NIASRA
NATIONAL INSTITUTE FOR APPLIED
STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong

Working Paper

16-14

A Comparison of Spatial Predictors when Datasets Could be Very Large

Jonathan R. Bradley, Noel Cressie and Tao Shi

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email:
anica@uow.edu.au

A Comparison of Spatial Predictors when Datasets Could be Very Large

Jonathan R. Bradley[1], Noel Cressie[2], Tao Shi[3]

Abstract

In this article, we review and compare a number of methods of spatial prediction. To demonstrate the breadth of available choices, we consider both traditional and more-recently-introduced spatial predictors. Specifically, in our exposition we review: traditional stationary kriging, smoothing splines, negative-exponential distance-weighting, Fixed Rank Kriging, modified predictive processes, a stochastic partial differential equation approach, and lattice kriging. This comparison is meant to provide a service to practitioners wishing to decide between spatial predictors. Hence, we provide technical material for the unfamiliar, which includes the definition and motivation for each (deterministic and stochastic) spatial predictor. We use a benchmark dataset of CO_2 data from NASA's AIRS instrument to address computational efficiencies that include CPU time and memory usage. Furthermore, the predictive performance of each spatial predictor is assessed empirically using a hold-out subset of the AIRS data.

Keywords: Best linear unbiased predictor; GIS; massive data; reduced rank statistical models; model selection

[1] (to whom correspondence should be addressed) Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211, bradleyjr@missouri.edu

[2] National Institute for Applied Statistics Research Australia, University of Wollongong, Australia

[3] Department of Statistics, The Ohio State University

1 Introduction

We are in an era of "big data," where the sizes of available datasets are becoming increasingly larger. For example, consider datasets on earnings from the US Census Bureau's Longitudinal Employer-Household Dynamics program, on weather from the National Oceanic and Atmospheric Administration (NOAA), and on public health from the Centers for Disease Control and Prevention (CDC). In the commercial sector, big data is now available using technology that allows companies to gather (anonymously) information on purchases (Hormozi and Giles, 2004). Pharmaceutical organizations amass large amounts of drug-testing data through combinatorial chemistry, medium-to-high-throughput screening (HTS), and other new technologies (Campbell, 2010). Many of these datasets can be very large in size; for example, the National Aeronautics and Space Administration (NASA) collects millions of atmospheric CO_2 measurements per month over the globe using the Atmospheric Infrared Sounder (AIRS) instrument on the Terra satellite.

As a result, big data is an important and growing topic in statistics. In the spatial-data setting, there are additional challenges. For example, AIRS CO_2 data have global extent, but they are spatially sparse. Additionally, they exhibit complex spatial dependencies that may be nonstationary. Thus, the complexity of "big spatial data" has motivated many to propose new statistical methodologies for spatial prediction (e.g., see Cressie and Wikle, 2011, Ch. 4; Sun et al., 2012, for reviews). In particular, there are methods that use separable covariance functions, tapered covariance matrices, composite likelihoods, and low-dimensional latent Gaussian processes. These methodologies are all motivated by the fact that the Gaussian likelihood is difficult to compute when the dataset is large. Specifically, the Gaussian likelihood involves the computation of an inverse and a determinant of an $n \times n$ covariance matrix, a task that is on the order of n^3 computations, where n represents the size of the spatial dataset.

Despite the growing number of spatial predictors that are becoming available, there has been no comprehensive comparison between (and among) both traditional and modern spatial predictors.

Such a comparison would be highly useful to the more general scientific community. In particular, the GIS community often use spatial interpolation and smoothing (e.g., see Xin et al., 2000), and would benefit from such a comparison. Hence, we shall review the parameterization, the algorithm, and the motivation of seven spatial predictors, also considered by Bradley et al. (2014a) in the context of local spatial predictor selection. We consider three traditional spatial predictors, namely traditional stationary kriging, smoothing splines, and negative-exponential distance-weighting; and we consider four more-recently-introduced spatial predictors, namely Fixed Rank Kriging, one based on modified predictive processes, one based on a stochastic partial differential equation, and lattice kriging. We use a benchmark dataset of CO_2 data from NASA's AIRS instrument to empirically compare the predictive performances, computation times, and memory usage of these key spatial predictors.

Kriging based on a stationary covariance function has become a method of spatial prediction covered in standard textbooks (e.g., Cressie, 1993; Banerjee et al., 2004; Schabenberger and Gotway, 2005; Cressie and Wikle, 2011) and has a rich history (see Cressie, 1990, and the references therein). Since this method of spatial prediction has become a staple, we consider it in our study of AIRS CO_2 and call the approach *traditional stationary kriging* (TSK). Another common approach is spatial interpolation using splines, which is obtained by minimizing a penalized-least-squares criterion (e.g., see Wahba, 1990; Nychka, 2001). Hence, we also consider *smoothing splines* (SSP) in our comparisons.

However, both TSK and SSP are not "scalable" to large datasets; for example, they cannot be computed for the entire AIRS dataset for computational reasons. One simple ad hoc solution to this "big data" problem is a spatial predictor based on *negative-exponential distance-weighting* (EDW) (see Cressie, 1993, p. 371, for a discussion on these types of deterministic methods). Here, a datum's negative log weight is proportional to the Euclidean distance from the prediction location to the datum's location (see Section 2.3 for more details on EDW).

Although EDW is computationally efficient, we are predominantly interested in spatial predictors that are derived from statistical models and are appropriate for big data. For example, low-rank statistical models provide a computationally efficient way to obtain the optimal kriging predictor and associated measures of error. For this reason, low-rank statistical modeling for spatially referenced data is a popular method in the literature. In the spatial univariate setting, see Cressie and Johannesson (2006), Shi and Cressie (2007), Banerjee et al. (2008), Cressie and Johannesson (2008), and Kang and Cressie (2011). In the spatio-temporal setting, see Wikle and Cressie (1999), Wikle et al. (2001), Cressie et al. (2010a), Cressie et al. (2010b), Katzfuss and Cressie (2011, 2012), and Bradley et al. (2014b). In this article, we focus on two low-rank spatial predictors that have motivated much of this literature: *Fixed Rank Kriging* (FRK), and the *Modified Predictive Process* (MPP) approach.

FRK seeks efficient calculation of the kriging predictor in the setting where n is very-large-to-massive. An advantage of FRK is that the inverse of the covariance matrix can be achieved efficiently using the Sherman-Morrison-Woodbury identity (e.g., Henderson and Searle, 1981), allowing FRK to be scalable (see Section 2.4 for more details). The approach taken by MPP is similar and starts by first predicting a low-rank random effect called the predictive process. Then, predictions of a latent process are found by multiplying the prediction of the random effect by a set of basis functions (see Section 2.5 for more details). Some have criticized the use of a low-rank representation of a latent Gaussian process and believe that in many settings much of the variability occurs at high frequencies (see Lindgren et al., 2011; Stein, 2014, for discussions). However, it should be noted that high-frequency or discontinuous basis functions can address this criticism.

The remaining two spatial predictors are based on imposing more parametric assumptions on the latent random process. One is based on a *stochastic partial differential equation* (SPD) approach proposed by Lindgren et al. (2011), and the other is *lattice kriging* (LTK) proposed by Nychka et al., 2014. These two methods of prediction achieve computational efficiency by placing structure on the precision matrix of the random-effects vector (see Sections 2.6 - 2.7 for more details). These four recent proposals, coupled with the fact that spatial prediction using big datasets is an important problem, adds additional motivation for our comparison.

In Section 2, we present seven methods of spatial prediction, ranging from the classical to the more recent ones designed to handle very-large-to-massive datasets; both deterministic and stochastic spatial predictors are considered. Details surrounding the predictors are presented systematically, along with the motivation behind each spatial predictor. In Section 3, we apply and compare these predictors using different-sized datasets of mid-tropospheric CO_2 measurements. We include the computation time and memory usage of each predictor in the comparison, along with an empirical comparison of predictive performance using a hold-out dataset. A concluding discussion is provided in Section 4.

2 Seven Spatial Predictors

In this section, we provide details on the spatial predictors considered. They are: traditional stationary kriging (TSK), smoothing splines (SSP), negative-exponential-distance-weighting (EDW), Fixed Rank Kriging (FRK), the modified predictive process approach (MPP), the SPDE approach (SPD), and lattice kriging (LTK). Notice that the spatial predictors could be deterministic or stochastic, and we have chosen several that have been proposed recently to handle big spatial datasets. Details of the seven predictors are set out according to: the parameterization associated with each spatial predictor; the algorithm used to compute the spatial predictor; and the motivation behind the spatial predictor.

Many of the spatial predictors that we consider can be motivated by a spatial mixed effects (SME) model (e.g., Cressie and Johannesson, 2006, 2008):

$$\text{Data Model}: \quad Z(\mathbf{u}) = Y(\mathbf{u}) + \varepsilon(\mathbf{u}) \tag{1}$$

$$\text{Process Model}: \quad Y(\mathbf{u}) = \mu(\mathbf{u}) + \nu(\mathbf{u}) + \xi(\mathbf{u}); \;\; \mathbf{u} \in D, \tag{2}$$

where $\varepsilon(\cdot)$ represents measurement error; $\mu(\cdot)$ is a deterministic mean function; $\nu(\cdot)$ models small-scale variation; $\xi(\cdot)$ is a term that captures (often non-smooth) micro-scale variation; and $D \equiv \{\mathbf{u}_j : j = 1,...,N\} \subset \mathbb{R}^d$ is a generic finite set of prediction locations. All stochastic components, $\varepsilon(\cdot)$, $\nu(\cdot)$, and $\xi(\cdot)$ are assumed mutually independent. A very flexible way to represent $\nu(\cdot)$ is through a basis-function expansion,

$$\nu(\mathbf{u}) = \mathbf{S}_r(\mathbf{u})'\boldsymbol{\eta};\ \ \mathbf{u} \in D, \tag{3}$$

where $\mathbf{S}_r(\cdot)$ is an r-dimensional vector of spatial basis functions and $\boldsymbol{\eta}$ is an r-dimensional vector of random coefficients.

The spatial random process $Z(\cdot)$ represents the "data," and it is observed over a subset of the spatial domain of interest $D \subset \mathbb{R}^d$; that is, $Z(\cdot)$ is observed at locations in the set $D_O \equiv \{\mathbf{s}_i : i = 1,...,n\} \subset D$. The latent process $Y(\cdot)$ is of principal interest, and one wishes to predict it from the data $\{Z(\mathbf{s}) : \mathbf{s} \in D_O\}$. It is assumed that $\varepsilon(\cdot)$ is a white-noise Gaussian process with mean zero and known var($\varepsilon(\cdot)$)= $\sigma_\varepsilon^2 V(\cdot)$, where $V(\cdot) > 0$ is a known function that captures heteroskedasticity. Note that often variance estimates are obtained from equipment calibration and quality assurance, in which case σ_ε^2 can be considered as known. Let $\mu(\cdot) \equiv \mathbf{x}(\cdot)'\boldsymbol{\beta}$, where $\mathbf{x}(\mathbf{s})$ is a p-dimensional vector of known spatial covariates defined on all $\mathbf{s} \in D$, and $\boldsymbol{\beta}$ is a p-dimensional vector of unknown regression coefficients.

The low-rank representation of $\nu(\cdot)$ requires further explanation. For $i = 1,...,r$, the i-th element of $\mathbf{S}_r(\cdot)$ is given by the function, $S_{i,r} : D \to \mathbb{R}$; and the r-dimensional random vector $\boldsymbol{\eta}$ is specified as a Gaussian process with mean zero and $r \times r$ covariance matrix $\mathbf{K}$. Finally, the random process $\xi(\cdot)$ is assumed to be a Gaussian white-noise process with mean zero and variance σ_ξ^2.

It will be seen below that the SME model motivates many of the stochastic predictors, although clearly not so for the deterministic predictors. Critically, it is not our intention in this article to fit a single stochastic model given by (1) and (2); rather, we look at each of the spatial predictors algorithmically, as it acts on the data $\{Z(\mathbf{s}_i) : i = 1,...,n\}$. We also consider the "central" spatial predictor in each case, recognizing that embellishments may be needed in a particular application. Our goal is to make the review and comparison as straightforward and transparent as possible.

