Tutorial on Fixed Rank Kriging (FRK) of CO_2 data

M. Katzfuss, *The Ohio State University*
N. Cressie, *The Ohio State University*

Technical Report No. 858

July, 2011

Department of Statistics
The Ohio State University
1958 Neil Avenue
Columbus, OH 43210-1247

Tutorial on Fixed Rank Kriging (FRK) of CO_2 Data

Matthias Katzfuss* Noel Cressie*

June 24, 2011

Abstract

In this document, we describe Fixed Rank Kriging (FRK), an approach to the analysis of very large spatial datasets. Such datasets now arise in many fields; our focus is on satellite measurements of CO_2. FRK predictors and standard errors can be computed rapidly, even for datasets with a million or more observations. FRK relies on a so-called spatial random effects (SRE) model, which assumes that the process of interest can be expressed as a linear combination of spatial basis functions, plus a fine-scale-variation component. Here, we describe in detail all steps involved in the analysis of a spatial dataset using FRK, we illustrate the steps using a synthetic dataset, and we provide Matlab code on an accompanying website.

1 Introduction

1.1 Spatial Analysis of Large Datasets

This document gives a detailed recipe of how a spatial analysis of a very large dataset can be conducted using a flexible class of geostatistical models called the spatial random effects (SRE) models. All steps of the analysis are laid out, from the exploratory analysis, to parameter estimation, to obtaining predictions and their accompanying standard errors.

The goal of most spatial analyses is to predict the distribution of a process of interest based on a number of observations containing measurement error. Traditional spatial statistics requires inversion of the covariance matrix of the data, which requires on the order of n^3 computations when there are n observations in the dataset. This is computationally feasible for n up to a few thousand, but it is not feasible when n is in the tens of thousands and above. In such large-data situations, one looks to dimension reduction. We do this here by making use of an SRE model, in which the process of interest is modeled as a linear combination of basis functions plus a fine-scale-variation term. Spatial prediction using this geostatistical model, termed Fixed Rank Kriging (FRK) by Cressie and Johannesson (2008), is then feasible even for very large n (up to a million or more), as long as the number of basis functions is much smaller (no more than a few thousand) than the number of observations.

The SRE model can be extended to a spatio-temporal version by specifying the dynamical evolution of the vector of basis-function coefficients over time. The resulting spatio-temporal

*Department of Statistics, The Ohio State University

random effects (STRE) model can then be used to exploit both spatial and temporal dependence of the process when predicting its distribution at locations in space and time. We shall not give any more details about the spatio-temporal case here and instead refer the interested reader to published articles, such as Cressie, Shi, and Kang (2010), Kang, Cressie, and Shi (2010), and Katzfuss and Cressie (2011a,b), that treat the spatio-temporal case (empirical-Bayesian and fully Bayesian) in detail.

1.2 Synthetic Dataset

Throughout this document, we shall demonstrate the generic recipe for FRK using a global dataset of CO_2 measurements. In principal, CO_2 varies with latitude, land/sea, and geopotential (pressure) height. Of particular scientific interest is the boundary layer closest to the Earth's surface, because this is usually where CO_2 is generated and absorbed (i.e., the sources and sinks). However, satellite instruments measure column-integrated versions of CO_2, denoted XCO2; that is, for a specific latitude-longitude footprint, they average over the CO_2 observed at all pressure heights, where the averaging weights are determined by a pressure weighting function. Carbon-cycle transport models are then used to look for sources and sinks. In what is to follow, we present FRK for XCO2 data.

Specifically, we use a global "dataset" of column-integrated CO_2 (XCO2) for 1pm local time on September 20, 2003. This dataset does not consist of actual measurements, but it was generated from the PCM-CSM Transition Model, or PCTM (Kawa, 2004; Strand, 2004; Chatterjee and Kawa, 2009). The PCTM provides values of column-integrated CO_2 on a grid of 1.25° longitude by 1° latitude, which results in a total of $288 \times 181 = 52,128$ grid cells on the globe. We consider these values to represent the true XCO2 field, which we denote by $Y(\cdot)$. Throughout most of this document, we assume that each observation is taken at the center point of the grid cell. In reality, each observation describes an integral over the respective grid cell, and we shall describe how to generalize our model to reflect this fact in Section 6. To obtain a dataset that resembles measurements that could have been made by the Orbiting Carbon Observatory (OCO; see Crisp and Johnson, 2005), we combine this PCTM dataset with track information and cloud/aerosol-density from the CALIPSO satellite (Winker et al., 2003). We discard all grid cells in the PCTM data that would *not* have been on the track of OCO during the 16-day time period of September 13-28, 2003 (orbit-geometry mask), and those that would *not* have been visible during that same time period, as determined by measurements of cloud density from the CALIPSO satellite (cloud mask). The criteria for a cell to be sampled are (i) the cell is on the OCO track and (ii) the cell is visible, meaning that both cloud and aerosol optical depth (from CALIPSO) are below 0.3. CALIPSO only takes measurements between -82° and 82° latitude, and so we simply assume that all grid cells closer to the poles are unobserved. Because of the orbit-geometry mask and cloud mask, the dataset we analyze here has only roughly half the number of observations (26,633 out of the 52,128 possible). Finally, to create the dataset $\mathbf{Z}$ that we analyze in this document, we added Gaussian white noise with zero mean and variance 0.25 as measurement error, to $Y(\cdot)$. The unit of measurement is parts per million (ppm). The resulting synthetic data $\mathbf{Z}$ are shown in Figure 1, and we refer to them as OCO-like (OCOL) data. (As a footnote, OCO was launched in February 2009 but failed to reach orbit. At the time of writing this document, OCO-2 is under development for a launch date in the 2013-2014 time window.)

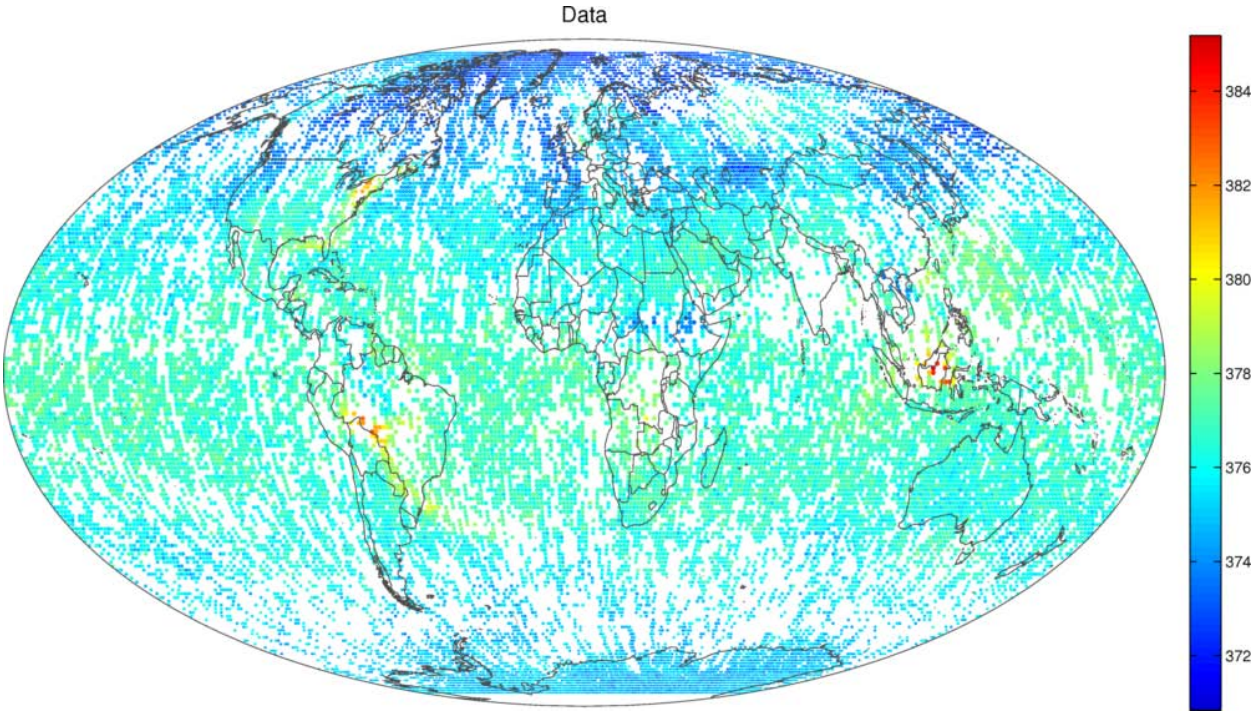

Figure 1: OCO-like data

1.3 Notation

We are interested in making inference on a hidden spatial process $\{Y(\mathbf{s}) : \mathbf{s} \in D_s\}$ on a spatial domain, or region of interest, D_s (e.g., the globe, or the United States). The n data,

$$Z(\mathbf{s}_i) \equiv Y(\mathbf{s}_i) + \epsilon(\mathbf{s}_i),\ i = 1, \ldots, n,$$

are observed with additive measurement error. The measurement-error process, $\epsilon(\cdot)$, is assumed to be statistically independent of $Y(\cdot)$ and distributed as,

$$\epsilon(\cdot) \sim N(0, \sigma^2_\epsilon v_\epsilon(\cdot)), \tag{1}$$

where $v_\epsilon(\cdot)$ is a function that is typically assumed to be known (see the end of this subsection). Then

$$\mathbf{Z} \equiv (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))'.$$

This defines a *data model* that can be equivalently written as,

$$\mathbf{Z}|\mathbf{Y} \sim N_n(\mathbf{Y}, \sigma^2_\epsilon V_\epsilon),$$

where $V_\epsilon \equiv diag(v_\epsilon(\mathbf{s}_1), \ldots, v_\epsilon(\mathbf{s}_n))$. The process $Y(\cdot)$ is decomposed into two components,

$$Y(\cdot) = \mu(\cdot) + \nu(\cdot),$$

where the first component $\mu(\cdot)$ is a deterministic large-scale trend $\mu(\cdot)$, and the second component $\nu(\cdot)$ is a random spatial-variation component $\nu(\cdot)$. We assume that the deterministic trend is a linear function of spatial covariates,

$$\mu(\cdot) = \mathbf{x}(\cdot)'\boldsymbol{\beta}, \tag{2}$$

where $\mathbf{x}(\cdot)$ is a p-dimensional vector of known covariates. Section 2.1 describes how suitable covariates can be identified for a particular dataset.

