ran an MCMC for 20 000 iterations using MATLAB, where one iteration of the MCMC took about 30 seconds to compute on an eight-core machine (Intel Xeon X5560, with 94.5 GB RAM). Thus, the 16 days worth of data can be processed in less than 1 week. Trace plots showed that convergence to stationarity had been reached rather quickly, so that we considered the first 2000 iterations as burn-in. We computed the posterior distribution of $\{Y_t(\cdot)\}$ at all $t = 1, \ldots, 16$ time points and all $m = 61\,236$ hexagons. The posterior means and standard

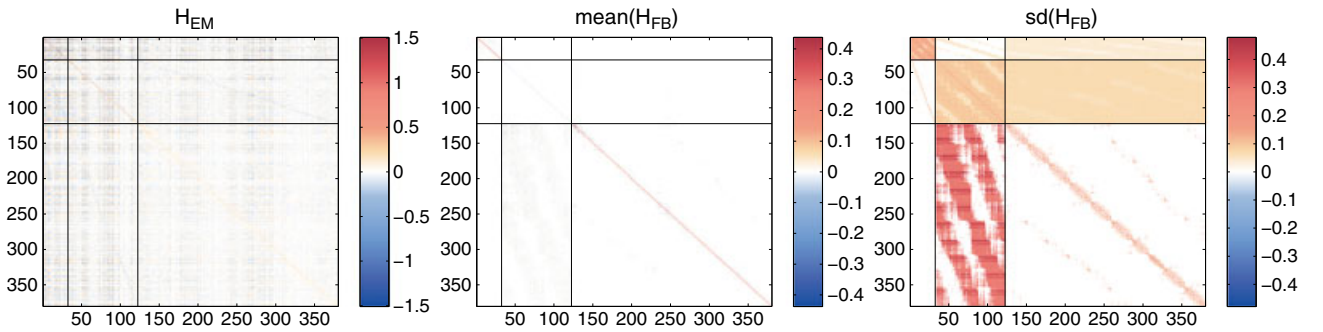

Figure 7. Expectation-maximization estimate, and mean and standard deviation of the posterior distribution of H from the Atmospheric InfraRed Sounder data. The black lines divide H as in (9).

deviations for $t = 1$ are shown in Figure 6. The posterior standard deviations are lowest at observed grid cells; they are highest around Siberia and Southeast Asia, and around $-60°$ latitude, which is likely caused by the lack of data south of that latitude.

To evaluate the prediction performance of our FB-FRS procedure and that of the EM-FRS procedure, 500 grid cells containing observations at time point $t = 10$ were randomly selected into a test set, $\mathcal{S}_{\text{test}} := \{S_1, \ldots, S_{500}\}$. These data were unavailable for model fitting and were used to measure out-of-sample-prediction accuracy. For EM-FRS, $\boldsymbol{\eta}_\delta$ was again considered a fixed parameter. As the true process $\{Y_t(\cdot)\}$ is not known in this real-data example, we used the measurements $\{Z_t(S_1), \ldots, Z_t(S_{500})\}$ directly as a reference, and we evaluated our predictions using an average-squared-distance criterion, ASD $:= \sum_{i=1}^{500}(\hat{Y}_{10}(S_i) - Z_{10}(S_i))^2/500$. For EM-FRS, we obtained $\text{ASD}_{\text{EM}} = 9.1992$, and for FB-FRS, we obtained $\text{ASD}_{\text{FB}} = 9.1045$ when using posterior means as predictors for $\{Y_{10}(S_1), \ldots, Y_{10}(S_{500})\}$. Thus, the FB-FRS approach had a small advantage. As a baseline predictor, we simply calculated, for a $1° \times 1°$ grid, the binned means of the data from all 16 days, resulting in $\text{ASD}_{\text{BM}} = 12.0924$.

Figure 7 shows the EM estimate and summaries of the posterior distribution of the propagator matrix H. The EM estimate of H exhibits very little structure, other than a strong (positive) diagonal and two faint lines of negative elements, where each element corresponds to two basis functions of two different resolutions that are close in space. The FB posterior mean of H is much more structured, and we can see the sparsity induced by the prior distribution by examining the elements with zero standard deviation in the panel on the right. Notice that the posterior standard deviations do not show symmetric behavior in H. The eigenvalues of the posterior mean are all smaller than one, indicating a non-explosiveness of the process.