2.1 Traditional Stationary Kriging (TSK)

Its parameterization: The statistical model from which TSK is an optimal spatial predictor can be defined hierarchically. The data model is given by (1) with $V(\cdot) \equiv 1$, and σ_ε^2 known. The process model is given by,

$$Y(\mathbf{u}) = \mathbf{x}(\mathbf{u})'\boldsymbol{\beta} + \nu(\mathbf{u}) + \xi(\mathbf{u});\ \ \mathbf{u} \in D, \tag{4}$$

where $\mathbf{x}(\mathbf{u})$ is a p-dimensional vector of known spatial covariates that describes the large-scale variation, $\nu(\mathbf{u})$ represents small-scale variation, and independently $\xi(\mathbf{u})$ represents fine-scale variation.

The spatial random process $\nu(\cdot)$ is specified to have mean zero and a second-order stationary covariance function,

$$\mathrm{cov}(\nu(\mathbf{u}+\mathbf{h}), \nu(\mathbf{u})) \equiv C(\mathbf{h});\ \ \mathbf{h} \in \mathbb{R}^d, \tag{5}$$

where the function $C(\cdot)$ is positive-definite (e.g., Cressie, 1993, p.68). Specifically, in Section 3, we use the exponential covariance function given by,

$$C(\mathbf{h}) = \sigma_0^2 \exp\left(-\frac{||\mathbf{h}||}{\theta}\right);\ \ \mathbf{h} \in \mathbb{R}^d, \tag{6}$$

where $\theta > 0$ and $\sigma_0^2 > 0$. We organize these unknown parameters into the set $\boldsymbol{\theta}^{\mathrm{TSK}} \equiv \{\boldsymbol{\beta}, \theta, \sigma_0^2, \sigma_\xi^2\}$.

The algorithm: To compute TSK for a given $\boldsymbol{\theta}^{\mathrm{TSK}}$, first construct the $n \times n$ covariance matrix,

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\mathrm{TSK}}) \equiv \left(\mathrm{cov}\left(\nu(\mathbf{s}_i), \nu(\mathbf{s}_j) | \theta, \sigma_0^2\right) : i,j = 1,...,n\right) + \sigma_\xi^2 \mathbf{I}_n + \sigma_\varepsilon^2 \mathbf{I}_n, \tag{7}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix. Also construct the n-dimensional vector,

$$\operatorname{cov}(\mathbf{Z}, Y(\mathbf{u})|\boldsymbol{\theta}^{\mathrm{TSK}}) = \operatorname{cov}(\mathbf{Z}, \nu(\mathbf{u})|\boldsymbol{\theta}^{\mathrm{TSK}}) + \sigma_\xi^2(I(\mathbf{u} = \mathbf{s}_1), ..., I(\mathbf{u} = \mathbf{s}_n))', \tag{8}$$

where $I(\cdot)$ represents the indicator function. Then define

$$\hat{Y}(\mathbf{u}, \mathbf{Z}|\boldsymbol{\theta}^{\mathrm{TSK}}) \equiv \mathbf{x}(\mathbf{u})'\boldsymbol{\beta} + \operatorname{cov}(\mathbf{Z}, Y(\mathbf{u})|\boldsymbol{\theta}^{\mathrm{TSK}})'\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\mathrm{TSK}})^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}), \tag{9}$$

where $\mathbf{X} \equiv (\mathbf{x}(\mathbf{s}_1), ..., \mathbf{x}(\mathbf{s}_n))'$.

Modifying (9) to be a function only of the data $\mathbf{Z}$, we substitute in the ordinary least squares (OLS) estimate for $\boldsymbol{\beta}$ and maximum likelihood (ML) estimates of the covariance parameters where the likelihood assumes mean zero, covariance (7), and it uses the detrended data (e.g., Cressie, 1993, p. 239 and pp. 291-292). The estimated parameters are denoted as $\hat{\boldsymbol{\theta}}^{\mathrm{TSK}}$. In Section 3, TSK is defined by the predictor,

$$\hat{Y}^{\mathrm{TSK}}(\mathbf{u}, \mathbf{Z}) \equiv \hat{Y}(\mathbf{u}, \mathbf{Z}|\hat{\boldsymbol{\theta}}^{\mathrm{TSK}}); \ \mathbf{u} \in D. \tag{10}$$

To compute $\hat{Y}^{\mathrm{TSK}}$, we use the R-package “geoR” version 1.7-4 (Ribeiro, Jr. and Diggle, 2012).

The motivation: The spatial predictor given by (9) minimizes the mean squared prediction error,

$$E\left((Y(\mathbf{u}) - \hat{Y}(\mathbf{u}, \mathbf{Z}))^2|\boldsymbol{\theta}^{\mathrm{TSK}}\right),$$

among the class of linear predictors, $\hat{Y}(\mathbf{u}, \mathbf{Z}) = \ell + \mathbf{k}'\mathbf{Z}$ (e.g., Cressie, 1993, Section 3.4.5).

2.2 Smoothing Splines (SSP)

Its parameterization: In our implementation of smoothing splines, there is a single parameter that trades off smoothness with goodness-of-fit, which we denote as $\theta^{\mathrm{SSP}} > 0$.

The algorithm: The smoothing spline predictor, for a given θ^{SSP}, is

$$\hat{\mathbf{Y}}(\mathbf{u},\mathbf{Z}|\theta^{\mathrm{SSP}}) \equiv \mathbf{x}(\mathbf{u})'\hat{\boldsymbol{\beta}}^{\mathrm{SSP}} + \mathbf{W}(\mathbf{u})'(\mathbf{W}+\theta^{\mathrm{SSP}}\mathbf{I}_n)^{-1}(\mathbf{Z}-\mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{SSP}}), \tag{11}$$

where $\mathbf{x}(\mathbf{u})$ is a p-dimensional vector of known spatial covariates, $\mathbf{X} \equiv (\mathbf{x}(\mathbf{s}_1),\ldots,\ \mathbf{x}(\mathbf{s}_n))'$ is an $n \times p$ matrix, and

$$\hat{\boldsymbol{\beta}}^{\mathrm{SSP}} \equiv (\mathbf{X}'(\mathbf{W}+\theta^{\mathrm{SSP}}\mathbf{I}_n)^{-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{W}+\theta^{\mathrm{SSP}}\mathbf{I}_n)^{-1}\mathbf{Z}.$$

In our implementation, the (i,j)-th entry of $\mathbf{W}$, say W_{ij}, is obtained from a radial basis function as follows,

$$||\mathbf{s}_i - \mathbf{s}_j||^2 \log\left(||\mathbf{s}_i - \mathbf{s}_j||\right), \tag{12}$$

and the n-dimensional vector $\mathbf{W}(\mathbf{u})$ has i-th entry $||\mathbf{u}-\mathbf{s}_i||^2\log(||\mathbf{u}-\mathbf{s}_i||)$ (e.g., Wahba, 1990, p. 31).

The value of θ^{SSP} is chosen based on minimizing a leave-one-out cross-validation error (Wahba, 1990, pp. 47 - 52). Denote this minimized value as $\hat{\theta}_{\mathrm{SSP}}$, and hence SSP is defined by the predictor,

$$\hat{Y}^{\mathrm{SSP}}(\mathbf{u},\mathbf{Z}) \equiv \hat{Y}(\mathbf{u},\mathbf{Z}|\hat{\theta}^{\mathrm{SSP}});\ \mathbf{u} \in D, \tag{13}$$

which is a function only of the data $\mathbf{Z}$. To compute $\hat{Y}^{\mathrm{SSP}}$, we use the Matlab (Version 8.0) function "griddata."

The motivation: The parameter θ^{SSP} is used to achieve a balance between goodness-of-fit and degree-of-smoothness of the spatial predictor (Wahba, 1990). In $\mathbb{R}^2$, the smoothing spline predictor is the function $f(\cdot)$ that minimizes the following penalized sum of squares (Wahba, 1990, p.31; Nychka, 2001),

$$\frac{1}{n}\sum_{i=1}^{n}(Z(\mathbf{s}_i)-f(\mathbf{s}_i))^2+\theta^{\text{SSP}}\int\int\left(\frac{\partial^2 f(\mathbf{u})}{\partial^2 u_1}+2\frac{\partial^2 f(\mathbf{u})}{\partial u_1 \partial u_2}+\frac{\partial^2 f(\mathbf{u})}{\partial^2 u_2}\right)du_1du_2, \tag{14}$$

for $\mathbf{u}=(u_1,u_2)'$. Its generalization to $\mathbb{R}^d$ for any positive integer d, is straightforward.

2.3 Negative-exponential-distance weighting (EDW)

Its parameterization: There is a single parameter used for controlling the weights in negative-exponential-distance weighting, which we denote as $\theta^{\text{EDW}} > 0$.

The algorithm: The data are weighted based on their Euclidean distance from the prediction location $\mathbf{u}$. Let $d_i(\mathbf{u}) \equiv ||\mathbf{u}-\mathbf{s}_i||$ be the Euclidean distance between $\mathbf{u}$ and $\mathbf{s}_i$. The negative-exponential-distance-weighting predictor, for a given θ^{EDW}, is

$$\hat{Y}(\mathbf{u},\mathbf{Z}|\theta^{\text{EDW}}) \equiv \frac{\sum_{i=1}^{n}\exp\{-\theta^{\text{EDW}}d_i(\mathbf{u})\}Z(\mathbf{s}_i)}{\sum_{i=1}^{n}\exp\{-\theta^{\text{EDW}}d_i(\mathbf{u})\}};\ \mathbf{u}\in D. \tag{15}$$

The value of θ^{EDW} is often prespecified in advance. In this article, we use $\theta^{\text{EDW}}=1$, although other choices are possible, resulting in more or less smoothness of the predicted surface. Then EDW is defined by the predictor,

$$\hat{Y}^{\text{EDW}}(\mathbf{u},\mathbf{Z}) \equiv \hat{Y}(\mathbf{u},\mathbf{Z}|1);\, \mathbf{u}\in D. \tag{16}$$

To compute $\hat{Y}^{\text{EDW}}$, we wrote a simple MATLAB script.

2.4 Fixed Rank Kriging (FRK)

Its parameterization: The statistical model from which FRK is derived as an optimal spatial predictor, can be defined hierarchically. The data model is given by (1) with both $V(\cdot)$ and σ_ε^2 known. The process model is,

$$Y(\mathbf{u}) = \mathbf{x}(\mathbf{u})'\boldsymbol{\beta} + \mathbf{S}_r^{\mathrm{BI}}(\mathbf{u})'\boldsymbol{\eta} + \xi(\mathbf{u});\ \ \mathbf{u} \in D, \tag{17}$$

where $\mathbf{x}(\mathbf{u})$ is a p-dimensional vector of known spatial covariates that describes the large-scale variation, $\mathbf{S}_r^{\mathrm{BI}}(\mathbf{u})'\boldsymbol{\eta}$ represents small-scale variation, and independently $\xi(\mathbf{u})$ represents fine-scale variation. The p-dimensional vector $\boldsymbol{\beta}$, the r-dimensional random vector $\boldsymbol{\eta}$, and the Gaussian white-noise process $\xi(\cdot)$ are all defined below (3). We organize the unknown parameters into the set $\boldsymbol{\theta}^{\mathrm{FRK}} \equiv \{\boldsymbol{\beta}, \mathbf{K}, \sigma_\xi^2\}$.

The term $\mathbf{S}_r^{\mathrm{BI}}(\cdot)$ is an r-dimensional vector function of bisquare basis functions (e.g., Cressie and Johannesson, 2008), and the value of r is specified to be much smaller than n. As will be discussed at the end of this section, specifying $r \ll n$ leads to computational advantages.