The random spatial-variation term $\nu(\cdot)$ is further assumed to follow the spatial-random-effects (SRE) model,

$$\nu(\cdot) = \mathbf{S}(\cdot)'\boldsymbol{\eta} + \xi(\cdot),$$

where $\mathbf{S}(\cdot) \equiv (S_1(\cdot), \ldots, S_r(\cdot))'$ is an r-dimensional ($r << n$) vector of spatial basis functions, and $\boldsymbol{\eta}$ is a random-effects vector. We assume that $\boldsymbol{\eta}$ is distributed as,

$$\boldsymbol{\eta} \sim N_r(\mathbf{0}, K),$$

where K is an unknown $r \times r$ symmetric positive-definite matrix. The fine-scale-variation process $\xi(\cdot)$ accounts for the error induced by the dimension reduction and is distributed as,

$$\xi(\mathbf{s}) \sim N(0, \sigma_\xi^2 v_\xi(\mathbf{s})), \tag{3}$$

independently for all $\mathbf{s} \in D_s$, and independently of $\boldsymbol{\eta}$, where $v_\xi(\cdot)$ is a known function.

By stacking all scalar quantities into vectors and all row vectors into matrices, we can write the data vector as,

$$\mathbf{Z} \equiv [Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)]' = \mathbf{Y} + \boldsymbol{\epsilon},$$

where

$$\mathbf{Y} = X\boldsymbol{\beta} + S\boldsymbol{\eta} + \boldsymbol{\delta}; \tag{4}$$

hence, for example, the i-th row of S is $\mathbf{S}(\mathbf{s}_i)'$. The covariance matrix of the data vector has the form,

$$\Sigma \equiv var(\mathbf{Z}) = var(S\boldsymbol{\eta}) + var(\boldsymbol{\delta} + \boldsymbol{\epsilon}) = SKS' + D, \tag{5}$$

where $D \equiv \sigma_\xi^2 V_\xi + \sigma_\epsilon^2 V_\epsilon$, $V_\xi \equiv diag(v_\xi(\mathbf{s}_1), \ldots, v_\xi(\mathbf{s}_n))$, and recall that $V_\epsilon \equiv diag(v_\epsilon(\mathbf{s}_1), \ldots, v_\epsilon(\mathbf{s}_n))$.

We shall assume for now that the functions $v_\xi(\cdot)$ and $v_\epsilon(\cdot)$ are known (see Section 6.1 for one possible way of relaxing this assumption). For $v_\epsilon(\cdot)$, if the observations $\mathbf{Z}$ are binned averages of some original measurements, one could set $v_\epsilon(\mathbf{s}_i) = 1/N_i$, where N_i is the number of original measurements that went into the average $Z(\mathbf{s}_i)$; $i = 1, \ldots, n$. If there is no reason to believe that the measurement-error variances should be different in different parts of D_s, one can simply assume $V_\epsilon = I_n$, the identity matrix. The same can be said of $v_\xi(\cdot)$.

We can also write our model in a hierarchical form: The data model is given by,

$$\mathbf{Z}|\mathbf{Y}^P, \sigma_\epsilon^2 \sim N_n(\mathbf{Y}, \sigma_\epsilon^2 V_\epsilon),$$

where $\mathbf{Y}^P$ is the vector of the hidden process for all m possible locations, observed or not. The *process model* is given by,

$$\mathbf{Y}^P|K, \sigma_\xi^2 \sim N_m(X^P\boldsymbol{\beta}, S^P K S^{P\prime} + \sigma_\xi^2 V_\xi^P),$$

where X^P, S^P, V^P are obtained by evaluating the corresponding quantities at all m possible locations. Alternatively, the process model can be written conditionally, in two stages:

$$\begin{aligned} \mathbf{Y}^P|\boldsymbol{\eta}, \sigma_\xi^2 &\sim N_m(X^P\boldsymbol{\beta} + S^P\boldsymbol{\eta}, \sigma_\xi^2 V_\xi^P) \\ \boldsymbol{\eta}|K &\sim N_r(\mathbf{0}, K). \end{aligned}$$

In Section 3, we shall describe how to estimate the unknown parameters σ^2_ϵ, σ^2_δ, and K from data. We then take an empirical-Bayesian approach and substitute the estimators into the hierarchical model given by the data model and the process model. Inference on $Y(\cdot)$ is obtained from Bayes Theorem; see Section 4. Section 5 gives diagnostics to check the fit of the hierarchical SRE model, and model extensions are discussed in Section 6.

2 Preliminary Steps

2.1 Detrending

The first step in the analysis of a given spatial dataset is the identification of important large-scale spatial variation (deterministic trend) and important covariates; that is, we would like to identify the components of $\mathbf{x}(\cdot)$ in (2). If no variables (other than the variable of interest) are available, we can at least examine whether $Z(\cdot)$ appears to be varying systematically with one or more of the components of the spatial-location vector $\mathbf{s} \in D_s$. This is usually done by plotting the data versus spatial directions (e.g., Easting and Northing, here longitude and latitude) and examining the plots. Other exploratory data analysis (EDA) techniques are described in Cressie (1993, Sect. 2.2). EDA tools can be used to identify trend, check model assumptions, and in general to "get a feel for the data."

Once we have identified the components of the matrix X in (4), we estimate $\boldsymbol{\beta}$ using ordinary least squares. That is, we find the $\boldsymbol{\beta}$ that minimizes the sum of the squared distance between the data points $\{\mathbf{Z}(\mathbf{s}_i) : i = 1, \ldots, n\}$ and the respective trend terms, $\{\mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} : i = 1, \ldots, n\}$. The solution to this minimization problem is given by $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{Z}$. We then detrend the data,

$$\tilde{\mathbf{Z}} \equiv \mathbf{Z} - X\hat{\boldsymbol{\beta}}; \tag{6}$$

we shall work with these detrended data (residuals) for the rest of this document.

In the XCO2-data example described in Section 1.2, we have $\mathbf{s} = (\text{longitude}, \text{latitude})'$; hence we plot the data versus longitude and latitude. From the plots and also from expert knowledge (Michalak, 2010), we know that the global distribution of XCO2 varies broadly according to a latitudinal gradient for most of the year. Of course, we also want to include an intercept, and so we set $\mathbf{x}(\mathbf{s}) = (1, s_2)'$. For the dataset we created in Section 1.2, we obtained the estimate $\hat{\boldsymbol{\beta}} = (375.6595, -0.0001)'$, which resulted in the detrended data $\tilde{\mathbf{Z}}$, as described in (6).

2.2 Examine Normality of the Data

As a second step, we see whether the normality assumptions in Section 1.3 are (at least approximately) met. However, even without normality of the data, predictors based on our model are still the best linear unbiased predictors (BLUPs). To check normality, we can look at normal-quantile-quantile plots and plots of kernel-density estimates or histograms of the (detrended) data. With very large datasets, it is almost always possible to find deviations from normality. If the data do not appear to follow a normal distribution, even approximately, it is possible that they will after a transformation, such as the log transformation or a transformation from the Box-Cox family of transformations (Box and Cox, 1964); Shi and Cressie (2007) used the log transformation for aerosol data from the MISR instrument on NASA's Terra satellite.

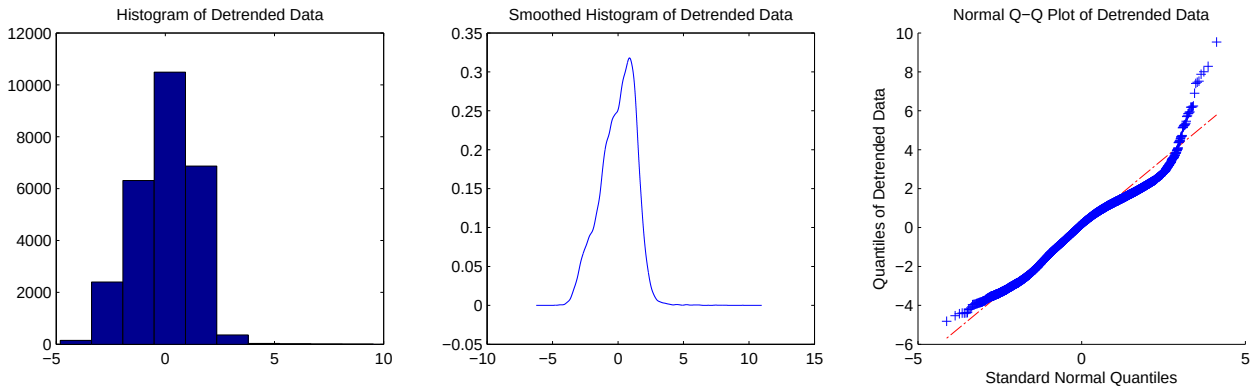

Figure 2: A histogram (left), a kernel density estimate (middle), and a plot of quantiles versus the corresponding quantiles of a normal distribution (right) for the detrended data.

However, in our case, $\mathbf{Y}^P$ has some large values, and hence the detrended data $\tilde{\mathbf{Z}}$ have a left-skewed distribution, with outliers on the right (see Figure 2). It does not seem possible to adjust for this with a simple transformation. Nevertheless, FRK is justified as a spatial Best Linear Unbiased Predictor (BLUP), and we proceed with the analysis. (The dataset is synthetic and used only for illustration.)

2.3 The Basis Functions

Finally, before we can proceed with the actual analysis, we must specify the basis functions used in the matrix S in (4). This usually means specifying the type, the number, and the locations of the basis functions. Many options are possible, and we give some guidelines here. First, it is clear that we want the number of basis functions, r, to be much smaller than the number of measurements, n, in order for the SRE model to have any computational advantage over traditional kriging. Second, it is recommended that the basis functions be of different resolutions, to capture different scales of spatial variation. Each resolution contains a group of basis functions with the same smoothness and range, but the range varies between resolutions. There are typically a few smooth basis functions with large support, and many "spiky" basis functions with small support. The locations of the basis functions within a resolution should ideally cover the entire spatial domain of interest, D_s, and the locations of the basis functions from two different resolutions should probably not be coincident. Cressie and Johannesson (2008) used coincident basis functions, but we recommend avoiding this since it may introduce small oscillatory patterns in maps of (square root) mean squared prediction errors. As for specifying the type of basis functions, possible choices include bisquare functions, wavelets, indicator functions, empirical orthogonal functions, and harmonic functions (see Antoulas, 2005; Wikle, 2010, for overviews). The bisquare basis functions are illustrated in Figures 3 and 4. Cressie and Johannesson (2008) recommend that the radius of a basis function of a particular resolution be equal to 1.5 times the shortest distance between the center points of any two basis functions of that resolution. (When the basis functions within a resolution are not equidistant, Kang et al., 2011, modify this slightly.)