For model validation, we compared empirical root-semivariograms with posterior medians of theoretical root-semivariograms for several spatio-temporal directions and for several reference locations. The two sets of root-semivariograms were reasonably close, but did not match exactly. The discrepancy is likely due to shrinkage through the priors and because of the fact that K_0, H, and U were held constant over time and hence had to account for the spatial and temporal covariance structure on all days. We also divided the residuals of the test data by the posterior standard errors of $Z_{10}(\cdot)$ at the corresponding locations. The posterior standard error of $Z_{10}(\mathbf{s})$ is given by the square-root of the sum of the posterior variance of $Y_{10}(\mathbf{s})$ and the variance of the measurement error ($\sigma^2_\epsilon = 5.6062$). The ratios roughly follow a normal distribution, but the tails of the empirical distribution are slightly lighter than the standard normal tails. Thus, the uncertainty estimation of our model seems to be somewhat conservative for the AIRS data.

5. DISCUSSION AND CONCLUSIONS

In this article, we have presented a fully Bayesian, hierarchical approach to spatio-temporal smoothing of very large datasets. By projecting the spatio-temporal process of interest onto a low-dimensional space spanned by basis functions, the STRE model makes Bayesian model fitting feasible, even when the number of observations is large. The temporal evolution of the process on the lower-dimensional space is governed by two covariance matrices and a propagator matrix. Apart from positive definiteness of the covariance matrices, we do not require any restrictions (e.g., diagonal matrices) on these potentially large matrix parameters, but instead we specify prior distributions to achieve regularization and identifiability. In the Appendix, we give detailed instructions for posterior inference, and we provide computational speedups to achieve feasible computation times.

Another key ingredient of our approach is that we do not ignore the error introduced by the dimension reduction. Instead, we attempt to separate the discrepancy between observations and the reduced-dimensional process into two types of error, one caused by the measurement process, and one caused by the dimension reduction. Estimating the spatial heterogeneity of the variance of the fine-scale variation is important, as can be seen from the example using AIRS mid-tropospheric CO_2 in Section 4. From Figure 6, there is an indication that there are different degrees of smoothness in different parts of the globe; see, in particular, the heterogeneity of variances in the bottom panel. As the basis-function space is limited in the amount of roughness it can exhibit, the variation in the excess roughness should be reflected in spatially heterogeneous fine-scale variation over the globe.

In the two comparisons of the fully Bayesian FB-FRS procedure to the empirical-Bayes EM-FRS procedure presented in Sections 3 and 4, FB-FRS results in better predictions than EM-FRS, particularly in the simulation experiment where the true process was available for comparison. The true worth of FB-FRS is apparent from its more accurate assessment of prediction uncertainty in the simulation. Here, FB-FRS

outperforms EM-FRS by up to 80% (in terms of the interval score; see Table 1). In Section 4, we expect that the relative prediction accuracy of FB-FRS would be even larger if the test set consisted of a contiguous region of the globe, so that there would be no data nearby.

Our current prior on H (Section 2.3) allows for sparsity and shrinkage on each element h_{ij} as a function of the distance of the two basis functions i and j, but it assumes prior independence conditional on the parameters in $\boldsymbol{\theta}_H$. Although inference as described in the Appendix should work even when the prior on H specifies conditional dependence (within rows of H), we encountered difficulties with our MCMC in that case.

It would be of interest to generalize further the joint-distributional assumptions on the fine-scale variation. In this article, we have assumed spatial independence and allowed the variance to vary spatially, but we could also allow for short-range spatial dependence. One would need to balance the requirement of rapid inversion of the data's covariance matrix (e.g., by using a strong taper) with a realistic covariance structure. Assuming spatial independence, the maps of predictions and standard errors look 'spiky' in locations where data were observed; short-range dependence would mitigate against this. Katzfuss (2011, Chap. 4) explores some of these suggestions.

Future work could also include further optimization of the computer code. The current code already exploits some computational speed-ups (as described in the Appendix), and parallelization is employed where possible to allow for efficient use of a multi-core computer. However, faster computation might be achieved by implementing the MCMC in a compiled language such as C++, instead of MATLAB.

Finally, an elegant solution to spatio-temporal *filtering* would be of interest; we have presented spatio-temporal smoothing here. If parameters are fixed across time (as they are in this article), their posterior distributions need to be updated when a new set of data becomes available. This can quickly become infeasible as one goes forward in time. Letting parameters such as H_t and U_t vary with t might provide a solution.