The algorithm: Define the $n \times n$ matrix $\mathbf{V}_\varepsilon \equiv \mathrm{diag}\{V(\mathbf{s}_1), ..., V(\mathbf{s}_n)\}$ and the $n \times r$ matrix $\mathbf{S}_r^{\mathrm{BI}} \equiv (\mathbf{S}_r^{\mathrm{BI}}(\mathbf{s}_1), ..., \mathbf{S}_r^{\mathrm{BI}}(\mathbf{s}_n))'$. To compute FRK, for a given $\boldsymbol{\theta}^{\mathrm{FRK}}$, first construct the $n \times n$ covariance matrix,

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\mathrm{FRK}}) \equiv \mathrm{cov}(\mathbf{Z}|\boldsymbol{\theta}^{\mathrm{FRK}}, \mathbf{S}_r^{\mathrm{BI}}) = \mathbf{S}_r^{\mathrm{BI}}\mathbf{K}(\mathbf{S}_r^{\mathrm{BI}})' + \sigma_\xi^2\mathbf{I}_n + \sigma_\varepsilon^2\mathbf{V}_\varepsilon,$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix. Also construct the n-dimensional vector,

$$\mathrm{cov}(\mathbf{Z}, Y(\mathbf{u})|\boldsymbol{\theta}^{\mathrm{FRK}}, \mathbf{S}_r^{\mathrm{BI}}) = \mathbf{S}_r^{\mathrm{BI}}\mathbf{K}\,\mathbf{S}_r^{\mathrm{BI}}(\mathbf{u}) + \sigma_\xi^2(I(\mathbf{u} = s_1), ..., I(\mathbf{u} = s_n))';\ \ \mathbf{u} \in D, \tag{18}$$

where recall that $I(\cdot)$ represents the indicator function. Then define

$$\hat{Y}(\mathbf{u},\mathbf{Z}|\boldsymbol{\theta}^{\mathrm{FRK}}) \equiv \mathbf{x}(\mathbf{u})'\boldsymbol{\beta} + \mathrm{cov}(\mathbf{Z},Y(\mathbf{u})|\boldsymbol{\theta}^{\mathrm{FRK}},\mathbf{S}_r^{\mathrm{BI}})'\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\mathrm{FRK}})^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta});\ \ \mathbf{u}\in D, \qquad (19)$$

where $\mathbf{X} \equiv (\mathbf{x}(\mathbf{s}_1),...,\mathbf{x}(\mathbf{s}_n))'$.

Modifying (19) to be a function only of the data $\mathbf{Z}$, we substitute in the OLS estimate for $\boldsymbol{\beta}$ and the expectation maximization (EM) estimates of the covariance parameters; here the likelihood from which the EM estimates are obtained assumes that the detrended data follow a Gaussian distribution with mean zero and covariance (18) (Katzfuss and Cressie, 2009). For a review of the EM algorithm in this setting, see Bradley et al. (2011). The estimated parameters are denoted as $\hat{\boldsymbol{\theta}}^{\mathrm{FRK}}$. Then FRK is defined by the predictor,

$$\hat{Y}^{\mathrm{FRK}}(\mathbf{u},\mathbf{Z}) \equiv \hat{Y}(\mathbf{u},\mathbf{Z}|\hat{\boldsymbol{\theta}}^{\mathrm{FRK}});\ \ \mathbf{u}\in D. \qquad (20)$$

To compute $\hat{Y}^{\mathrm{FRK}}$, we use Matlab code that is available on the website //niasra.uow.edu.au/cei/ webprojects/UOW175995.html.

The motivation: The spatial predictor given by (19) minimizes the mean squared prediction error,

$$E\left((Y(\mathbf{u})-\hat{Y}(\mathbf{u},\mathbf{Z}))^2|\boldsymbol{\theta}^{\mathrm{FRK}}\right),$$

among the class of linear predictors, $\hat{Y}(\mathbf{u},\mathbf{Z}) = \ell + \mathbf{k}'\mathbf{Z}$ (Cressie and Johannesson, 2008).

The primary motivation for FRK, as described in Cressie and Johannesson (2008), is that $\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\mathrm{FRK}})^{-1}$ can be computed efficiently using the Sherman-Morrison-Woodbury formula (e.g., Henderson and Searle, 1981):

$$\begin{aligned}\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\mathrm{FRK}})^{-1} &= (\sigma_\xi^2\mathbf{I}_n+\sigma_\varepsilon^2\mathbf{V}_\varepsilon)^{-1}-(\sigma_\xi^2\mathbf{I}_n+\sigma_\varepsilon^2\mathbf{V}_\varepsilon)^{-1}\mathbf{S}_r^{\mathrm{BI}}\\ &\quad\times\{\mathbf{K}^{-1}+(\mathbf{S}_r^{\mathrm{BI}})'(\sigma_\xi^2\mathbf{I}_n+\sigma_\varepsilon^2\mathbf{V}_\varepsilon)^{-1}\mathbf{S}_r^{\mathrm{BI}}\}^{-1}(\mathbf{S}_r^{\mathrm{BI}})'(\sigma_\xi^2\mathbf{I}_n+\sigma_\varepsilon^2\mathbf{V}_\varepsilon)^{-1}.\end{aligned}\tag{21}$$

Equation (21) allows efficient computation of $\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\mathrm{FRK}})^{-1}$ in (19), since (21) involves inverses of $r\times r$ matrices and a diagonal $n\times n$ matrix. Specifically, the computation involved with computing the right-hand side of (21) is of order nr^2, which is linear in n (Cressie and Johannesson, 2008).

2.5 Modified Predictive Process Approach (MPP)

Its parameterization: The statistical model from which MPP is derived as an optimal spatial predictor, can be defined hierarchically. The data model is given by (1) with $V(\cdot)\equiv 1$, and σ_ε^2 is unknown. The process model is given by,

$$Y(\mathbf{u})=\mathbf{x}(\mathbf{u})'\boldsymbol{\beta}+\mathbf{S}_r^{\mathrm{PP}}(\mathbf{u};\kappa,\sigma_\nu^2)'\boldsymbol{\eta}+\xi(\mathbf{u});\;\;\mathbf{u}\in D,\tag{22}$$

where $\mathbf{x}(\mathbf{u})$ is a p-dimensional vector of known spatial covariates, $\mathbf{S}_r^{\mathrm{PP}}(\mathbf{u};\kappa,\sigma_\nu^2)'\boldsymbol{\eta}$ represents small-scale variability, both κ and σ_ν^2 are unknown parameters, and independently $\xi(\mathbf{u})$ represents fine-scale variability. The p-dimensional vector $\boldsymbol{\beta}$, the r-dimensional random vector $\boldsymbol{\eta}$, and the Gaussian white-noise process $\xi(\cdot)$ are all defined below (3).

Let $\{\mathbf{u}_1^*,\ldots,\mathbf{u}_r^*\}\equiv D^*\subset D$ be a set of $(r\ll n)$ knots over the spatial domain D. The r-dimensional random vector $\boldsymbol{\eta}$ is taken to be Gaussian with mean zero and covariance matrix $\mathbf{K}^*$, where $\mathbf{K}^*\equiv\left\{C(\mathbf{u}_i^*,\mathbf{u}_j^*)\right\}$ and $C(\cdot)$ is the exponential covariance function with scaling parameter $\kappa>0$ and variance σ_ν^2. The term $\mathbf{S}_r^{\mathrm{PP}}(\cdot;\kappa,\sigma_\nu^2)$ is an r-dimensional vector function defined as,

$$\mathbf{S}_r^{\mathrm{PP}}(\mathbf{u};\kappa,\sigma_\nu^2)'\equiv\mathbf{k}(\mathbf{u})'\left(\mathbf{K}^*\right)^{-1},\tag{23}$$

where $\mathbf{k}(\mathbf{u}) \equiv (C(\mathbf{u},\mathbf{u}_i^*) : i = 1,...,r)'$ also depends on parameters κ and σ_ν^2.

The original predictive process approach, proposed by Banerjee et al. (2008), did not include $\xi(\cdot)$, and this leads to a variance of the hidden process that is underestimated. Later Finley et al. (2009) introduced the fine-scale variability term $\xi(\cdot)$ into the model, resulting in the modified predictive process approach. They model the spatial random process $\xi(\cdot)$ as a mean zero independent Gaussian process such that $\mathrm{var}(\xi(\mathbf{u})) = C(\mathbf{u},\mathbf{u}) - \mathbf{k}(\mathbf{u})'(\mathbf{K}^*)^{-1}\mathbf{k}(\mathbf{u})$. This leads to

$$\begin{aligned}\mathrm{var}(Y(\mathbf{u})) &= \mathrm{var}\left(\mathbf{S}_r^{\mathrm{PP}}(\mathbf{u};\kappa,\sigma_\nu^2)'\boldsymbol{\eta} + \xi(\mathbf{u})\right)\\ &= \mathbf{k}(\mathbf{u})'(\mathbf{K}^*)^{-1}\mathbf{k}(\mathbf{u}) + C(\mathbf{u},\mathbf{u}) - \mathbf{k}(\mathbf{u})'(\mathbf{K}^*)^{-1}\mathbf{k}(\mathbf{u}) = C(\mathbf{u},\mathbf{u}).\end{aligned}$$

Thus, the variance of the exponential covariance function is preserved. We organize the unknown parameters into the set $\boldsymbol{\theta}^{\mathrm{MPP}} \equiv \{\boldsymbol{\beta},\kappa,\sigma_\nu^2,\sigma_\varepsilon^2\}$.

The algorithm: Markov Chain Monte Carlo (MCMC) techniques are used for inference on parameters in this setting (Banerjee et al., 2008; Finley et al., 2009). The prior distributions are taken as $\sigma_\nu^2 \sim \mathrm{IG}(a_\eta,b_\eta)$, $\kappa \sim \mathrm{U}(a_\kappa,b_\kappa)$, $\sigma_\varepsilon^2 \sim \mathrm{IG}(a_\varepsilon,b_\varepsilon)$, and $\boldsymbol{\beta}$ has a flat prior, where σ_η^2, κ, σ_ε^2, and $\boldsymbol{\beta}$ are assumed mutually independent, $\mathrm{IG}(a,b)$ represents an inverted gamma distribution with parameters a and b, and $\mathrm{U}(a,b)$ represents a uniform distribution with parameters a and b. Choices for the hyperparameters depend on the application, but in Section 3 we use the suggestions from Finley et al. (2012), who also give details of the MCMC computations.

A difference between MPP and the other stochastic spatial predictors under consideration is that MPP predicts the process $Z(\cdot)$. Recall that the data model is given by,

$$Z(\mathbf{u}) = Y(\mathbf{u}) + \varepsilon(\mathbf{u});\ \ \mathbf{u} \in D, \tag{24}$$

and hence MPP predicts the process with the measurement error included. Consequently, MPP predictions will be exactly equal to the training data at training data locations $\{\mathbf{s}_i\}$, which is an undesirable property when $\sigma_\varepsilon^2 > 0$. Typically, scientific interest is in $Y(\cdot)$, and $\varepsilon(\cdot)$ in (24) should be filtered out.

The MCMC generates samples $\{Z(\mathbf{u})_1, ..., Z(\mathbf{u})_L\}$ from the posterior distribution of $Z(\mathbf{u})$. Then MPP is defined by the predictor,

$$\hat{Y}^{\text{MPP}}(\mathbf{u}, \mathbf{Z}) \equiv \frac{1}{L} \sum_{\ell=1}^{L} Z(\mathbf{u})_\ell; \;\; \mathbf{u} \in D. \tag{25}$$

To compute $\hat{Y}^{\text{MPP}}$, we use the R-package "spBayes" (Finley et al., 2012).

The motivation: The spatial predictor given by (25) minimizes the mean squared prediction error,

$$E(Z(\mathbf{u}) - \hat{Y}(\mathbf{u}, \mathbf{Z}))^2; \;\; \mathbf{u} \in D, \tag{26}$$

where here the expectation is taken over $\mathbf{Z}$, $Z(\mathbf{u})$, and $\boldsymbol{\theta}^{\text{MPP}}$. As we noted above, instead of $Y(\mathbf{u})$, the scientifically-less-interesting quantity $Z(\mathbf{u})$ appears in (26). The primary motivation of this approach is that since $r \ll n$ the Sherman-Woodbury-Morrison formula can be used to compute the precision matrix efficiently, and thus it should be scalable for large spatial datasets.