In addition, one might want to add a constant basis function despite the fact that the mean has already been subtracted as part of the trend removal. This serves as an extra check, to make sure the

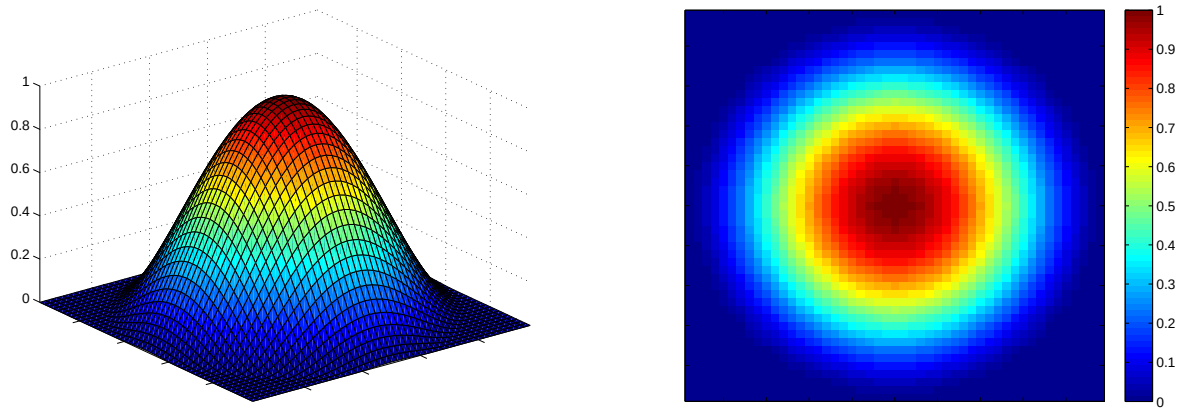

Figure 3: A two-dimensional bisquare function as a 3-D plot (left) and as an image plot (right).

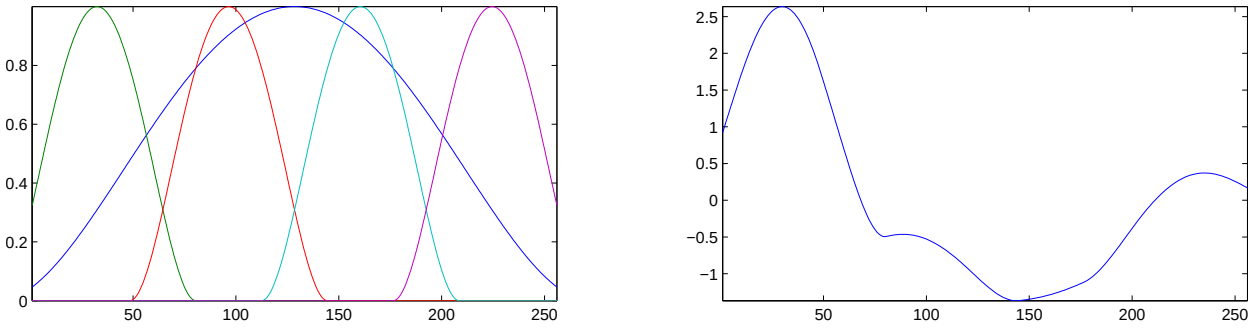

Figure 4: Five one-dimensional bisquare functions from two resolutions (left) and a random linear combination of them (right).

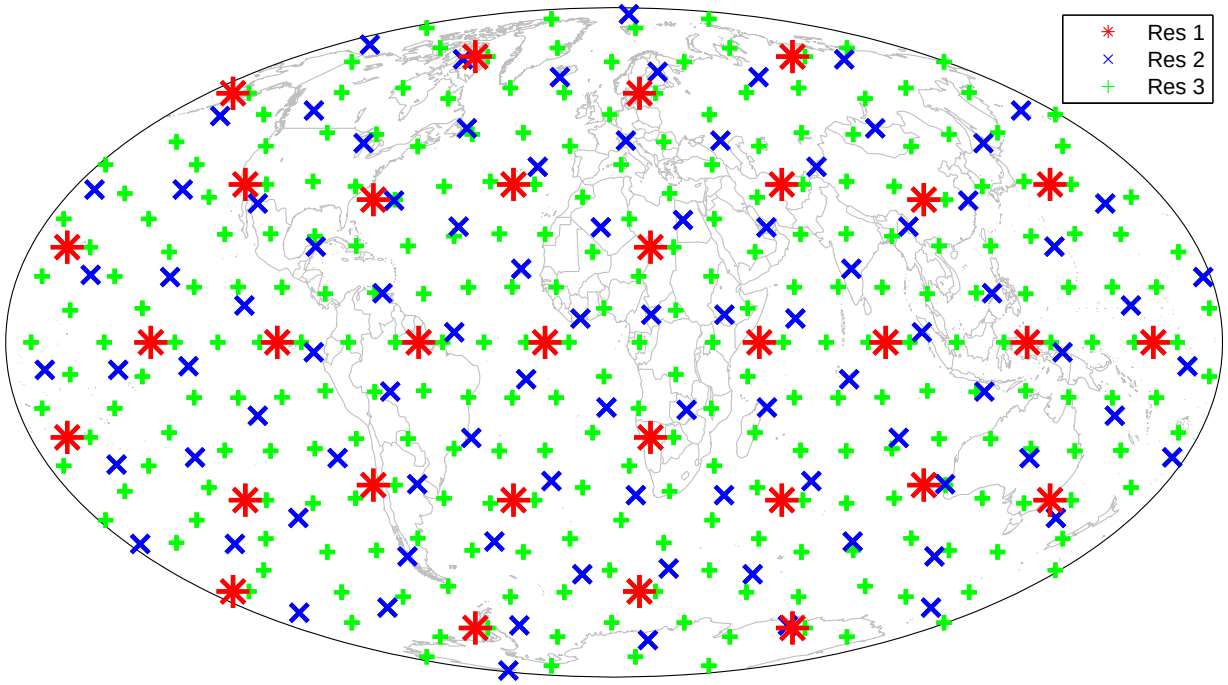

Figure 5: The locations of the centers of the 396 basis functions of the three resolutions.

rest of the random effects (not corresponding to the constant basis function) indeed have a mean of zero as specified in the model assumption. We have carried out some experiments that indicated that, while the estimated variance of the random-effect coefficient corresponding to the constant basis function is much lower than the estimated variances of the other random-effect coefficients, adding the the extra constant basis function does not cause numerical instabilities in the estimation of K or in the subsequent FRK that uses $\hat{K}$ in place of K.

In the XCO_2 data example described in Section 1.2, we chose $r = 396$ bisquare functions of 3 different resolutions. The 32 basis functions of resolution 1 have a great-arc radius of $6241km$, the 92 functions of resolution 2 have a great-arc radius of $3491km$, and the 292 functions of resolution 3 have a great-arc radius of $2047km$. The locations of the basis-function centers are shown in Figure 5.

3 Parameter Estimation

3.1 Specification/Estimation of the Measurement-Error Variance

The measurement-error variance consists of two components, σ^2_ϵ and $v_\epsilon(\cdot)$. The latter quantity is always assumed to be known, as explained in Section 1.3. It is also often the case that σ^2_ϵ can be specified in advance (e.g., from experiments with the measurement instrument). In cases where no prior information on the variance is available, it is possible to estimate it from the semivariogram near the origin. Define $N(h) \equiv \{(i,j) : \ ||\mathbf{s}_i - \mathbf{s}_j|| \in [h-\delta, h+\delta]\}$, the set of all indices corresponding to measurement-location pairs roughly a distance of h apart, for an arbitrarily chosen

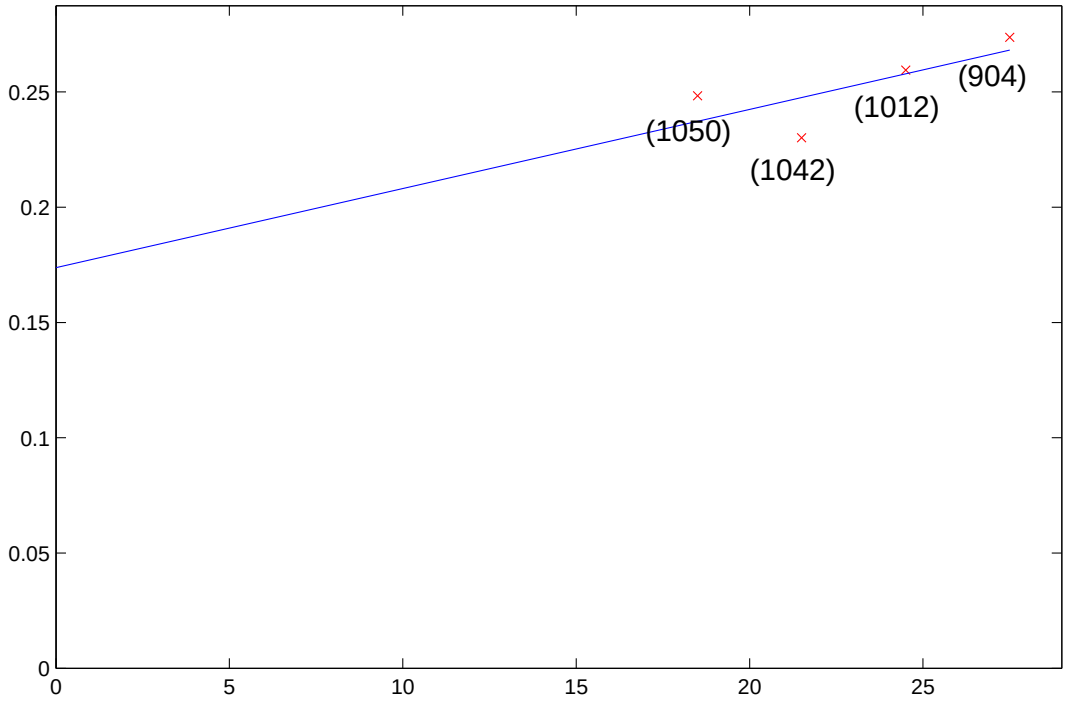

Figure 6: The estimated semivariogram, together with a straight line semivariogram fitted using WLS (Cressie, 1985). The number in parentheses shows the number of pairs of data, $|N(h)|$, that went into a bin located at h. Units on the x-axis are km (great-arc distance) and units on the y-axis are $(\text{ppm})^2$.