Acknowledgements

This research was supported by NASA under grant NNX08AJ92G issued through the ROSES Carbon Cycle Science Program and grant NNH08ZDA001N issued through the Advanced Information Systems Technology ROSES 2008 Solicitation. Katzfuss' research was also partially supported by the Mathematics Center Heidelberg. We would like to thank the following: Anna Michalak and Amy Braverman for advice on spatio-temporal variability expected in global CO_2 data; Kevin Sahr for providing the DGGRID software and for advice on how to shift the basis-function centres; the AIRS Project CO_2 team, particularly Dr Moustafa T. Chahine, Dr Edward T. Olsen, and Mr Luke L. Chen for their helpful comments on our analysis of the AIRS data; and the Associate Editor and two anonymous referees for their constructive feedback.

REFERENCES

Antoulas A. 2005. *Approximation of Large-Scale Dynamical Systems*. SIAM: Philadelphia, PA.
Aubry N, Lian W, Titi E. 1993. Preserving symmetries in the proper orthogonal decomposition. *SIAM Journal on Scientific Computing* **14**: 483–505.
Banerjee S, Gelfand AE, Finley AO, Sang H. 2008. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society, Series B* **70**(4): 825–848.
Bishop CM. 2006. *Pattern Recognition and Machine Learning*. Springer: New York, NY.
Calder CA, Holloman C, Higdon D. 2002. Exploring space-time structure in ozone concentration using a dynamic process convolution model. In *Case Studies in Bayesian Statistics*, Vol. VI, C Gatsonis, R Kass, A Carriquiry, A Gelman, D Higdon, D Pauler, I Verdinelli (eds). Springer: New York, NY; 165–176.
Cangelosi A, Hooten MB. 2009. Models for bounded systems with continuous dynamics. *Biometrics* **65**(3): 850–856.
Carter C, Kohn R. 1994. On Gibbs sampling for state space models. *Biometrika* **81**(3): 541–553.
Chahine M, Pagano TS, Aumann H, Atlas R, Barnet C, Blaisdell J, Chen L, Divakarla M, Fetzer E, Goldberg M, Gautier C, Granger S, Hannon S, Irion F, Kakar R, Kalnay E, Lambrigtsen B, Lee SY, Marshall JL, Mcmillian WW, Mcmillin L, Olsen E, Revercomb H, Rosenkranz P, Smith W, Staelin D, Strow L, Susskind J, Tobin D, Wolf W, Zhou L. 2006. AIRS - Improving weather forecasting and providing new data on greenhouse gases. *Bulletin of the American Meteorological Society* **87**(7): 911–926.
Cressie N. 1993. *Statistics for Spatial Data*, (revised edn.) John Wiley & Sons: New York, NY.
Cressie N, Johannesson G. 2006. Spatial prediction of massive datasets. In *Mastering the Data Explosion in the Earth and Environmental Sciences: Proceedings of the Australian Academy of Science Elizabeth and Frederick White Conference*. Australian Academy of Science: Canberra, Australia.
Cressie N, Johannesson G. 2008. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society, Series B* **70**(1): 209–226.
Cressie N, Kang EL. 2010. *High-resolution Digital Soil Mapping: Kriging for Very Large Datasets*. Springer: Dordrecht, NL. pp. 49–63.
Cressie N, Shi T, Kang EL. 2010. Fixed rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics* **19**(3): 724–745.
Cressie N, Wikle CK. 2011. *Statistics for Spatio-Temporal Data*. Wiley: Hoboken, NJ.
Dewar M, Scerri K, Kadirkamanathan V. 2009. Data-driven spatio-temporal modeling using the integro-difference equation. *IEEE Transactions on Signal Processing* **57**(1): 83–91.
Fassò A, Cameletti M. 2009. The EM algorithm in a distributed computing environment for modelling environmental space-time data. *Environmental Modelling & Software* **24**(9): 1027–1035.
Frühwirth-Schnatter S. 1994. Applied state space modelling of non-Gaussian time series using integration-based Kalman filtering. *Statistics and Computing* **4**(4): 259–269.
Gelpke V, Künsch HR. 2001. Estimation of motion from sequences of images: daily variability of total ozone mapping spectrometer ozone data. *Journal of Geophysical Research* **106**(D11): 11825–11834.
Geman S, Geman D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**(6): 721–741.
George EI, Sun D, Ni S. 2008. Bayesian stochastic search for VAR model restrictions. *Journal of Econometrics* **142**(1): 553–580.
Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**(477): 359–378.
Haario H, Saksman E, Tamminen J. 2001. An adaptive metropolis algorithm. *Bernoulli* **7**(2): 223–242.
Hamilton J. 1994. *Time Series Analysis*. Princeton University Press: Princeton, NJ.
Hastings W. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**(1): 97–109.