2.6 SPDE Approach (SPD)

Its parameterization: The statistical model from which SPD is derived as an optimal spatial predictor can be defined hierarchically. The data model is given by (1) with $V(\cdot) \equiv 1$, and σ_ε^2 unknown. The process model is given by,

$$Y(\mathbf{u}) = \mathbf{x}(\mathbf{u})'\boldsymbol{\beta} + \mathbf{S}_r^{\text{PL}}(\mathbf{u})'\boldsymbol{\eta}; \;\; \mathbf{u} \in D, \tag{27}$$

where $\mathbf{x}(\mathbf{u})$ is a p-dimensional vector of known spatial covariates that describes the large-scale variation, $\mathbf{S}_r^{\mathrm{PL}}(\mathbf{u})'\boldsymbol{\eta}$ represents small-scale variability, and the fine-scale variability term $\xi(\cdot) \equiv 0$. The p-dimensional vector $\boldsymbol{\beta}$ and the $r(> n)$-dimensional vector $\boldsymbol{\eta}$ are defined below (3). Here the term $\mathbf{S}_r^{\mathrm{PP}}(\cdot)$ is an r-dimensional vector function whose elements are piecewise-linear basis functions; and in contrast to FRK and MPP, $r > n$.

On the Euclidean space, define a set of r knots $\{\mathbf{u}_1^*, ..., \mathbf{u}_r^*\} \equiv D^*$, which contains the n locations of D_O; that is, $r > n$. The r-dimensional random vector $\boldsymbol{\eta}$ is specified to be a mean-zero Gaussian Markov random field defined on D^*. The precision matrix associated with $\boldsymbol{\eta}$ (i.e., $\mathbf{K}^{-1} \equiv \mathrm{cov}(\boldsymbol{\eta})^{-1}$) is based on parameters κ and σ_V^2. The functional form of this precision matrix, and hence the neighborhood structure of the elements in $\boldsymbol{\eta}$, is found by solving a stochastic partial differential equation, which we describe below. We organize the unknown parameters into the set $\boldsymbol{\theta}^{\mathrm{SPD}} \equiv \{\boldsymbol{\beta}, \mathbf{K}^{-1}, \sigma_\varepsilon^2\}$.

The algorithm: Bayesian inference proceeds without using MCMC; it is based on Integrated nested Laplacian approximations (INLA) in this setting (Rue et al., 2009; Lindgren et al., 2011). The INLA algorithm is derived from Laplace approximations of integrals of probability density functions.

First, priors are chosen for $\boldsymbol{\theta}^{\mathrm{SPD}}$. As a default in the R-INLA package, $\boldsymbol{\beta} \sim \mathrm{Gau}(\mathbf{0}, \tau_\beta^2 \mathbf{I})$, and $\log\left(1/\sigma_V^2\right)$, $\log\left(\sqrt{8}/\kappa\right)$, and $\log\left(\sigma_\varepsilon^2\right)$ are distributed as Log-Gamma. Further, $\boldsymbol{\beta}$, σ_V^2, κ, and σ_ε^2 are assumed to be mutually independent. The values of hyperparameters of the prior distribution are chosen heuristically (Rue, 2012, personal communication) based on default settings of the R-INLA package.

Denote the posterior probability density function of $Y(\mathbf{u})$ as $\pi(Y(\mathbf{u})|\mathbf{Z})$, and the INLA-approximated version is denoted as $\tilde{\pi}(Y(\mathbf{u})|\mathbf{Z})$ (e.g., Rue et al., 2009, Section 3). Rejection sampling is then used to generate L values $\{Y(\mathbf{u})_1, ..., Y(\mathbf{u})_L\}$ from $\tilde{\pi}(Y(\mathbf{u})|\mathbf{Z})$. Then SPD is defined by the predictor,

$$\hat{Y}^{\text{SPD}}(\mathbf{u},\mathbf{Z}) \equiv \frac{1}{L}\sum_{\ell=1}^{L} Y(\mathbf{u})_{\ell};\;\; \mathbf{u} \in D. \tag{28}$$

To compute $\hat{Y}^{\text{SPD}}$, we use the R-package "inla" (Rue et al., 2009; Rue, 2012).

The motivation: The spatial predictor given by (28) minimizes the (approximate) posterior mean squared prediction error,

$$\int (Y(\mathbf{u}) - \hat{Y}(\mathbf{u},\mathbf{Z}))^2 \bar{\pi}(Y(\mathbf{u})|\mathbf{Z}) dY(\mathbf{u}). \tag{29}$$

Computational efficiency is obtained through a connection between Gaussian Markov random fields and Gaussian processes that have a Matérn covariance function,

$$\frac{\sigma_\nu^2}{\Gamma(\alpha)2^{\alpha-1}}(\kappa||\mathbf{h}||)^{\alpha} K_{\alpha}(\kappa||\mathbf{h}||);\, \mathbf{h} \in \mathbb{R}^d, \tag{30}$$

where $K_\alpha(\cdot)$ is the modified Bessel function of the second kind of order $\alpha > 0$. Here, $0 < \alpha < \infty$ is a smoothing parameter, $\kappa > 0$ is a scaling parameter, and σ_ν^2 is the variance parameter.

A random process $\nu(\cdot)$ in $\mathbb{R}^d$ with covariance function given by (30) is a solution to the following stochastic partial differential equation (Whittle, 1963):

$$(\kappa^2 - \Delta)^{\zeta/2}\nu(\mathbf{u}) = W(\mathbf{u});\, \mathbf{u} \in \mathbb{R}^d, \tag{31}$$

where $W(\cdot)$ is a Gaussian white-noise process with mean zero and variance 1, and $\zeta \equiv \alpha + d/2$ is a positive integer, $\kappa > 0$, and $\sigma_\nu^2 > 0$. In (31), the Laplacian Δ is defined by,

$$\Delta \equiv \sum_{i=1}^{d} \frac{\partial^2}{\partial^2 u_i}. \tag{32}$$

The precision matrix associated with $\boldsymbol{\eta}$ (i.e., $\mathbf{K}^{-1} \equiv \text{cov}(\boldsymbol{\eta})^{-1}$) is specified to be a GMRF and is found by substituting $\nu(\mathbf{u}) = \sigma_\nu^2 \mathbf{S}_r^{\text{PL}}(\mathbf{u})'\boldsymbol{\eta}$ into Equation (31) and solving the stochastic partial differential equation. This solution, which is only for ζ a positive integer, can be found in Section 2.3 of Lindgren et al. (2011).

Lindgren et al. (2011) extend this modeling approach to handle nonstationarity by letting some of the parameters depend on spatial coordinates; they find the precision matrix associated with the random vector $\boldsymbol{\eta}$ that solves the following stochastic partial differential equation,

$$(\kappa^2(\mathbf{u}) - \Delta)^{\zeta/2}\{\sigma_\nu^2(\mathbf{u})\mathbf{S}_r^{\mathrm{PL}}(\mathbf{u})'\boldsymbol{\eta}\} = W(\mathbf{u}); \, \mathbf{u} \in \mathbb{R}^d, \tag{33}$$

where $\zeta \equiv \alpha + d/2$ is a positive integer, $\kappa(\mathbf{u}) > 0$, and $\sigma_\nu^2(\mathbf{u}) > 0$. Lindgren et al. (2011) propose the model,

$$\log\left(\sigma_\nu^2(\mathbf{u})\right) \equiv \sum_i \beta_i^{(1)} B_i^{(1)}(\mathbf{u}) \tag{34}$$

and

$$\log\left(\kappa^2(\mathbf{u})\right) \equiv \sum_i \beta_i^{(2)} B_i^{(2)}(\mathbf{u}), \tag{35}$$

where $\{B_i^{(1)}\}$ and $\{B_i^{(2)}\}$ represent two different finite sets of smooth basis functions.

Finally, in $\mathbb{R}^2$, $\mathbf{K}^{-1}$ is specified as follows: $\alpha = 1$ and hence $\zeta = 2$, since $d = 2$; $\{B_i^{(1)}\}$ is a set of four spherical basis functions of order three; and $\{B_i^{(2)}\}$ is a set of seven spherical basis function of order six (e.g., Lindgren et al., 2011).

2.7 Lattice Kriging (LTK)

Its parameterization: The statistical model defining lattice kriging can be defined hierarchically. The data model is given by (1) with $V(\cdot) \equiv 1$; and σ_ε^2 is assumed known. The process model is given by,

$$Y(\mathbf{u}) = \mathbf{x}(\mathbf{u})'\boldsymbol{\beta} + \mathbf{S}_r^{\mathrm{WL}}(\mathbf{u})'\boldsymbol{\eta}; \;\; \mathbf{u} \in D, \tag{36}$$

where $\mathbf{x}(\mathbf{u})$ is a p-dimensional vector of known spatial covariates, $\mathbf{S}_r^{\mathrm{WL}}(\mathbf{u})'\boldsymbol{\eta}$ represents small-scale variability, and the fine-scale variability term $\xi(\cdot) \equiv 0$. The p-dimensional vector $\boldsymbol{\beta}$ and the

$r(>n)$-dimensional vector $\boldsymbol{\eta}$ are defined below (3). Here the term $\mathbf{S}_r^{\mathrm{WL}}(\cdot)$ is an r-dimensional vector function whose elements are "smooth" Wendland basis functions; notice that $r > n$.

From Nychka et al. (2014), define a set of r knots $\{\mathbf{u}_1^*, ..., \mathbf{u}_r^*\} \equiv D^*$ on a regular grid contained in D. Then define the r-dimensional random vector $\boldsymbol{\eta} \equiv \mathbf{B}^{-1}\mathbf{e}$, where $\mathbf{e}$ is an r-dimensional Gaussian random vector with mean zero and variance $\sigma_\eta^2 \mathbf{I}_r$. Note that $\mathbf{B}\boldsymbol{\eta} = \mathbf{e}$, which is the form of a simultaneous autoregressive (SAR) model, and

$$\mathbf{B} \equiv \begin{pmatrix} 4+\kappa^2 & -1 & 0 & \cdots & 0 \\ -1 & 4+\kappa^2 & -1 & \cdots & 0 \\ 0 & -1 & & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ \vdots & & & -1 & 0 \\ \vdots & & -1 & 4+\kappa^2 & -1 \\ 0 & \cdots & & -1 & 4+\kappa^2 \end{pmatrix},$$

for $\kappa \geq 0$. The elements of $\boldsymbol{\eta}$ are arbitrarily ordered based on the locations of the knots. We organize the unknown parameters into the set $\boldsymbol{\theta}^{\mathrm{LTK}} \equiv \{\boldsymbol{\beta}, \sigma_\eta^2, \kappa\}$.

The algorithm: Define $\mathbf{S}_r^{\mathrm{WL}} \equiv (\mathbf{S}^{\mathrm{WL}}(\mathbf{s}_1), ..., \mathbf{S}^{\mathrm{WL}}(\mathbf{s}_n))'$. To compute LTK, for a given $\boldsymbol{\theta}^{\mathrm{LTK}}$, first construct the $n \times n$ covariance matrix,

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\mathrm{LTK}}) \equiv \mathrm{cov}(\mathbf{Z}|\mathbf{K}, \mathbf{S}_r^{\mathrm{WL}}) = \mathbf{S}_r^{\mathrm{WL}}\mathbf{K}(\mathbf{S}_r^{\mathrm{WL}})' + \sigma_\varepsilon^2 \mathbf{I}_n, \tag{37}$$

where recall that $\mathbf{I}_n$ is the $n \times n$ identity matrix and $\mathbf{K} \equiv \mathrm{cov}(\boldsymbol{\eta})$. Also construct the n-dimensional vector,

$$\mathrm{cov}(\mathbf{Z}, Y(\mathbf{u})|\boldsymbol{\theta}^{\mathrm{LTK}}, \mathbf{S}_r^{\mathrm{WL}}) = \mathbf{S}_r^{\mathrm{WL}}\mathbf{K}\,\mathbf{S}_r^{\mathrm{WL}}(\mathbf{u}).$$

Then define,

$$\hat{Y}(\mathbf{u},\mathbf{Z}|\boldsymbol{\theta}^{\text{LTK}}) \equiv \mathbf{x}(\mathbf{u})'\boldsymbol{\beta} + \text{cov}(\mathbf{Z}, Y(\mathbf{u})|\boldsymbol{\theta}^{\text{LTK}}, \mathbf{S}_r^{\text{WL}})'\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\text{LTK}})^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\beta});\;\; \mathbf{u}\in D, \qquad (38)$$

where $\mathbf{X} \equiv (\mathbf{x}(\mathbf{s}_1),...,\mathbf{x}(\mathbf{s}_n))'$.