(discrete) range of values for the spatial lag h close to zero. The choice of the tolerance value $\delta > 0$ is a bias-variance-tradeoff problem (as are many problems in Statistics). The smaller the δ, the smaller the systematic bias, but also the less observations there are in the set $N(h)$ (which increases the variance). It is advisable to try several values and to see which gives results that do not result in highly fluctuating variogram estimates. The robust variogram estimator (Cressie and Hawkins, 1980) is,

$$2\gamma(h) \equiv \left(\frac{1}{|N(h)|} \sum_{N(h)} \left| \frac{\tilde{Z}(\mathbf{s}_i)}{\sqrt{v_\epsilon(\mathbf{s}_i)}} - \frac{\tilde{Z}(\mathbf{s}_j)}{\sqrt{v_\epsilon(\mathbf{s}_j)}} \right|^{0.5} \right)^4 \Big/ \Big(0.457 + 0.494/|N(h)| \Big),$$

where $|N(h)|$ is the number of elements in $N(h)$. To obtain an estimate of σ_ϵ^2 , we fit a straight line to the estimated semivariogram using weighted least squares (WLS; see Cressie, 1985, for details), and estimate σ_ϵ^2 as the line's intercept (Kang et al., 2010) at $h = 0$. That is, we fit the straight line, $\hat{\gamma}(h) = \hat{\gamma}(0+) + bh$, and estimate σ_ϵ^2 with:

$$\hat{\sigma}_\epsilon^2 = \hat{\gamma}(0+).$$

In the XCO2-data example, the measurement-error variance is known to be $\sigma_\epsilon^2 = 0.25$, since the measurement error was simulated using that value. For illustration purposes, we estimated the term from the variogram as described, allowing us to compare the estimate to what we know is

the truth, as a check. Using great-arc distances (Cressie, 1993, p. 265), we estimated the semi-variogram and fitted a straight line as shown in Figure 6, where the bins' endpoints are given by $\{17, 20, 23, 26, 29\}$ (km). The resulting estimate of σ_ϵ^2 was $\hat{\sigma}_\epsilon^2 = 0.1738$, an underestimate likely caused by the relatively large grid cells associated with the PCTM; recall they are 1.25° longitude by 1° latitude. For the remainder of this document, we used the true value, $\sigma_\epsilon^2 = 0.25$.

3.2 Method-of-Moments Estimation of Spatial-Dependence Parameters

In this document, we shall feature two possible estimation procedures for the covariance matrix K and the fine-scale-variation variance σ_ξ^2: binned method-of-moments (MM) estimation (described here) and maximum-likelihood (ML) estimation via the expectation-maximization (EM) algorithm (described in Section 3.3). If we decide to estimate the parameters using binned method of moments as introduced in Cressie and Johannesson (2008), we have found the procedure to be much more stable when $\hat{\sigma}_\xi^2$ is estimated from the variogram. Let h_s be the closest intersite distance between measurement-location pairs in the dataset, defined by the first bin chosen for the variogram estimation in Section 3.1. Following Kang et al. (2010), we set,

$$\hat{\sigma}^2_{\xi,\text{MM}} \equiv \frac{1}{2} \sum_{N(h_s)} \Big((\tilde{Z}(\mathbf{s}_i) - \tilde{Z}(\mathbf{s}_j))^2 - \hat{\sigma}_\epsilon^2 (v_\epsilon(\mathbf{s}_i) + v_\epsilon(\mathbf{s}_j)) \Big) / |N(h_s)|. \tag{7}$$

Should this quantity be negative, we simply set $\hat{\sigma}^2_{\xi,\text{MM}} = 0$.

To estimate K using MM, we first divide D_s into M bins, $\mathcal{B}_1, \ldots, \mathcal{B}_M$. To ensure stable estimation, we recommend a minimum of 15 measurements in each bin. Each $\mathcal{B}_j$ has associated with it an index set $\mathcal{I}_j \equiv \{i\colon\ \mathbf{s}_i \in \mathcal{B}_j\}$. We then define the matrix,

$$\hat{\Sigma}_M \equiv (\hat{\sigma}_{jl})_{j,l=1,\ldots,M}$$

through its elements,

$$\hat{\sigma}_{jl} \equiv \begin{cases} avg\{\tilde{Z}(\mathbf{s}_i)^2\colon\ i \in \mathcal{I}_j\}, & j = l, \\ avg\{\tilde{Z}(\mathbf{s}_i)\colon\ i \in \mathcal{I}_j\} \cdot avg\{\tilde{Z}(\mathbf{s}_i)\colon\ i \in \mathcal{I}_l\}, & j \neq l, \end{cases} \tag{8}$$

where $avg\{x_1, \ldots, x_k\} \equiv \sum_{i=1}^k x_i/k$. We define binned versions of the other quantities in (5), namely $\bar{S} \equiv [\bar{\mathbf{S}}_1', \ldots, \bar{\mathbf{S}}_M']'$, and $\bar{D} \equiv diag(\bar{d}_1, \ldots, \bar{d}_M)$, as follows:

$$\bar{\mathbf{S}}_j \equiv avg\{\mathbf{S}(\mathbf{s}_i)\colon\ i \in \mathcal{I}_j\},\ j = 1, \ldots, M,$$

and

$$\bar{d}_j \equiv avg\{\hat{\sigma}_\xi^2 v_\xi(\mathbf{s}_i) + \hat{\sigma}_\epsilon^2 v_\epsilon(\mathbf{s}_i)\colon\ i \in \mathcal{I}_j\},\ j = 1, \ldots, M.$$

The MM estimate of the matrix K is the value of K that makes

$$\bar{\Sigma}_M \equiv \bar{S} K \bar{S}' + \bar{D} \tag{9}$$

as close as possible to $\hat{\Sigma}_M$, as measured by the Frobenius norm. The Frobenius norm of a generic matrix $A \equiv (a_{ij})$ is defined as $\|A\| \equiv \sum_{i,j} a_{ij}^2$. The solution to this optimization problem is given by,

$$\hat{K}_{\text{MM}} = R^{-1} Q' (\hat{\Sigma}_M - \bar{D}) Q (R^{-1})',$$

where $\bar{S} = QR$ is the Q-R decomposition of $\bar{S}$ (i.e., Q is an orthogonal matrix and R is an upper-triangular matrix).

If $(\hat{\Sigma}_M - \bar{D})$ is not positive-definite, then $\hat{K}_{\text{MM}}$ will not be positive-definite either. In this case, we need to "lift" the eigenvalues, while at the same time we preserve the total variability (Kang et al., 2010). Specifically, we lift the eigenvalues, $\lambda_1, \ldots, \lambda_M$, of

$$A \equiv \bar{D}^{-1/2}(\hat{\Sigma}_M - \bar{D})\bar{D}^{-1/2},$$

such that the new eigenvalues, $\lambda_1^*, \ldots, \lambda_M^*$, are given by,

$$\lambda_i^* = \begin{cases} \lambda_0 \exp\{a(\lambda_i - \lambda_0)\}, & \lambda_i < \lambda_0 \\ \lambda_i, & \lambda_i \geq \lambda_0, \end{cases}$$

where $\lambda_0 > 0$ is chosen as the $[(M-q)/M]$-quantile of $\{\lambda_1, \ldots, \lambda_M\}$. Further, $a > 0$ is chosen such that $\text{tr}(\hat{\Sigma}_M^*) = \text{tr}(\hat{\Sigma}_M)$, where $\hat{\Sigma}_M^* \equiv \bar{D}^{1/2} A^* \bar{D}^{1/2} + \bar{D}$, and A^* is the same as A but with the new eigenvalues, $\{\lambda_1^*, \ldots, \lambda_M^*\}$. Kang et al. (2010) recommend choosing $q = r$, which leaves the $r+1$ largest eigenvalues unchanged, the smaller positive eigenvalues are decreased, and the negative eigenvalues are increased to be positive (thus ensuring positive-definiteness). However, sometimes λ_0 is negative and/or it is impossible to preserve the total variability (i.e., there exists no $a > 0$ such that $\text{tr}(\hat{\Sigma}_M^*) = \text{tr}(\hat{\Sigma}_M)$) when $q = r$. In this case, we reduce q by one and try again. We do this iteratively, until we have found a value for q (and therefore a value for λ_0), for which the total variability can be preserved. In extreme cases, we might even have to choose $\lambda_0 > \max\{\lambda_1, \ldots, \lambda_M\}$, which means that all of the original eigenvalues will be "lifted."

Once we have succesfully modified all the eigenvalues to be positive, the positive-definite MM estimate of K is then given by,

$$\hat{K}_{\text{MM}} = R^{-1} Q' (\hat{\Sigma}_M^* - \bar{D}) Q (R^{-1})'.$$

For the XCO2 data, the smallest bin for variogram estimation was $[17, 20]$ (in km of great-arc distance). We consider all pairs of locations falling into this bin as approximately exhibiting the distance h_s in (7). This results in the following estimate of the fine-scale-variation variance: $\hat{\sigma}^2_{\xi,\text{MM}} = 0.0889$.

To obtain the MM estimate of K, we combined the original $1.25^\circ \times 1^\circ$ grid into bins of size $7.5^\circ \times 5^\circ$ (i.e., $6 \times 5 = 30$ grid cells per bin), resulting in potentially 1,728 bins. To achieve stable estimation, we deleted all bins containing less than 15 observations. The remaining 953 bins were used for estimation of K.

The original estimate, $\hat{\Sigma}_M - \bar{D}$, was not positive-definite, so we lifted the eigenvalues as described above. However, it was not possible to preserve total variability with the default value $\lambda_0 = [(M-r)/M]$-quantile. Applying the iterative procedure described above, we chose $\lambda_0 = \max\{\lambda_1, \ldots, \lambda_M\} + 1$, so that all eigenvalues are lifted and the traces are matched. This resulted in a valid estimate, $\hat{K}_{\text{MM}}$, shown in Figure 7.

3.3 Parameter Estimation via the EM Algorithm

An alternative approach to parameter estimation in the SRE model is through maximum likelihood (ML) estimation via the expectation-maximization (EM) algorithm (Katzfuss and Cressie, 2009).