Higdon D. 1998. A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics* **5**(2): 173–190.
Higdon D. 2002. Space and space-time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues*, C Anderson, V Barnett, P Chatwin, A El-Shaarawi (eds). Springer: London; 37–56.
Johannesson G, Cressie N, Huang HC. 2007. Dynamic multi-resolution spatial models. *Environmental and Ecological Statistics* **14**(1): 5–25.
Jun M, Stein ML. 2008. Nonstationary covariance models for global data. *Annals of Applied Statistics* **2**(4): 1271–1289.
Kalman R. 1960. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* **82**(1): 35–45.
Kang EL, Cressie N. 2011. Bayesian inference for the spatial random effects model. *Journal of the American Statistical Association* **106**(495): 972–983.
Kang EL, Cressie N, Shi T. 2010. Using temporal variability to improve spatial mapping with application to satellite data. *Canadian Journal of Statistics* **38**(2): 271–289.
Kang EL, Liu D, Cressie N. 2009. Statistical analysis of small-area data based on independence, spatial, non-hierarchical, and hierarchical models. *Computational Statistics & Data Analysis* **53**(8): 3016–3032.
Kaplan A, Cane M, Kushnir Y, Clement A, Blumenthal M, Rajagopalan B. 1998. Analyses of global sea surface temperature 1856–1991. *Journal of Geophysical Research* **103**(18): 18567–18589.
Katzfuss M. 2011. Hierarchical spatial and spatio-temporal modeling of massive datasets, with application to global mapping of CO_2. *PhD Dissertation*, The Ohio State University.
Katzfuss M, Cressie N. 2011. Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets. *Journal of Time Series Analysis* **32**(4): 430–446.
Kot M, Lewis M, van Den Driessche P. 1996. Dispersal data and the spread of invading organisms. *Ecology* **77**: 2027–2042.
Litterman R. 1986. Forecasting with Bayesian vector autoregressions: ffve years of experience. *Journal of Business & Economic Statistics* **4**(1): 25–38.
Lopes HF, Salazar E, Gamerman D. 2008. Spatial dynamic factor analysis. *Bayesian Analysis* **3**(4): 759–792.
Mardia K, Goodall C, Redfern E, Alonso F. 1998. The kriged Kalman filter. *Test* **7**(2): 217–282.
Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E. 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21**(6): 1087–1092.
Nychka DW, Wikle CK, Royle JA. 2002. Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling* **2**(4): 315–331.
Sahr K. 2003. DGGRID Software. http://webpages.sou.edu/~sahrk/dgg/dggrid/dggrid.html. Version 4.3b.
Shumway R, Stoffer D. 2006. *Time Series Analysis and Its Applications: With R Examples*, (2nd edn.) Springer: New York, NY.
Smith T, Reynolds R, Livezey R, Stokes D. 1996. Reconstruction of historical sea surface temperatures using empirical orthogonal functions. *Journal of Climate* **9**(6): 1403–1420.
Stroud J, Müller P, Sansó B. 2001. Dynamic models for spatiotemporal data. *Journal of the Royal Statistical Society, Series B* **63**(4): 673–689.
Stroud J, Stein ML, Lesht B, Schwab D. 2010. An ensemble Kalman filter and smoother for satellite data assimilation. *Journal of the American Statistical Association* **105**(491): 978–990.
Wickerhauser M. 1994. *Adapted Wavelet Analysis from Theory to Software*. A K Peters: Wellesley, MA.
Wikle CK. 2003. Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology* **84**(6): 1382–1394.
Wikle CK. 2010. Low-rank representations for spatial processes. In *Handbook of Spatial Statistics*, AE Gelfand, M Fuentes, P Guttorp, P Diggle (eds). Chapman and Hall/CRC: Boca Raton, FL; 107–118.
Wikle CK, Berliner L, Cressie N. 1998. Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics* **5**(2): 117–154.
Wikle CK, Cressie N. 1999. A dimension-reduced approach to space-time Kalman filtering. *Biometrika* **86**(4): 815–829.
Wikle CK, Hooten MB. 2010. A general science-based framework for dynamical spatio-temporal models. *Test* **19**(3): 417–451.
Wikle K, Milliff R, Nychka DW, Berliner L. 2001. Spatiotemporal hierarchical Bayesian modeling: tropical ocean surface winds. *Journal of the American Statistical Association* **96**(454): 382–397.
Xu B, Wikle CK, Fox N. 2005. A kernel-based spatio-temporal dynamical model for nowcasting radar precipitation. *Journal of the American Statistical Association* **100**(472): 1133–1144.
Xu K, Wikle CK. 2007. Estimation of parameterized spatio-temporal dynamic models. *Journal of Statistical Planning and Inference* **137**(2): 567–588.

APPENDIX

This is included as supplementary material to the published article.