Modifying (38) to be a function only of the data $\mathbf{Z}$, we substitute in the maximum likelihood estimate of $\boldsymbol{\theta}^{\text{LTK}}$ (denoted $\hat{\boldsymbol{\theta}}^{\text{LTK}}$). Then LTK is defined by the predictor,

$$\hat{Y}^{\text{LTK}}(\mathbf{u},\mathbf{Z}) \equiv \hat{Y}(\mathbf{u},\mathbf{Z}|\hat{\boldsymbol{\theta}}^{\text{LTK}});\;\; \mathbf{u}\in D. \qquad (39)$$

To compute $\hat{Y}^{\text{LTK}}$, we use the R package “LatticeKrig” (Nychka et al., 2014).

The motivation: The spatial predictor given by (38) minimizes the mean squared prediction error,

$$E\left((Y(\mathbf{u}) - \hat{Y}(\mathbf{u},\mathbf{Z}))^2|\boldsymbol{\theta}^{\text{LTK}}\right), \qquad (40)$$

among the class of linear predictors, $\hat{Y}(\mathbf{u},\mathbf{Z}) = \ell + \mathbf{k}'\mathbf{Z}$. A numerical motivation for LTK is that $\boldsymbol{\Sigma}(\boldsymbol{\theta}^{\text{LTK}})^{-1}$ can be found using sparse-matrix techniques (Nychka et al., 2014).

3 A Comparison of the Seven Spatial Predictors: Mid-Tropospheric CO_2 Measurements

The Aqua satellite is part of the Earth Observing System (EOS), which is administered by the National Aeronautics and Space Administration (NASA). The Atmospheric Infrard Sounder (AIRS) is an instrument on board the Aqua satellite that retrieves information on atmospheric CO_2. Specifically, the AIRS instrument collects measurements in the form of spectra that are then converted to mid-tropospheric CO_2 values in parts per million (ppm) (Chahine et al., 2006). This information on global CO_2 has been used to great effect in raising public awareness on greenhouse gases and in determining policy regarding climate change (e.g., see https://www.ipcc.ch/).

The AIRS instrument records data over swaths (or paths) of the Earth's surface (roughly 800 km wide) and extends from -60° to 90° latitude. Data are collected on a daily cycle from 1:30 pm to 1:30 am. We use AIRS data recorded from February 1 through February 9, 2010. The collected data are then reported at different spatial resolutions. In this article, we analyze AIRS's level-2 CO_2 data, which is reported at a 17.6 km by 17.6 km spatial resolution.

The resulting AIRS CO_2 dataset consists of 74,361 total observations. We would like to compare both the predictive performance and the computational performance of each spatial predictor. However, not every predictor can be computed using all 74,361 observations. For example, it is well known that the traditional predictors TSK and SSP cannot handle datasets this large (or larger). Hence, we subset the globe (i.e., D) into a study region that contains a smaller number of data points than found in $\{Z(\mathbf{s}) : \mathbf{s} \in D_O\}$.

In Figure 1, we display Study Region 1, which covers the Midwest US. Here, there is a total of just 71 observations available, which we separate into two subsets of size $n = 57$ and $m = 14$. The n observations are the "training" data (top panel of Figure 1) used to fit the spatial predictors, and the m observations are the "validation" data used to assess the predictive performance of each spatial predictor (bottom panel of Figure 1); notice that we reserve roughly 20% of the data for validation. Our main reason for analyzing this small study region is to compare the predictive performance of *all* seven spatial predictors, which we do in Section 3.1.

Although we are interested in comparing all the predictors, a number of them are designed to handle larger datasets. In particular, EDW, FRK, MPP, SPD, and LTK are relatively straightforward (but non-trivial) predictors that are intended for large spatial datasets. Consider Study Region 2 in Figure 2, which covers the Americas and western Sahara between longitudes -125° to 3° and latitudes -20° to 44° (this is the same study region used in Cressie et al., 2010b). There are $n = 12,358$ observations used to train each spatial predictor (top panel of Figure 2), and $m = 3,090$

observations used for validation (bottom panel of Figure 2). In Section 3.2, we use the data in Figure 2 to compare these five spatial predictors.

Finally, in Section 3.3, we use the entire dataset in Figure 3, which is computationally feasible only for EDW, FRK, SPD, and LTK, but no longer for MPP. There are $n = 59,488$ observations used to train each spatial predictor (top panel of Figure 3), and $m = 14,873$ observations used for validation (bottom panel of Figure 3). This is by no means an unusually large dataset (with spatially correlated observations) that one might process; for example, Sengupta et al. (2012) and Bradley et al. (2014b) process datasets on the order of millions.

The training (validation) data are referenced by their locations, $D^{\text{trn}} \equiv \{\mathbf{s}_j^{\text{trn}} : j = 1,...,n\}$ ($D^{\text{val}} \equiv \{\mathbf{s}_j^{\text{val}} : j = 1,...,m\}$), where $D_O = D^{\text{trn}} \cup D^{\text{val}}$ and $D^{\text{trn}} \cap D^{\text{val}} = \emptyset$. Hence, the total size of the dataset is $n+m$. We use the validation datasets to assess the predictive performance of each spatial predictor. Define the root average squared testing error (RSTE) associated with the predictor $\hat{Y}^{\text{PRD}}$ as,

$$\text{RSTE}(\hat{Y}^{\text{PRD}}, m) \equiv \left(\frac{1}{m}\sum_{j=1}^{m}(Z(\mathbf{s}_j^{\text{val}}) - \hat{Y}^{\text{PRD}}(\mathbf{s}_j^{\text{val}}, \mathbf{Z}))^2\right)^{1/2}, \tag{41}$$

where “PRD” notates a generic predictor. The RSTE will be used to compare each spatial predictor (small values are desirable), PRD = TSK, SSP, EDW, FRK, MPP, SPD, and LTK.

Another criterion that we consider is the predictive model choice criterion (PMCC) from Gneiting and Raftery (2007, see Equation (27)),

$$\text{PMCC}(\hat{Y}^{\text{PRD}}, m) \equiv \frac{1}{m}\sum_{j=1}^{m}\frac{(Z(\mathbf{s}_j^{\text{val}}) - \hat{Y}^{\text{PRD}}(\mathbf{s}_j^{\text{val}}, \mathbf{Z}))^2}{\hat{\sigma}^{\text{PRD}}(\mathbf{s}_j^{\text{val}}, \mathbf{Z})^2} - \log\left(\hat{\sigma}^{\text{PRD}}(\mathbf{s}_j^{\text{val}}, \mathbf{Z})^2\right), \tag{42}$$

where $\hat{\sigma}^{\text{PRD}}(\,\cdot\,,\,\cdot\,)^2$ is the model-based posterior variance, and hence we can only compute the PMCC for PRD = TSK, FRK, MPP, SPD, and LTK. Notice that for the SME model in (1) and

PRD = TSK, FRK, and LTK,

$$\hat{\sigma}^{\text{PRD}}(\mathbf{s},\mathbf{Z})^2$$
$$= \text{var}(\nu(\mathbf{s})|\hat{\boldsymbol{\theta}}^{\text{PRD}}) + \text{var}(\xi(\mathbf{s})|\hat{\boldsymbol{\theta}}^{\text{PRD}}) - \text{cov}(\mathbf{Z},Y(\mathbf{s})|\hat{\boldsymbol{\theta}}^{\text{PRD}})'\text{cov}(\mathbf{Z}|\hat{\boldsymbol{\theta}}^{\text{PRD}})^{-1}\text{cov}(\mathbf{Z},Y(\mathbf{s})|\hat{\boldsymbol{\theta}}^{\text{PRD}}). \tag{43}$$

The posterior variance for the predictors that are derived using a fully Bayesian approach are estimated by

$$\hat{\sigma}^{\text{PRD}}(\mathbf{s},\mathbf{Z})^2 = \begin{cases} \text{var}\,(Z(\mathbf{s})_\ell : \ell = 1,...,L) & \text{if PRD} = \text{MPP}, \\ \text{var}\,(Y(\mathbf{s})_\ell : \ell = 1,...,L) & \text{if PRD} = \text{SPD};\ \ \mathbf{s} \in D^{\text{val}}, \end{cases}$$

where recall $\{Z(\cdot)_\ell\}$ and $\{Y(\cdot)_\ell\}$ are samples from their respective posterior distributions defined in Sections 2.5 and 2.6, respectively. The PMCC is useful for comparing predictors (small values are desirable) because it incorporates information on the implicit model-based prediction errors. However, it has the limitation of not allowing a comparison to deterministic predictors, which can be done with RSTE.

We are interested in evaluating other properties of the predictors in addition to its predictive performance. In particular, to assess the amount of smoothness in PRD, consider the lag-1 semivariogram,

$$\frac{1}{2|C(1)|}\sum_{C(1)}(\hat{Y}^{\text{PRD}}(\mathbf{u}_i) - \hat{Y}^{\text{PRD}}(\mathbf{u}_j))^2, \tag{44}$$

where PRD = TSK, SSP, EDW, FRK, MPP, SPD, and LTK, $C(h) \equiv \{(i,j) : ||\mathbf{u}_i - \mathbf{u}_j|| = h\}$, $|C(h)|$ denotes the number of distinct elements in the set $C(h)$, h denotes the spatial lag, and $h = 1$ is in a unit of distance defined by the smallest lag at which a semivariogram can be computed. In Study Regions 1, 2, and 3 the unit of distance is $1.41°$, $1.5°$, and $1.5°$, respectively. A large (small) lag-1 semivariogram in (44) suggests that the map of PRD is non-smooth (smooth).

The exact specifications of each of the seven spatial predictors can be found in Section 2.

Here the covariates are $\mathbf{x}((\text{latitude}, \text{longitude})') \equiv (1, \text{latitude})$, since it is well known that mid-tropospheric CO_2 values display a latitudinal gradient (Hammerling et al., 2012); that is, there are $p = 2$ covariates. Additionally, the measurement-error variances are assumed known for TSK, FRK, and LTK; in practice, these variances are estimated using a variogram-extrapolation technique used by Cressie et al. (2010b) and Katzfuss and Cressie (2012). We use their estimate of σ_ε^2 = 5.6062 ppm^2 and, for simplicity, we shall take $V(\cdot) \equiv 1$.

In Sections 3.1 through 3.3, all of our computations are performed on a Dell Optiplex 7010 Desktop Computer with a quad-Core 3.40 GHz processor and 8 Gbytes of memory. It is important to note that the timing and memory-usage results may be different for different machines; however, to illustrate what someone might expect in practice, we use a computer that has the specification of a "typical personal desktop."

3.1 Comparison using a Small Dataset of Mid-Tropospheric CO_2

In this section, we use the data in Study Region 1 displayed in Figure 1, which we process using all seven spatial predictors, namely $\hat{Y}^{\text{TSK}}$, $\hat{Y}^{\text{SSP}}$, $\hat{Y}^{\text{EDW}}$, $\hat{Y}^{\text{FRK}}$, $\hat{Y}^{\text{MPP}}$, $\hat{Y}^{\text{SPD}}$, and $\hat{Y}^{\text{LTK}}$. Maps of the seven spatial predictors are given in Figure 4.

Each spatial predictor displays similar general patterns, with lower CO_2 values near the Great Lakes. In general, we can separate the predictors in Figure 4 into two categories: smooth and non-smooth. The two deterministic predictors (SSP and EDW) appear non-smooth, whereas the stochastic spatial predictors appear quite smooth; this is also seen in the lag-1 semivariograms in Table 1. This may be because the stochastic predictors can rely on an underlying stationary process in this setting, where the dataset is small and fairly sparse over the prediction region D.

The RSTE results for this example (given in Table 1) indicate that FRK is the individual predictor that appears to have the highest predictive performance, while LTK has the least-favorable predictive performance among the seven spatial predictors; however, it should be noted that the

RSTE values are fairly similar across different choices of PRD. The PMCC results for this example (given in Table 1) indicate that TSK, MPP, and FRK appear to have the highest predictive performance, while SPD and LTK have the least-favorable predictive performance among the five stochastic spatial predictors. As expected, there are no difficulties with CPU time and memory usage for this small dataset, and each of the seven spatial predictors were computed in a matter of seconds.