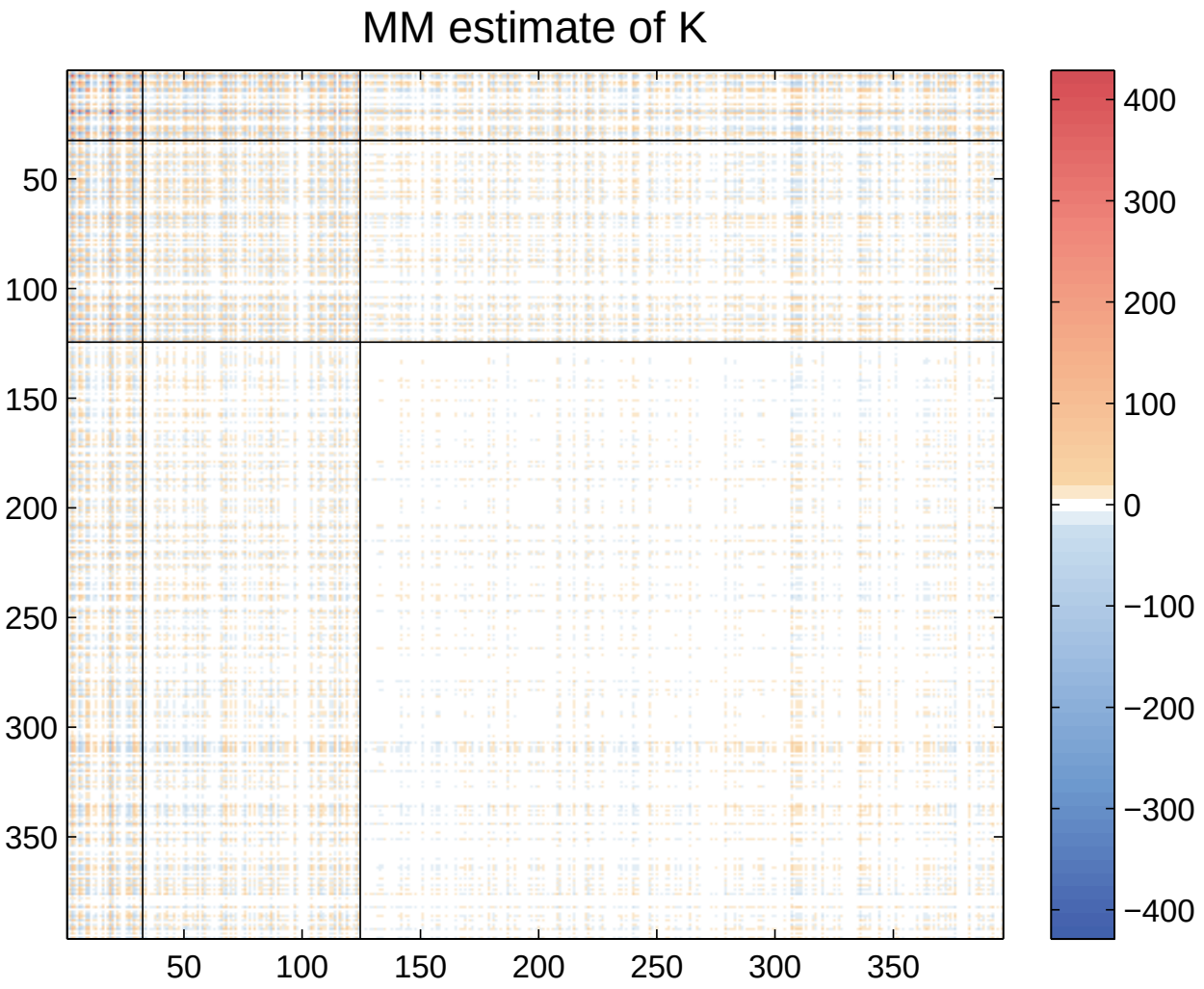

Figure 7: The MM estimate of the covariance matrix K in the XCO2-data example. (The matrix is blocked according to the resolutions.)

The EM algorithm is an iterative procedure that attempts to find the value of the parameter vector that maximizes the likelihood function, which is defined as the probability density function of the observed data as a function of the unknown parameters. To be more precise, the EM algorithm finds a solution to the likelihood estimating equations. (The likelihood estimating equations are obtained by setting the first derivative of the likelihood function with respect to the parameters to zero.)

The EM algorithm begins with starting values, $K^{[0]}$ and $\sigma_\xi^{2[0]}$, for the two parameters to be estimated. Then, we iteratively update both parameters for $t = 1, 2, \ldots$ (until convergence):

$$K^{[t+1]} = K^{[t]} - K^{[t]}S'\Sigma^{[t]-1}SK^{[t]} + (K^{[t]}S'\Sigma^{[t]-1}\tilde{\mathbf{Z}})(K^{[t]}S'\Sigma^{[t]-1}\tilde{\mathbf{Z}})'$$
$$\sigma_\xi^{2[t+1]} = \sigma_\xi^{2[t]} + (\sigma_\xi^{2[t]})^2 tr(\Sigma^{[t]-1}[\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}'\Sigma^{[t]-1} - I_n]V_\xi/n),$$

where $\Sigma^{[t]} \equiv SK^{[t]}S' + \sigma_\xi^{2[t]}V_\xi + \sigma_\epsilon^2 V_\epsilon$, and $\Sigma^{[t]-1}$ is shorthand for $(\Sigma^{[t]})^{-1}$. If σ_ϵ^2 is unknown, we replace it by its estimate from Section 3.1; importantly, we do not attempt to estimate it within the EM algorithm. Applying a Sherman-Morrison-Woodbury formula, inversion of $\Sigma^{[t]}$ only requires inversion of the $r \times r$ matrix $K^{[t]}$ and the diagonal matrix $D^{[t]} \equiv \sigma_\xi^{2[t]}V_\xi + \sigma_\epsilon^2 V_\epsilon$:

$$\Sigma^{[t]-1} = D^{[t]-1} - D^{[t]-1}S[K^{[t]-1} + S'D^{[t]-1}S]^{-1}S'D^{[t]-1}.$$

Using this result and rearranging the resulting expression, the update of σ_ξ^2 can be written as:

$$\sigma_\xi^{2[t+1]} = \sigma_\xi^{2[t]} + (\sigma_\xi^{2[t]})^2 tr([S'D^{[t]-1}V_\xi D^{[t]-1}S][K^{[t]-1} + S'D^{[t]-1}S]^{-1})/n$$
$$- (\sigma_\xi^{2[t]})^2 tr(D^{[t]-1}V_\xi)/n + (\sigma_\xi^{2[t]}V_\xi\Sigma^{[t]-1}\tilde{\mathbf{Z}})'V_\xi^{-1}(\sigma_\xi^{2[t]}V_\xi\Sigma^{[t]-1}\tilde{\mathbf{Z}}).$$

Since this is an iterative algorithm, there needs to be a stopping rule. We want to stop the algorithm at the iterate $t = T$, say, when any further change in how close the parameter estimates are to the solution to the likelihood equations, is not large enough to warrant the computational effort of carrying out another EM update. At this point, we say that the algorithm has converged. We then set $K_{\text{EM}} = K^{[T]}$ and $\sigma_{\xi,\text{EM}}^2 = \sigma_\xi^{2[T]}$. The easiest way to monitor convergence is to stack all unique elements of $K^{[t]}$ and $\sigma_\xi^{2[t]}$ into a vector, $\boldsymbol{\theta}^{[t]}$, and then stop the algorithm when $||\boldsymbol{\theta}^{[t+1]} - \boldsymbol{\theta}^{[t]}|| < \zeta$, for some small pre-specified value $\zeta > 0$. In the XCO2-data example, we set $\zeta = 10^{-6}\, r^2 \approx 0.157$, since there are on the order of r^2 parameters in $\boldsymbol{\theta}$.

The choice of starting values in the EM algorithm can be important. For example, one could use the MM estimates, from the previous subsection, or some other estimate or "guess" of the parameter values. However, one does not always want to carry out the MM estimation procedure, and thinking about what an appropriate starting value of K might look like can often be a daunting task for large r. In those cases, one should at least ensure that the starting values are valid (i.e., $K^{[0]}$ must be symmetric and positive-definite, and $\sigma_\xi^{2[0]}$ must be positive); a default choice might be $K^{[0]} = (0.9)\mathcal{V}^2 I_r$ and $\sigma_\xi^{2[0]} = (0.1)\mathcal{V}^2$, where $\mathcal{V}^2 \equiv \tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}/n$.

If the assumptions of normal distributions for all quantities are correct, the EM estimators should be more efficient than the MM estimators described in Section 3.2, resulting in a smaller mean-squared error (MSE). Of course, real data never follow any distributional assumptions exactly, and so questions of robustness of the estimators arise. In principle, since the MM estimators do not rely on any distributional assumptions, they should be more robust to departures from normal assumptions than the likelihood-based EM estimators (Zimmerman and Zimmerman, 1991).

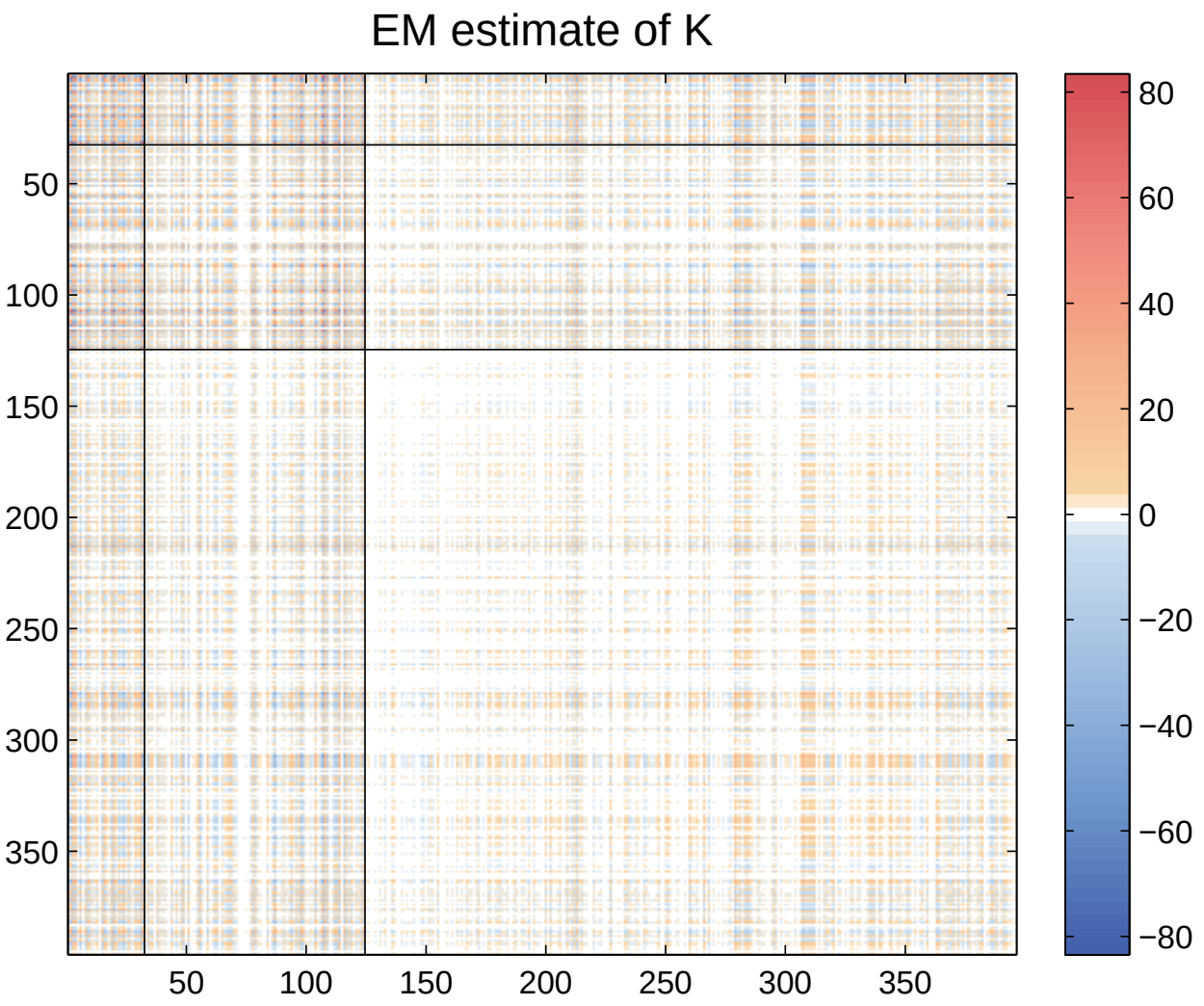

Figure 8: The EM estimate of the covariance matrix K in the XCO2-data example. (The matrix is blocked according to the resolutions.)