3.2 Comparison using a Large Dataset of Mid-Tropospheric CO_2

It is well known that the inversion of a large $n \times n$ matrix makes TSK and SSP computationally impractical. Hence, for this large dataset in Study Region 2 (see Figure 2) we consider the five spatial predictors that can be computed, namely EDW, FRK, MPP, LTK, and SPD.

Maps of the five spatial predictors $\hat{Y}^{\mathrm{EDW}}$, $\hat{Y}^{\mathrm{FRK}}$, $\hat{Y}^{\mathrm{MPP}}$, $\hat{Y}^{\mathrm{SPD}}$, and $\hat{Y}^{\mathrm{LTK}}$ are given in Figure 5. Each spatial predictor displays similar general patterns; however, in contrast to the results in Section 3.1, the large dataset used in this section shows clearly that MPP is the smoothest predictor, EDW is the least smooth, and FRK, LTK, and SPD have similar patterns of smoothness. These results are further corroborated by inspecting the lag-1 semivariograms in Table 2.

The RSTE results for this example (see Table 2) are fairly constant across different choices of PRD, with MPP (EDW) having the highest (least-favorable) predictive performance as measured by RSTE; recall that MPP is the smoothest spatial predictor. Similar conclusions can be made from the PMCC in Table 2, which indicates that the reduced-rank prediction methods appear to have the highest predictive performances, whereas the full-rank prediction methods appear to have less-favorable predictive performances. The CPU time and memory usage are manageable except for MPP, which has a CPU time of approximately 3.5 hours.

3.3 Comparison using a Very Large Dataset of Mid-Tropospheric CO_2

In this section, we use the data in Study Region 3 (the entire dataset) displayed in Figure 3, and the four spatial predictors that can process a dataset of this size; that is, we compare EDW, FRK, SPD, and LTK. Note that the MPP predictor is computed using a Metropolis-within-Gibbs sampler, making it too computationally intensive for very large spatial datasets. Coincidentally, the four spatial predictors that can handle datasets of this size do not use MCMC algorithms for statistical inference. Specifically, FRK and LTK are empirical Bayesian, SPD uses a fully Bayesian approach based on Rue et al. (2009)'s INLA algorithm, and EDW does not use a statistical model for inference.

Maps of the four spatial predictors, $\hat{Y}^{\mathrm{EDW}}$, $\hat{Y}^{\mathrm{FRK}}$, $\hat{Y}^{\mathrm{SPD}}$, and $\hat{Y}^{\mathrm{LTK}}$ are given in Figure 6. Similar to the results in Sections 3.1 and 3.2, each prediction method displays very similar patterns. The lag-1 semivariograms indicate that SPD is now the least smooth among the four predictors; LTK retains its property of being much smoother than FRK, EDW, and SPD.

The RSTE results for this example (see Table 3) are fairly constant across different choices of PRD (similar to the results in Sections 3.1 and 3.2), with FRK (EDW) having the highest (least-favorable) predictive performance as measured by RSTE. As in Sections 3.1 and 3.2, PMCC indicates that the reduced-rank method FRK has higher predictive performance than the full-rank methods, SPD and LTK. The CPU time for both FRK and SPD indicate that both of these methods are highly computationally efficient for spatial prediction. Moreover, the memory usage for each predictor is modest. However, EDW and LTK require a significant wait-time to obtain spatial predictions (around 1.5 hours).

4 Discussion

In this article, we present a comparison of spatial predictors from an algorithmic viewpoint. In particular, we systematically layout the parameterization, the algorithm, and the motivation of three traditional methods of spatial prediction and four more-recently-introduced spatial predictors. The traditional spatial predictors include: traditional stationary kriging (TSK), smoothing splines (SSP), and negative-exponential distance-weighting (EDW). The more-recently-introduced spatial predictors include: Fixed Rank Kriging (FRK), a modified predictive processes approach (MPP), a stochastic partial differential equation approach (SPD), and lattice kriging (LTK). Additionally, we use a benchmark of small, large, and very large mid-tropospheric CO_2 datasets to compare computation time, memory-usage, and the prediction performance of each spatial predictor.

Recent advances in technology, such as remote sensing, have made large-to-massive spatial datasets more available, making spatial prediction using big datasets an important and growing problem in the statistics literature. Consequently, the algorithmic concerns of CPU time and memory-usage are featured in our comparison along with predictive performance.

Of the seven predictors we consider, FRK and SPD perform extremely well in terms of CPU time and memory-usage. However, the remaining five spatial predictors are not as efficient. Both EDW and LTK were scalable to the very large benchmark dataset, but the wait-time was rather large (approximately 1.5 hours for each). It is well known that TSK and SSP have very poor CPU time and memory-usage properties for large datasets and, hence, we were only able to use these predictors using the small benchmark dataset. The MPP predictor also has limitations in CPU time; consequently, we were only able to use MPP on the small and large benchmark datasets, the latter dataset resulting in a significant wait-time (around 3.5 hours).

When visually comparing each of the seven spatial predictors, we see that they each display roughly the same general pattern. From an algorithmic point-of-view, this is to be expected, since if the signal-to-noise ratio is “large enough,” then any local-averaging scheme should be able to pick large-scale patterns. These visual patterns are further corroborated using the lag-1 semivariogram, which is consistently smaller (larger) for MPP (EDW and SPD). Of the three predictors for which PMCC can be computed for all study regions, FRK has much higher predictive performance than SPD and LTK. The other predictor that can be computed for all three study regions, EDW, had least-favorable predictive performance (among FRK, SPD, LTK and EDW) according to the RSTE criterion.

Section 3 has filled a need for empirical comparisons between reduced-rank and full-rank spatial predictors; the results shed light on the recent criticisms of reduced-rank statistical modeling (Lindgren et al., 2011; Stein, 2014) despite the fact that it has been shown to do well in other settings (see, e.g., Wikle and Cressie, 1999; Cressie and Johannesson, 2006; Shi and Cressie, 2007; Banerjee et al., 2008; Cressie and Johannesson, 2008; Finley et al., 2009; Katzfuss and Cressie, 2011; Cressie et al., 2010a,b; Kang and Cressie, 2011; Katzfuss and Cressie, 2012). In terms of predictive performance as measured by RSTE and PMCC, our results on a benchmark dataset of CO_2 data from NASA's AIRS instrument showed that reduced-rank methods outperform the viable full-rank alternatives.

Acknowledgments

This research was partially supported by the U.S. National Science Foundation (NSF) and the U.S. Census Bureau under NSF grant SES−1132031, funded through the NSF-Census Research Network (NCRN) program. It was also partially supported by NASA's Earth Science Technology Office through its Advanced Information Systems Technology program Grant NNH08ZDA001N. This research is partially supported by NSF grant DMS-1007060 and DMS-1308458.

References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. London, UK: Chapman and Hall.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). "Gaussian predictive process models for large spatial data sets." *Journal of the Royal Statistical Society Series B*, 70, 825–848.

Bradley, J., Cressie, N., and Shi, T. (2014a). "Comparing and Selecting Spatial Predictors Using Local Criteria." *TEST*, forthcoming.

Bradley, J., Holan, S., and Wikle, C. (2014b). "Mixed effects modeling for areal data that exhibit multivariate-spatio-temporal dependencies." *arXiv preprint arXiv: 1407.7479*.

Bradley, J. R., Cressie, N., and Shi, T. (2011). "Selection of rank and basis functions in the Spatial Random Effects model." In *Proceedings of the 2011 Joint Statistical Meetings*, 3393–3406. Alexandria, VA: American Statistical Association.

Campbell, J. (2010). "Improving lead generation success through integrated methods: transcending 'drug discovery by numbers'." *IDrugs: the Investigational Drugs Journal*, 21, 62 – 71.

Chahine, M., Pagano, T., Aumann, H., Atlas, R., Barnet, C., Blaisdell, J., Chen, L., Divakarla, M., Fetzer, E., Goldberg, M., Gautier, C., Granger, S., Hannon, S., Irion, F. W., Kakar, R., Kalnay, E., Lambrigtsen, B., Lee, S., Marshall, J. L., McMillian, W. W., McMillin, L., Olsen, E. T., Revercomb, H., Rosenkranz, P., Smith, W. L., Staelin, D., Strow, L. L., Susskind, J., Tobin, D., Wolf, W., and Zhou, L. (2006). "AIRS: Improving weather forecasting and providing new data on greenhouse gases." *Bulletin of the American Meteorological Society*, 87, 911–926.

Cressie, N. (1990). "The origins of kriging." *Mathematical Geology*, 22, 239–252.

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

Cressie, N. and Johannesson, G. (2006). "Spatial prediction for massive data sets." In *Australian Academy of Science Elizabeth and Frederick White Conference*, 1–11. Australian Academy of Science, Canberra.

— (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. (2010a). " Fixed Rank Filtering for spatio-temporal data." *Journal of Computational and Graphical Statistics*, 19, 724–745.

— (2010b). " Using temporal variability to improve spatial mapping with application to satellite data." *Canadian Journal of Statistics*, 38, 271–289.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Hoboken, NJ: Wiley.

Finley, A. O., Banerjee, S., and Carlin, B. (2012). "Package 'spBayes'." http://cran.r-project.org/web/packages/spBayes/spBayes.pdf. Retrieved January, 2013.

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009). "Improving the performance of predictive process modeling for large datasets." *Computational Statistics and Data Analysis*, 53, 2873–2884.

Gneiting, T. and Raftery, A. (2007). "Strictly proper scoring rules, prediction, and estimation." *Journal of the American Statistical Association*, 102, 359–378.

Hammerling, D. M., Michalak, A. M., and Kawa, S. R. (2012). "Mapping of CO_2 at high spatiotemporal resolution using satellite observations: Global distributions from OCO-2." *Journal of Geophysical Research*, 117, 1–10.

Henderson, H. V. and Searle, S. R. (1981). "On deriving the inverse of a sum of matrices." *SIAM Review*, 23, 53–60.

Hormozi, A. and Giles, S. (2004). "Data mining: A competitive weapon for banking and retail industries." *Information Systems Management*, 21, 62 – 71.

Kang, E. L. and Cressie, N. (2011). "Bayesian inference for the Spatial Random Effects model." *Journal of the American Statistical Association*, 106, 972 – 983.

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*, 3378–3390. Alexandria, VA: American Statistical Association.

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

— (2012). "Bayesian hierarchical spatio-temporal smoothing for very large datasets." *Environmetrics*, 23, 94–107.

Lindgren, F., Rue, H., and Lindström, J. (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach." *Journal of the Royal Statistical Society, Series B*, 73, 423–498.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2014). "A multiresolution Gaussian process model for the analysis of large spatial data sets." *Journal of Computational and Graphical Statistics*, DOI: 10.1080/10618600.2014.914946.

Nychka, D. W. (2001). "Spatial process estimates as smoothers." In *Smoothing and Regression: Approaches, Computation and Applications,* rev. edn, ed. M. G. Schmiek, 393–424. New York, NY: Wiley.

Ribeiro, Jr., P. J. and Diggle, P. J. (2012). "Package 'geoR'." http://cran.r-project.org/web/packages/geoR/geoR.pdf. Retrieved November, 2012.

Rue, H. (2012). "The R-INLA Project." http://www.r-inla.org/. Retrieved November, 2012.

Rue, H., Martino, S., and Chopin, N. (2009). "Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations." *Journal of the Royal Statistical Society, Series B*, 71, 319–392.

Schabenberger, O. and Gotway, C. (2005). *Statistical Methods for Spatial Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC Press.

Sengupta, A., Cressie, N., Frey, R., and Kahn, B. (2012). "Statistical modeling of MODIS cloud data using the Spatial Random Effects model." In *Proceedings of the Joint Statistical Meetings*, 3111–3123. Alexandria, VA: American Statistical Association.

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.

Stein, M. (2014). "Limitations on low rank approximations for covariance matrices of spatial data." *Spatial Statistics*, 8, 1–19.

Sun, Y., Li, B., and Genton, M. (2012). "Geostatistics for large datasets." In *Space-Time Processes and Challenges Related to Environmental Problems*, eds. E. Porcu, J. M. Montero, and M. Schlather, 55–77. Berlin: Springer.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

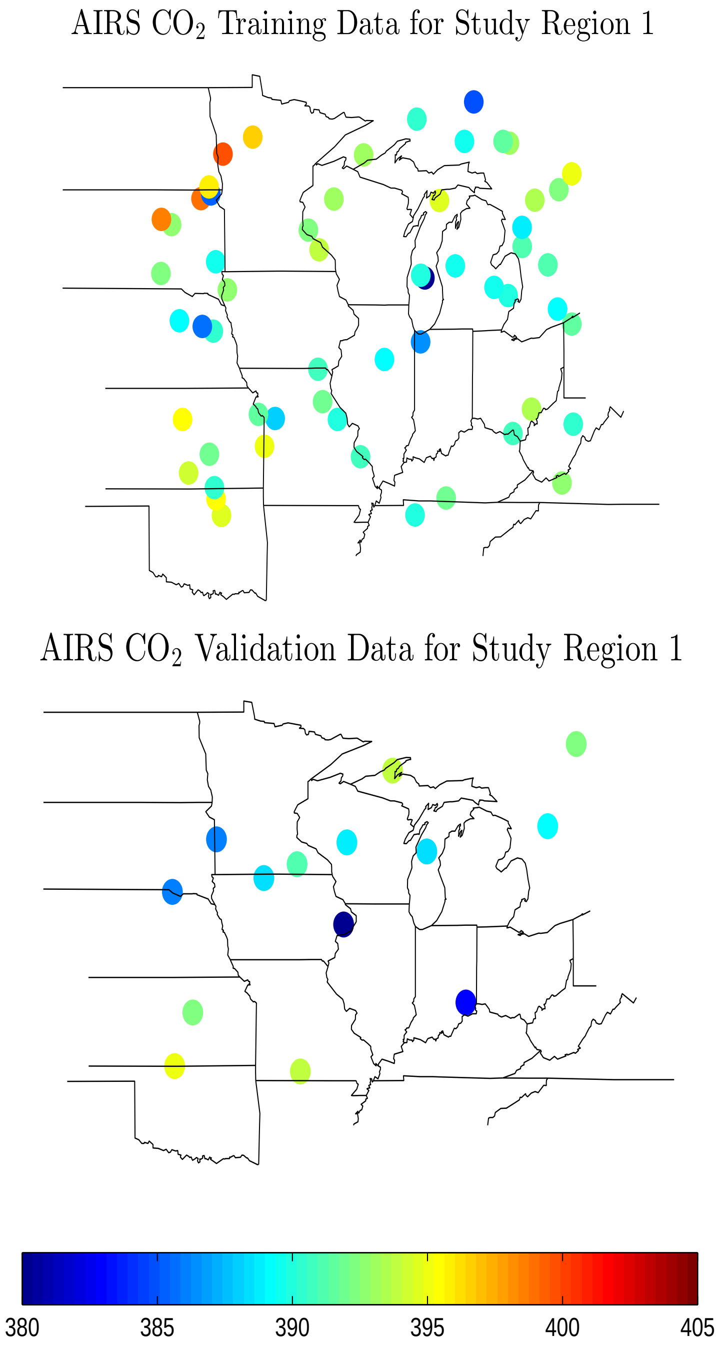
Whittle, P. (1963). "Stochastic processes in several dimensions." *Bulletin of the International Statistical Institute*, 40, 974–994.

Wikle, C. and Cressie, N. (1999). "A dimension-reduced approach to space-time Kalman filtering." *Biometrika*, 86, 815–829.

Wikle, C., Milliff, R., Nychka, D., and Berliner, L. (2001). “Spatiotemporal hierarchical Bayesian modeling tropical ocean surface winds.” *Journal of the American Statistical Association (Theory and Methods)*, 96, 382–397.

Xin, L., Guodong, C., and Ling, L. (2000). “Comparison of spatial interpolation methods.” *Advance in Earth Sciences*, 15, 260–265.

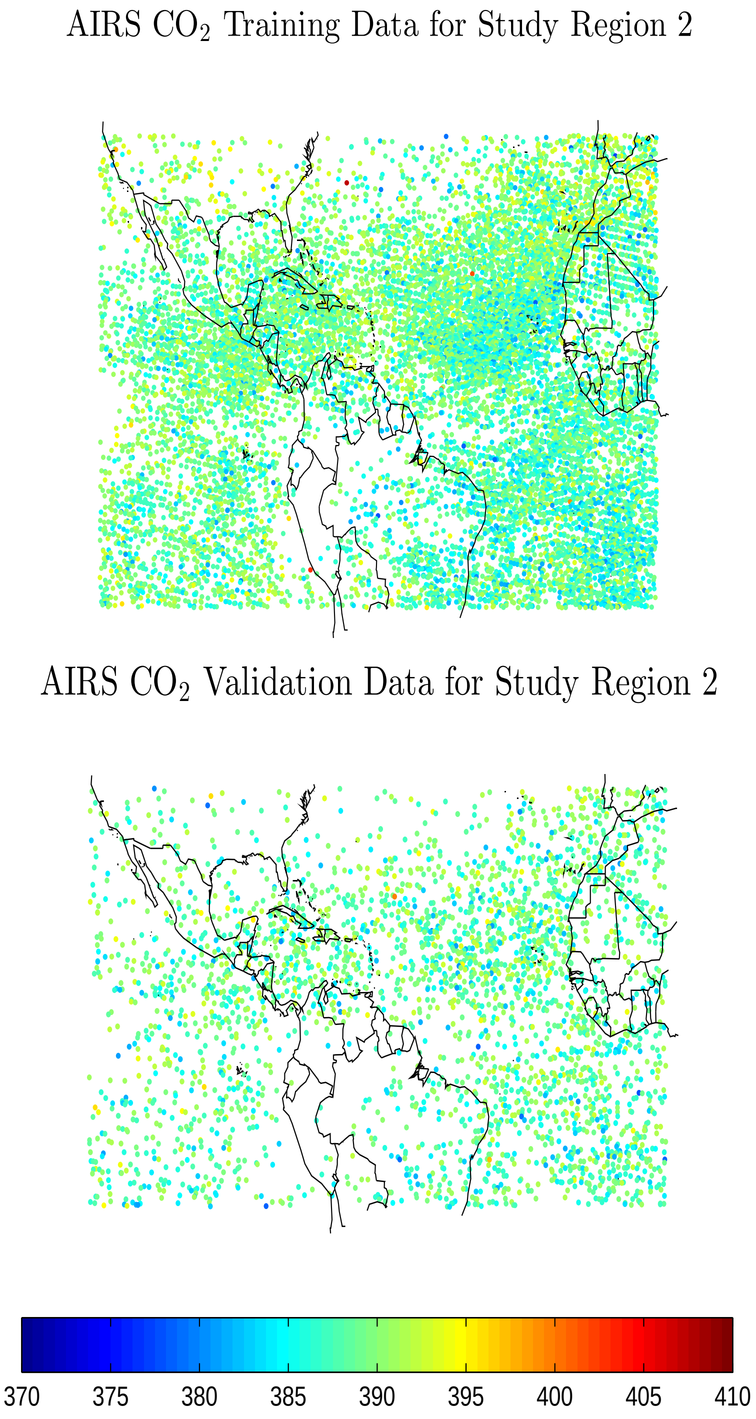
Figures and Tables



AIRS CO$_2$ Training Data for Study Region 1

AIRS CO$_2$ Validation Data for Study Region 1

Figure 1: A spatial dataset made up of 9 days of measurements of mid-tropospheric CO$_2$ in parts per million (ppm). The data considered are between -49° degrees and 36° degrees latitude and -80° degrees and -99.5° degrees longitude, from Februrary 1 through Februrary 9, 2010. The data are randomly split into observed and testing datasets with $n = 57$ and $m = 14$, respectively.


AIRS CO$_2$ Training Data for Study Region 2

AIRS CO$_2$ Validation Data for Study Region 2

Figure 2: A spatial dataset made up of 9 days of measurements of mid-tropospheric CO$_2$ in parts per million (ppm). The data considered are between $-60°$ degrees and $90°$ degrees latitude from Februrary 1 through Februrary 9, 2010. The data are randomly split into observed and testing datasets with $n = 12,358$ and $m = 3,090$, respectively.

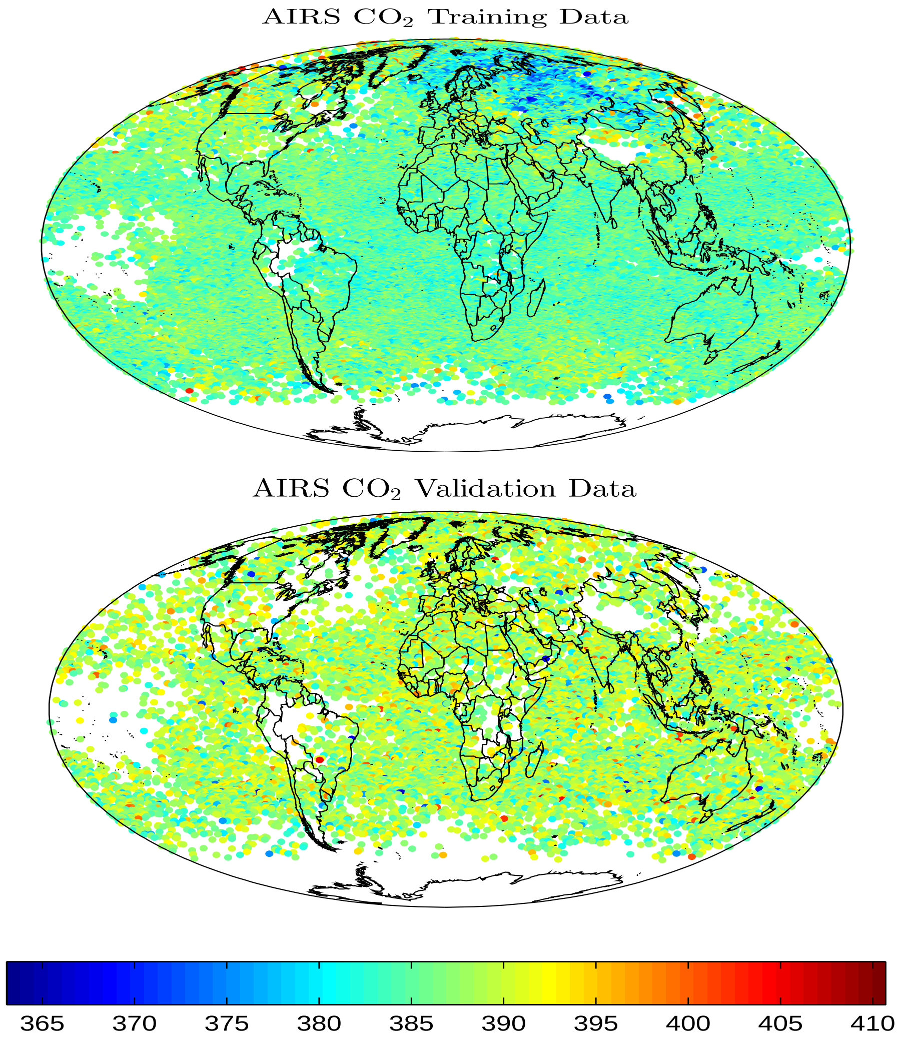


Figure 3: A spatial dataset made up of 9 days of measurements of global mid-tropospheric CO$_2$ in parts per million (ppm). The data considered are between -60° degrees and 90° degrees latitude from Februrary 1 through Februrary 9, 2010. The data are randomly split into observed and testing datasets with $n = 44,621$ and $m = 2,000$, respectively.