However, our previous experience (Katzfuss and Cressie, 2009, 2011b) indicates that, in practice, the EM estimators exploit spatial dependence in the data more accurately and precisely, and they seem to be somewhat more stable in some situations where the data are sparse.

For the XCO2-data example, we assume no prior knowledge of the parameter values, and we ran the EM algorithm with the default starting values $K^{[0]} = (0.9)\mathcal{V}^2 I_r \approx 1.61 I_r$ and $\sigma_\xi^{2[0]} = (0.1)\mathcal{V}^2 = 0.18$. The resulting estimates are $\hat{\sigma}^2_{\xi,\text{EM}} = 0.1018$ and $\hat{K}_{\text{EM}}$, which is shown in Figure 8. Compare this to $\hat{\sigma}^2_{\xi,\text{MM}} = 0.0889$, and the plot of $\hat{K}_{\text{MM}}$ shown in Figure 7; the difference between the two estimates will be explored in Section 5.

4 Optimal Spatial Prediction (Kriging)

Once all parameters have been estimated, it is straightforward to obtain optimal spatial predictions of the process of interest at any spatial location $\mathbf{s}_0 \in D_s$:

$$\hat{Y}(\mathbf{s}_0) \equiv E(Y(\mathbf{s}_0)|\mathbf{Z}) = \mu(\mathbf{s}_0) + \mathbf{k}(\mathbf{s}_0)'\Sigma^{-1}\tilde{\mathbf{Z}}. \tag{10}$$

In the case of the SRE model (4), this is known as the Fixed Rank Kriging (FRK) predictor. In this case, Σ is given by (5), and

$$\mathbf{k}(\mathbf{s}_0) \equiv SK\mathbf{S}(\mathbf{s}_0) + \sigma_\xi^2 v_\xi(\mathbf{s}_0) I(\mathbf{s}_0 \in \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}), \tag{11}$$

where all parameters are replaced by their (MM or EM) estimates obtained earlier. The FRK predictor (10) is the posterior mean of $Y(\mathbf{s}_0)$ (i.e., the mean of the true process at location $\mathbf{s}_0$ given the data $\mathbf{Z}$). It is an optimal predictor in that, if our model is correct, it minimizes the mean squared error between the predictor and the true value. Again, the inverse of $\Sigma = var(\mathbf{Z}) = SKS' + D$ is given by the Sherman-Morrison-Woodbury identity:

$$\Sigma^{-1} = D^{-1} - D^{-1}S[K^{-1} + S'D^{-1}S]^{-1}S'D^{-1}.$$

For each prediction, we can obtain an accompanying mean squared prediction error (MSPE),

$$\hat{\sigma}^2(\mathbf{s}_0) \equiv E(\hat{Y}(\mathbf{s}_0) - Y(\mathbf{s}_0))^2 = \mathbf{S}(\mathbf{s}_0)'K\mathbf{S}(\mathbf{s}_0) + \sigma_\xi^2 v_\xi(\mathbf{s}_0) - \mathbf{k}(\mathbf{s}_0)'\Sigma^{-1}\mathbf{k}(\mathbf{s}_0).$$

Recalling the definition of $\mathbf{k}(\mathbf{s}_0)$ in (11), we see that the MSPE is decreased if the prediction location coincides with one of the data locations. By letting $\mathbf{s}_0$ vary over a fine grid defined on the domain D_s, we obtain a map of predictions and an accompanying map of MSPEs.

In the XCO2-data example, we obtained FRK predictions and accompanying standard errors (root MSPEs) using both MM estimates (Figure 9) and EM estimates (Figure 10). The difference between the results from the two estimation methods are likely due to the fact that $\hat{K}_{\text{MM}}$ is close to singular, and so inversion of this matrix is numerically unstable. We shall explore the differences between the two estimation methods' results in Section 5.

5 Diagnostics

It is important to check the success of the estimation and the FRK predictions. This can be done in various ways. Here, we have used two different estimation methods, MM and EM estimation, and our diagnostics will involve a comparison of the two. We see that the MM and EM estimates of $K \equiv (k_{ij})$ are very different; the magnitude of the elements of $\hat{K}_{\text{MM}} \equiv (\hat{k}_{ij,\text{MM}})$ in Figure 7 is around 430, and the magnitude of the elements of $\hat{K}_{\text{EM}} \equiv (\hat{k}_{ij,\text{EM}})$ in Figure 8 is around 80. Figure 11 gives a plot of $\{\hat{k}_{ij,\text{MM}}\}$ versus $\{\hat{k}_{ij,\text{EM}}\}$. The magnitudes are clearly different, although the pattern of relative values is similar.

We calculated the Frobenius norm described below (9) for both estimates of K: the norm achieved by the EM estimates is 391.2822, which is higher than the value 332.5180 achieved by the MM estimates, although this is to be expected.

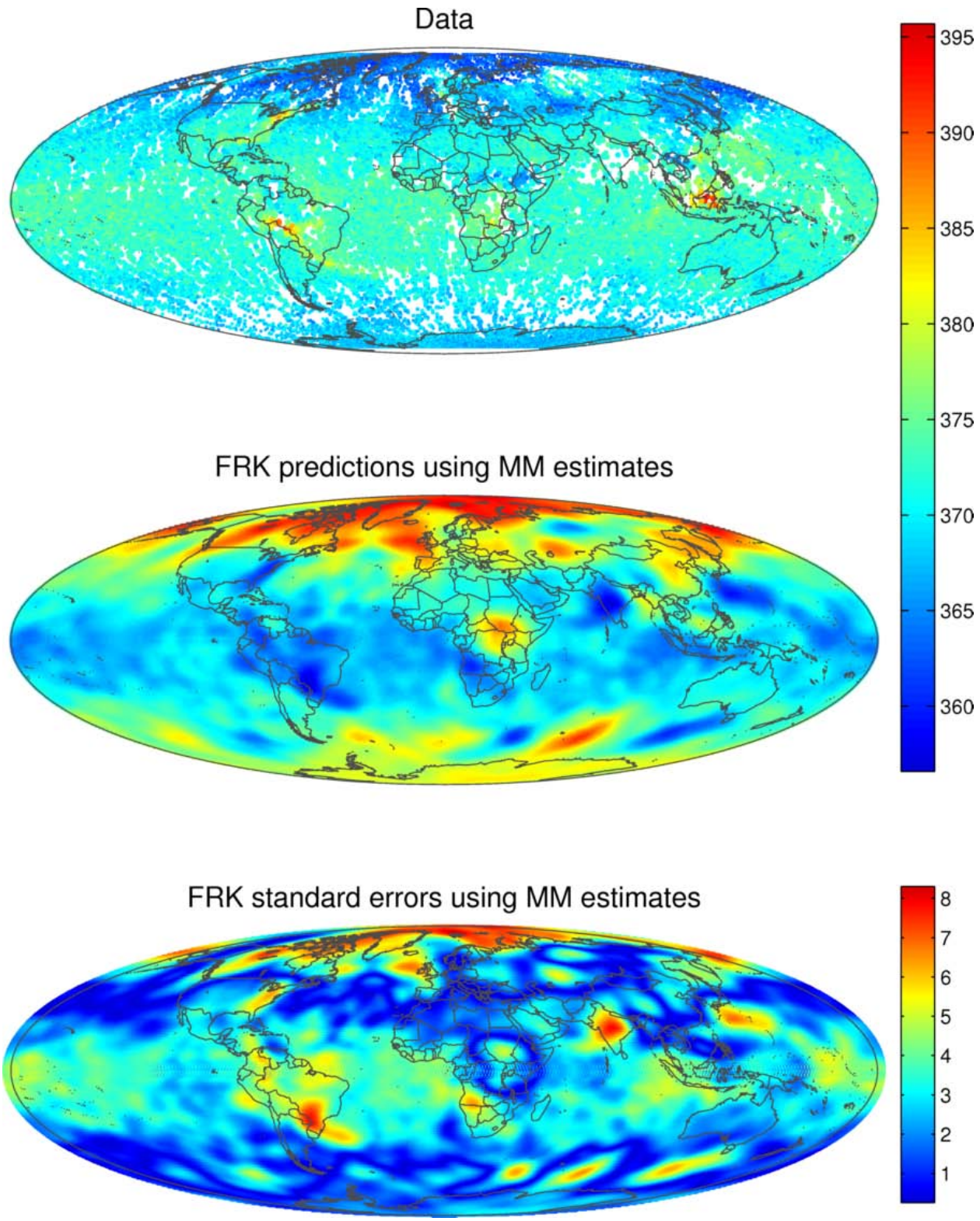

Figure 9: The FRK predictions and standard errors (root MSPEs) using the MM estimates.

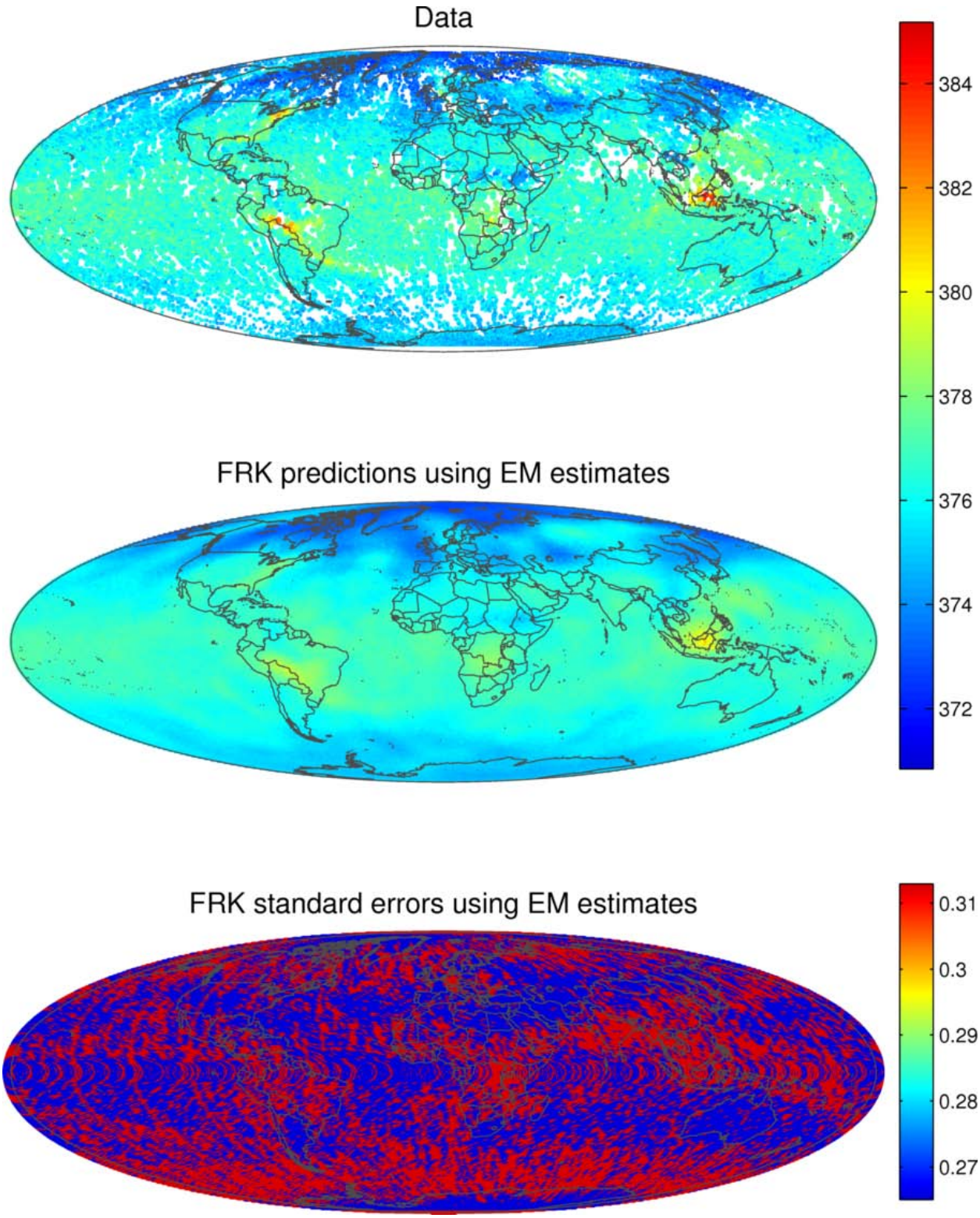

Figure 10: The FRK predictions and standard errors (root MSPEs) using the EM estimates. Note that the color scales are different from those of Figure 9.

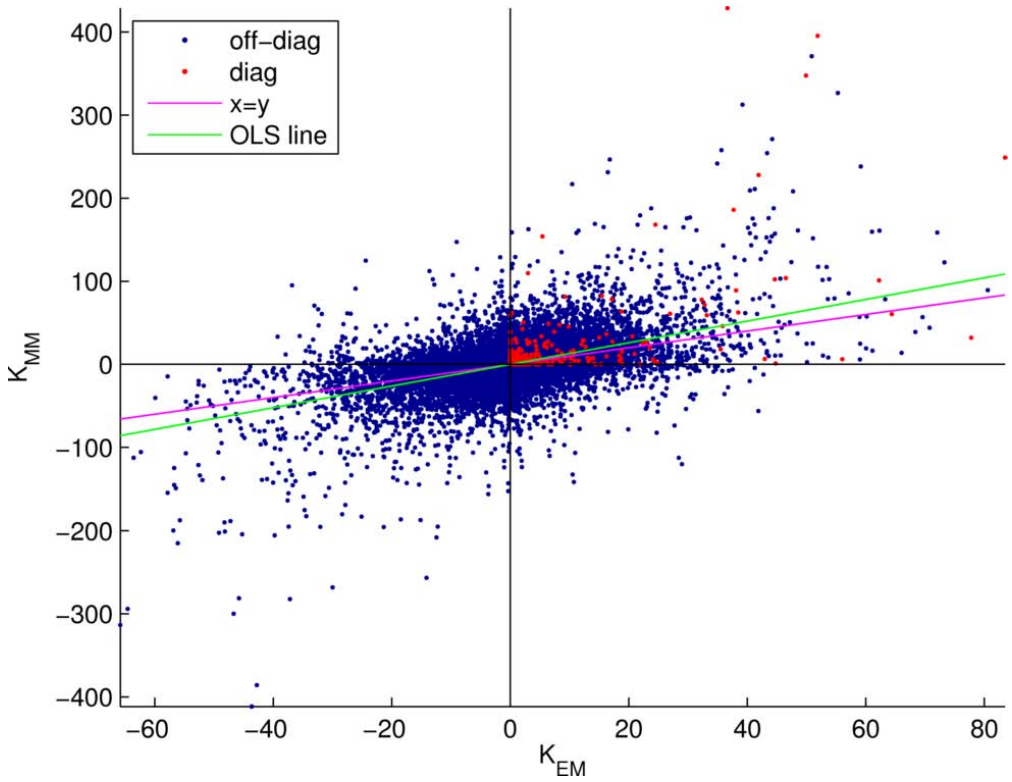

Figure 11: Plot of $\{\hat{k}_{ij,\text{MM}}\}$ versus $\{\hat{k}_{ij,\text{EM}}\}$.

Note that the EM estimates are derived based on assuming that the data follow a joint normal distribution. We have seen in Section 2 that this assumption might not be justified for the XCO2 data, due to rather heavy tails in the empirical distribution of the data. Hence, we decided to ascertain the effect of deleting the extreme observations in the tails (the 20 data points that were farther than four standard deviations away from the mean). For this truncated dataset, we re-ran the EM algorithm and found that the magnitude of the elements of the resulting $\hat{K}_{\text{EM}}$ were still around 80 (i.e., much lower than for $\hat{K}_{\text{MM}}$).

Of course, $cov(Y(\mathbf{s}), Y(\mathbf{u}))$ is given by $\mathbf{S}(\mathbf{s})'K\mathbf{S}(\mathbf{u})' + \sigma_\xi^2 I(\mathbf{s} = \mathbf{u})$, and so the magnitude of the estimated elements of K do not directly tell us what the estimated covariance structure is. We compared $diag(S\hat{K}_{\text{MM}}S')$ to $diag(S\hat{K}_{\text{EM}}S')$, and the elements of the two vectors were of roughly equal magnitude. As this summary only contains marginal variances (as opposed to covariances), we chose to follow the diagnostics of Cressie and Johannesson (2008) and explore the estimated covariance structures using plots of empirical directional root-semivariograms, together with their theoretical counterparts (as estimated using MM and EM, respectively), at several reference points on the globe (Figure 12).

The theoretical root-semivariogram evaluated at two locations, $\mathbf{s}_1$ and $\mathbf{s}_2$, for the SRE model is

given by,

$$\begin{aligned}\sqrt{\gamma(\mathbf{s}_1,\mathbf{s}_2)} &\equiv \{\frac{1}{2}E([Z(\mathbf{s}_1)-Z(\mathbf{s}_2)]^2)\}^{0.5}\\ &= \frac{1}{\sqrt{2}}\big[((\mathbf{x}(\mathbf{s}_1)-\mathbf{x}(\mathbf{s}_2))'\boldsymbol{\beta})^2+(\mathbf{S}(\mathbf{s}_1)-\mathbf{S}(\mathbf{s}_2))'K(\mathbf{S}(\mathbf{s}_1)-\mathbf{S}(\mathbf{s}_2))\\ &\quad + I(\mathbf{s}_1\neq\mathbf{s}_2)\big(\sigma^2_\xi(v_\xi(\mathbf{s}_1)+v_\xi(\mathbf{s}_2))+\sigma^2_\epsilon(v_\epsilon(\mathbf{s}_1)+v_\epsilon(\mathbf{s}_2))\big)\big]^{0.5}.\end{aligned}$$

The theoretical quantities are estimated and plotted in Figure 12; they are obtained by replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$ (from Section 2.1) and K and σ^2_ξ with their MM and EM estimates, respectively. Also, $\gamma(\mathbf{s},\mathbf{s})\equiv 0$ for any $\mathbf{s}$, but the first value of each of the root-semivariograms in Figure 12 is almost on the vertical axis, not exactly so.

It can be seen from Figure 12 that the estimated covariance structures are more similar than the (very different) MM and EM estimates of K would suggest. However, overall, the estimated covariance structure using the EM estimates is closer to the empirical structure implied by the data, than the estimated structure using the MM estimates. The problems with the latter (e.g., in the top left panel) might be due to the eigenvalue lifting that was rather severe, in order to ensure that $\hat{K}_{\text{MM}}$ was positive-definite and total variability was preserved (see Section 3.2).

6 Possible Extensions

6.1 Generalization of the Distribution of the Fine-Scale Variation

The function $v_\xi(\cdot)$, which determines the variance heterogeneity of the fine-scale variation, is assumed known, but this assumption could be relaxed. One could assume that $v_\xi(\cdot)$ is a function of the form,

$$v_\xi(\cdot)=\exp\{\mathbf{S}_\xi(\cdot)'\boldsymbol{\psi}\},$$

where $\mathbf{S}_\xi(\cdot)$ is a vector of r_ξ basis functions, much like $\mathbf{S}(\cdot)$, and $\boldsymbol{\psi}$ are unknown parameters. For example, we could set $\mathbf{S}_\xi(\cdot)$ to be the first resolution of basis functions in $\mathbf{S}(\cdot)$. For the vector of basis-function coefficients, we could then estimate the parameters in an EM algorithm. A good starting value is $\boldsymbol{\psi}^{[0]}=\mathbf{0}$ and $\boldsymbol{\psi}$ could be estimated in the EM algorithm by adding one extra update:

$$\boldsymbol{\psi}^{[t+1]}=\boldsymbol{\psi}^{[t]}-A^{-1}\mathbf{b},$$

where

$$\begin{aligned}\mathbf{b}&=S'_\xi(I_n-\Lambda/\sigma^2_\xi)\mathbf{1}_n\\ A&=S'_\xi\Lambda S_\xi/\sigma^2_\xi,\end{aligned}$$

and $\Lambda\equiv diag(\{E_{\boldsymbol{\theta}^{[t]}}(\xi(\mathbf{s}_i)^2|\mathbf{z})\exp(-\mathbf{S}_\xi(\mathbf{s}_i)'\boldsymbol{\psi}^{[t]}):\ i=1,\ldots,n\})$. For more details, see Katzfuss and Cressie (2011a).