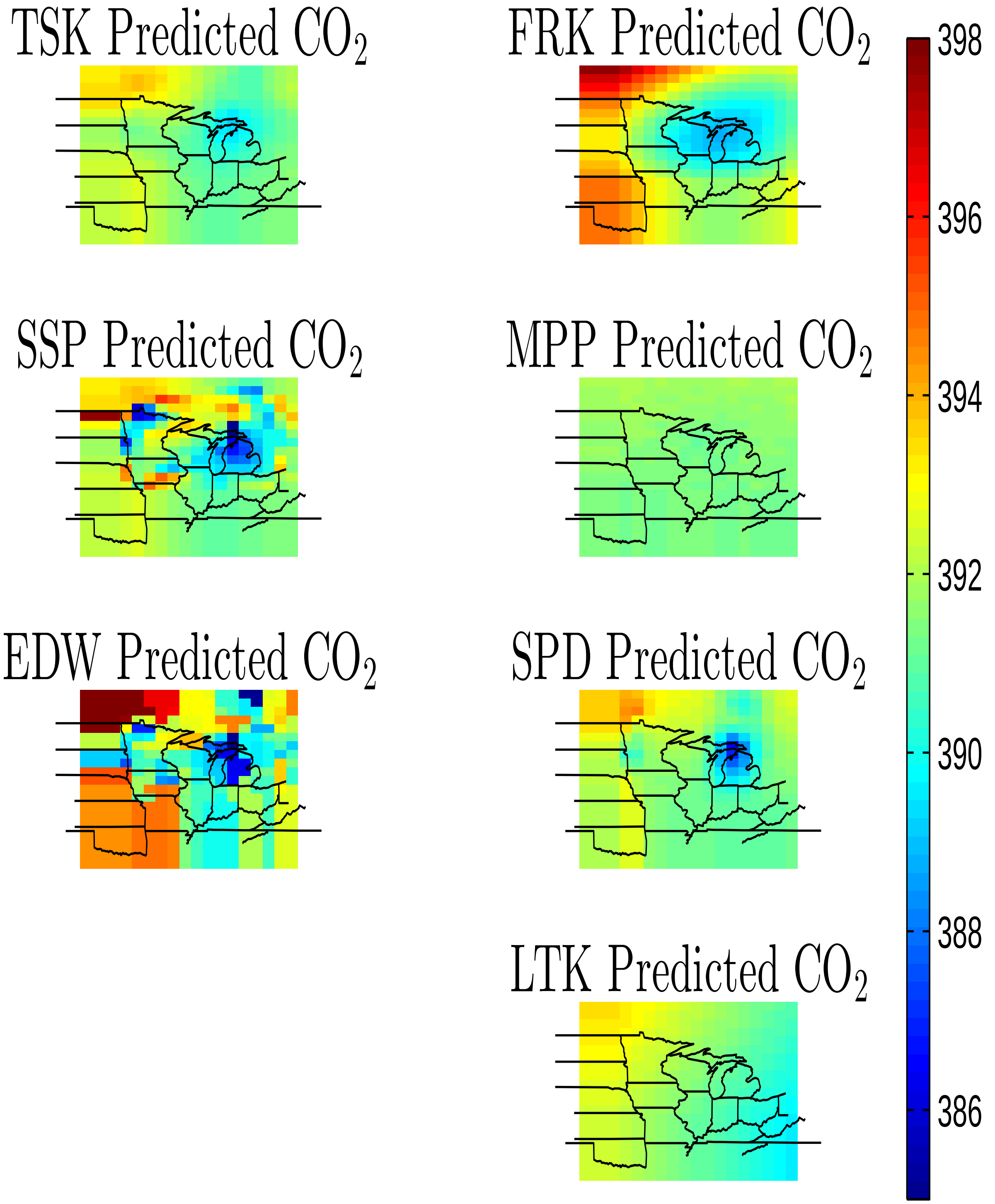


Figure 4: Spatial prediction of mid-tropospheric CO_2 concentrations using TSK, SSP, EDW, FRK, MPP, SPD, and LTK. Predictions are indicated in the title headings and are mapped over Study Region 1.

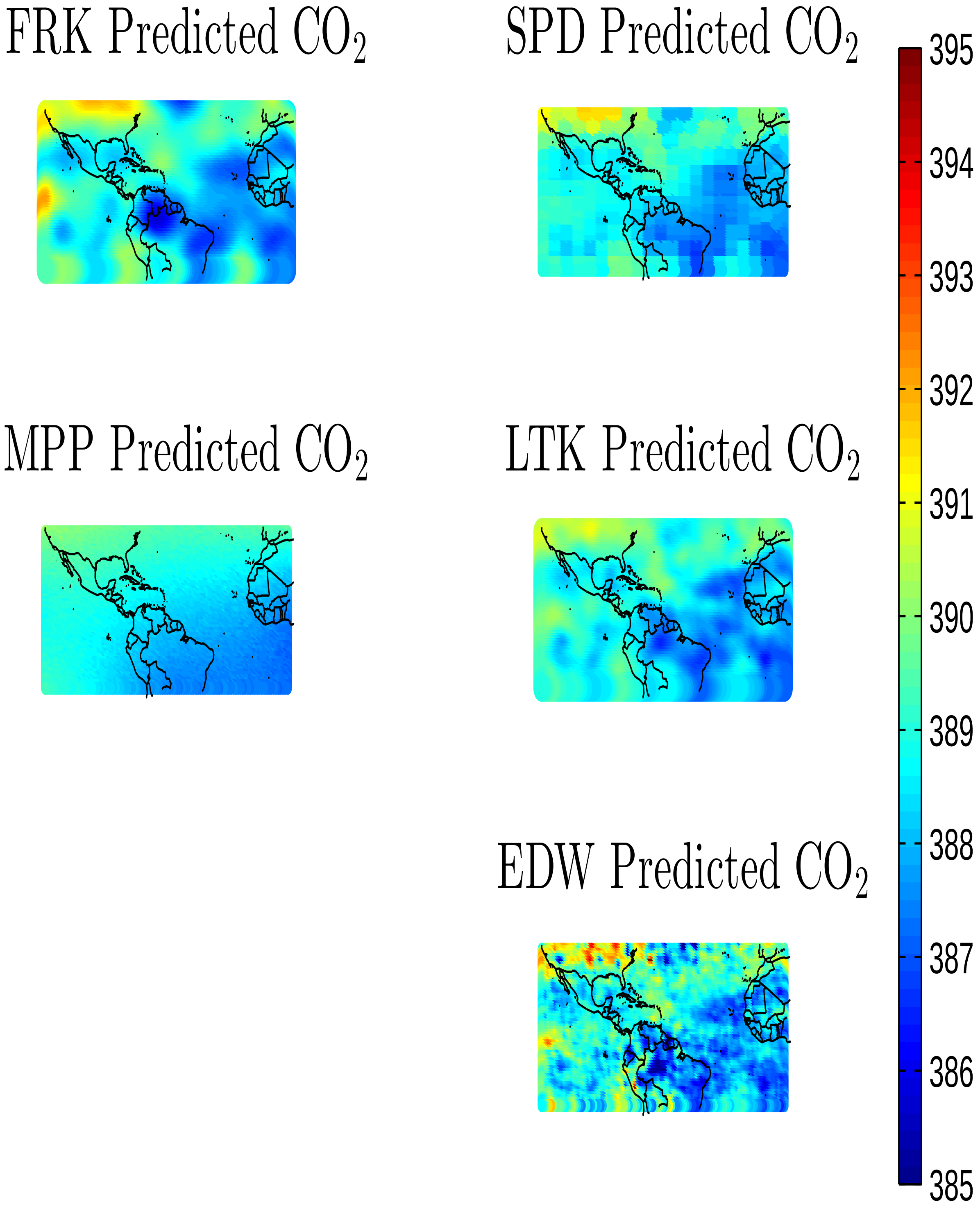


Figure 5: Spatial prediction of mid-tropospheric CO_2 concentrations using FRK, MPP, SPD, LTK, and EDW. Predictions are indicated in the title headings and are mapped over Study Region 2.

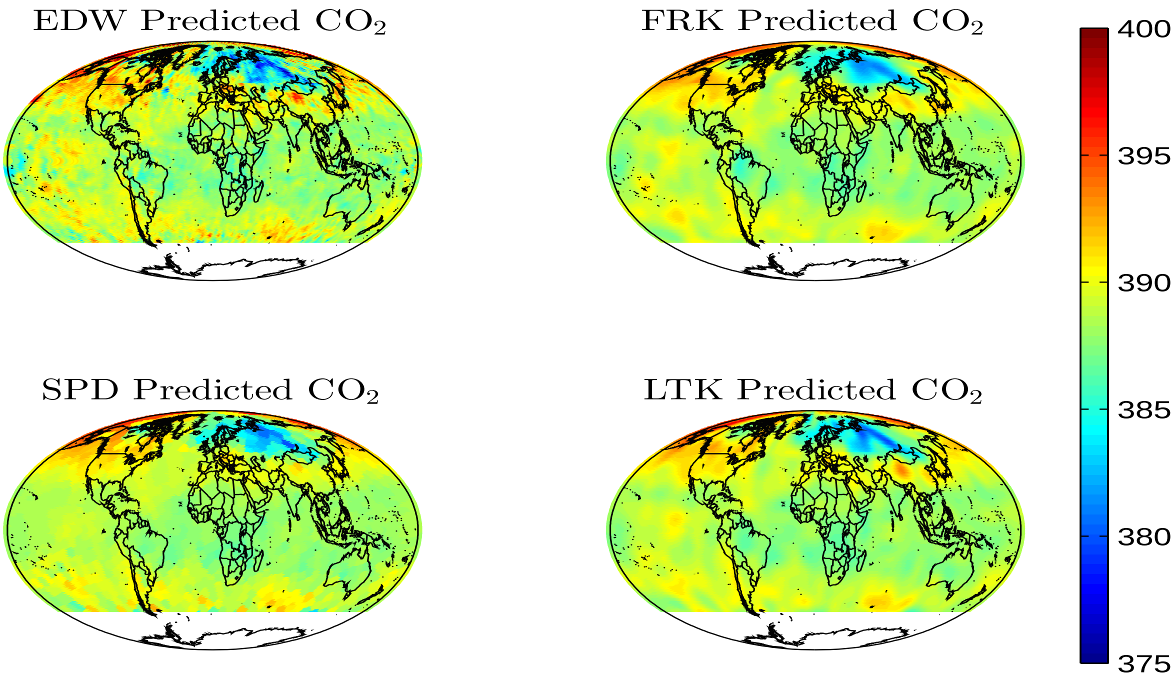


Figure 6: Global spatial prediction of mid-tropospheric CO_2 concentrations using EDW, SPD, FRK, and LTK. Predictions are indicated in the title headings and are mapped over Study Region 3. Note that there is no prediction given below latitude $-60°$, since AIRS has not released any observations there.

Table 1: Results from Study Region 1 (Section 3.1) for the root average squared testing error (RSTE), CPU time, and peak memory-usage by predictor. These quantities are produced using the data in Figure 1.

Predictor	RSTE	PMCC	Lag-1 Semivariogram	CPU Time (in minutes)	Peak Memory Usage (in MB)
TSK	4.7063	-0.4845	0.5739	0.20	171.08
SSP	4.7151	N/A	4.9746	0.02	1,043.80
EDW	4.7703	N/A	7.5466	0.30	733.39
FRK	4.3097	12.5612	2.1298	1.01	791.12
MPP	4.9084	-0.5873	0.0339	3.37	239.51
SPD	4.7399	26.2548	1.1271	0.24	143.14
LTK	5.0163	39.5806	0.2536	2.73	205.84

Table 2: Results from Study Region 2 (Section 3.2) for the root average squared testing error (RSTE), CPU time, and peak memory-usage by predictor. These quantities are produced using the data in Figure 2.

Predictor	RSTE	PMCC	Lag-1 Semivariogram	CPU Time (in minutes)	Peak Memory Usage (in MB)
EDW	3.0396	N/A	0.6966	6.36	850.0470
FRK	3.0067	12.6155	0.2075	0.52	841.0030
MPP	2.9243	-1.0327	0.0164	216.79	2042.6
SPD	2.9630	70.0529	0.1243	0.47	111.18
LTK	2.9855	27.3636	0.1470	1.72	1,971.8

Table 3: Results from Study Region 3 (Section 3.3) for the root average squared testing error (RSTE), CPU time, and peak memory-usage by predictor. These quantities are produced using the data in Figure 3.

Predictor	RSTE	PMCC	Lag-1 Semivariogram	CPU Time (in minutes)	Peak Memory Usage (in MB)
EDW	4.0799	N/A	1.6088	90.68	691.57
FRK	3.9841	12.0974	0.5080	0.51	1,025.40
SPD	3.9882	53.1760	2.1121	4.72	165.19
LTK	4.0026	45.1762	0.1440	85.13	490.60