6.2 Non-Point Support

Until now, we have assumed that both measurements and predictions are at point-level support $\mathbf{s}$. In reality, measurements are often made at an aggregated level. For example, satellite instruments

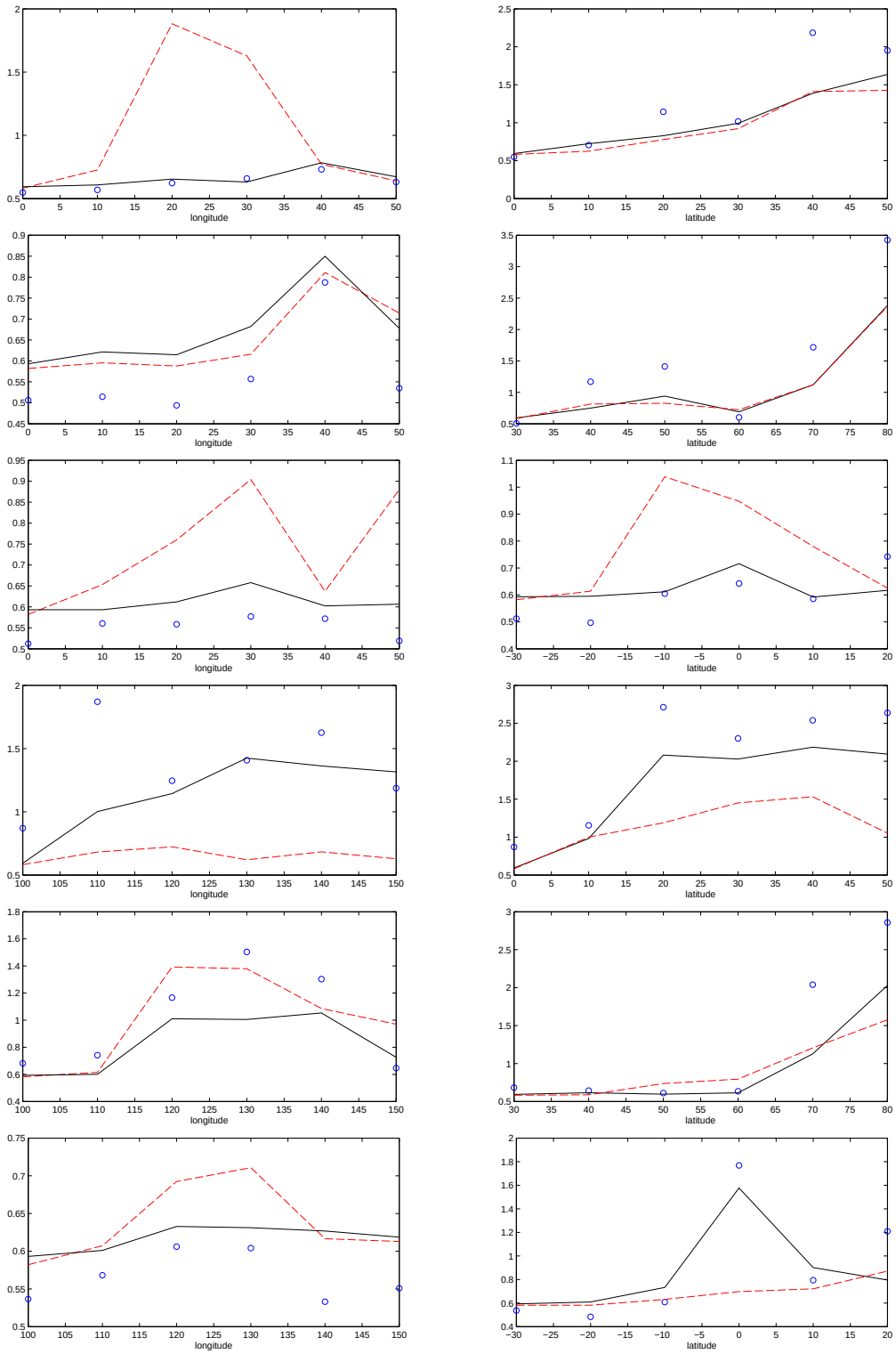

Figure 12: Directional root-semivariograms at various reference points in longitudinal (left column) and latitudinal (right column) directions. Blue circles: Empirical root-semivariograms. Red dashed lines: Theoretical root-semivariograms using the MM estimates. Black solid lines: Theoretical root-semivariograms using the EM estimates.

typically have a "footprint"; that is, their measurements are really averages or integrals over small areas. To accommodate this in the model, we can specify the model at a very fine grid of areas that are so small that they can be considered to be the same as point locations, for all practical purposes. These very small areas are called basic areal units (BAUs) by Nguyen et al. (2010). After evaluating the trend functions, $\mathbf{x}(\cdot)$, and the basis functions, $\mathbf{S}(\cdot)$, at each of the BAUs, we then simply average over all BAUs within a given footprint $\mathcal{S}$ for which a measurement is available:

$$\mathbf{x}(\mathcal{S}) \equiv \frac{1}{|\tilde{D}_s \cap \mathcal{S}|} \sum_{\mathbf{s} \in \tilde{D}_s \cap \mathcal{S}} \mathbf{x}(\mathbf{s})$$

and

$$\mathbf{S}(\mathcal{S}) \equiv \frac{1}{|\tilde{D}_s \cap \mathcal{S}|} \sum_{\mathbf{s} \in \tilde{D}_s \cap \mathcal{S}} \mathbf{S}(\mathbf{s}),$$

where $\tilde{D}_s$ denotes the grid of BAUs over D_s. This same averaging can also be done for $\xi(\cdot)$ in (3) and $\epsilon(\cdot)$ in (1), so that $v_\xi(\mathcal{S})$ in (3) and $v_\epsilon(\mathcal{S})$ in (1) are then inversely proportional to the number of BAUs in a given footprint $\mathcal{S}$.

If predictions at an aggregated level are also of interest, we can make predictions for each of the BAUs and then average over all predicted values corresponding to the BAUs within a prediction region of interest. The prediction variances are obtained from the BAUs' prediction variances and covariances. Further details of the use of BAUs can be found in Nguyen et al. (2010).

7 Matlab Code

We have made available Matlab code that can be used to carry out the computations described above; the website

`http://www.stat.osu.edu/~sses/collab_co2.html`

can be consulted. Specifically, the files `FRK.m`, `EM.m`, `variogram_estimation.m`, and `binest.m` contain Matlab functions implementing, respectively, FRK, EM estimation for FRK, estimation of σ^2_ϵ and σ^2_ξ based on the variogram, and MM estimation for FRK. The function `Create_S.m` creates the matrix of basis functions. The function `binest.m` relies on support functions `bin.m` and `trvar.m`.

In addition, we have included a file called `xco2analysis.m`, which contains code for the analysis of the XCO2-data example (called `xco2data.mat`) used in this document, and it gives an example of how to call the functions above.

Acknowledgment

This research was supported by NASA under grant NNX08AJ92G issued through the ROSES Carbon Cycle Science Program.

References

Antoulas, A. (2005), *Approximation of Large-Scale Dynamical Systems*, Philadelphia, PA: SIAM.

Box, G. and Cox, D. (1964), "An analysis of transformations," *Journal of the Royal Statistical Society, Series B*, 26, 211–252.

Chatterjee, A. and Kawa, S. R. (2009), personal communication.

Cressie, N. (1985), "Fitting variogram models by weighted least squares," *Mathematical Geology*, 17, 563–586.

— (1993), *Statistics for Spatial Data*, rev. edn., New York, NY: John Wiley & Sons.

Cressie, N. and Hawkins, D. (1980), "Robust estimation of the variogram: I," *Mathematical Geology*, 12, 115–125.

Cressie, N. and Johannesson, G. (2008), "Fixed rank kriging for very large spatial data sets," *Journal of the Royal Statistical Society, Series B*, 70, 209–226.

Cressie, N., Shi, T., and Kang, E. L. (2010), "Fixed rank filtering for spatio-temporal data," *Journal of Computational and Graphical Statistics*, 19, 724–745.

Crisp, D. and Johnson, C. (2005), "The orbiting carbon observatory mission," *Acta Astronautica*, 56, 193–197.

Kang, E. L., Cressie, N., and Sain, S. R. (2011), "Combining outputs from the NARCCAP regional climate models using a Bayesian hierarchical model," *Applied Statistics*, under revision.

Kang, E. L., Cressie, N., and Shi, T. (2010), "Using temporal variability to improve spatial mapping with application to satellite data," *Canadian Journal of Statistics*, 38, 271–289.

Katzfuss, M. and Cressie, N. (2009), "Maximum likelihood estimation of covariance parameters in the spatial-random-effects model," in *Proceedings of the Joint Statistical Meetings*, Alexandria, VA: American Statistical Association, pp. 3378–3390.

— (2011a), "Bayesian hierarchical spatio-temporal smoothing for massive datasets," *Environmetrics*, under revision.

— (2011b), "Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets," *Journal of Time Series Analysis*, 32, 430–446.

Kawa, S. R. (2004), "Global CO_2 transport simulations using meteorological data from the NASA data assimilation system," *Journal of Geophysical Research*, 109, 1–17.

Michalak, A. (2010), personal communication.

Nguyen, H., Cressie, N., and Braverman, A. (2010), "Spatial statistical data fusion for remote-sensing applications," Technical Report No. 849, Department of Statistics, The Ohio State University, Columbus, OH.

Shi, T. and Cressie, N. (2007), "Global statistical analysis of MISR aerosol data: A massive data product from NASA's Terra satellite," *Environmetrics*, 18, 665–680.

Strand, G. (2004), "Development of the PCM-CSM Transition Model (PCTM)," http://www.cgd.ucar.edu/pcm/admin/pctm.html.

Wikle, C. K. (2010), "Low-rank representations for spatial processes," in *Handbook of Spatial Statistics*, eds. Gelfand, A. E., Fuentes, M., Guttorp, P., and Diggle, P., Boca Raton, FL: Chapman and Hall/CRC, pp. 107 – 118.

Winker, D., Pelon, J., and McCormick, M. (2003), "The CALIPSO mission: Spaceborne lidar for observation of aerosols and clouds," in *Lidar Remote Sensing for Industry and Environment Monitoring III*, eds. Singh, U. N., Itabe, T., and Liu, Z., Bellingham, WA: SPIE, vol. 4893 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Proceedings*, pp. 1–11.

Zimmerman, D. and Zimmerman, M. (1991), "A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors," *Technometrics*, 33, 77–91.