2013

Comparing and selecting spatial predictors using local criteria

Jonathan R. Bradley
Ohio State University

Noel Cressie
University of Wollongong

Tao Shi
Ohio State University

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

21-13

Comparing and Selecting Spatial Predictors Using Local Criteria

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

Comparing and Selecting Spatial Predictors Using Local Criteria

Jonathan R. Bradley [*] [‡] Noel Cressie [†] Tao Shi [‡]

Abstract

Remote sensing technology for the study of Earth and its environment has led to "Big Data" that, paradoxically, have global extent but may be spatially sparse. Furthermore, the variability in the measurement error and the latent process error may not fit conveniently into the Gaussian linear paradigm. In this paper, we consider the problem of selecting a predictor from a finite collection of spatial predictors of a spatial random process defined on D, a subset of d-dimensional Euclidean space. Critically, we make no statistical distributional assumptions other than additive measurement error. In this nonparametric setting, one could use a criterion based on a validation dataset to select a spatial predictor for all of D. Instead, we propose local criteria based on validation data to select a predictor at each spatial location in D; the result is a hybrid combination of the spatial predictors, which we call a locally selected predictor (LSP). We consider selection from a collection of some of the classical and more recently proposed spatial predictors currently available. In a simulation study, the relative performances of various LSPs, as well as the performance of each of the individual spatial predictors in the collection, are assessed. "Big Data" are always challenging, and here we apply LSP to a very large global spatial dataset of atmospheric CO_2 measurements.

1 Introduction

With the advent of remote sensing technology, larger spatial datasets are becoming more available and, consequently, there has been an increasing number of spatial predictors proposed for large spatial datasets. Some of the more recent spatial predictors include: predictive processes (Banerjee et al., 2008; Finley et al., 2009), Fixed Rank Kriging (Cressie and Johannesson, 2006, 2008; Shi and Cressie, 2007), a stochastic partial differential equation approach (Lindgren et al., 2011), and lattice kriging (Nychka et al., 2012).

[*]Department of Statistics, University of Missouri, Contact: bradleyjr@missouri.edu
[†]National Institute for Applied Statistics Research Australia, University of Wollongong, Australia
[‡]Department of Statistics, The Ohio State University

Furthermore, there are many other methods for spatial prediction currently available that one could use (see, e.g., Cressie, 1993, see, Section 5.9).

Hence, there is a clear practical need for developing criteria to select spatial predictors; that is, there are many different spatial predictors available and, for a given dataset, we would like to base our choice of predictor on statistical criteria rather than familiarity. Fortunately, there is a rich literature available on predictor (or model) selection using criteria (see, e.g., Mallows, 1973; Akaike, 1974; Stein, 1981; Efron, 1983, 1986; Ronchetti and Staudte, 1994; Konishi and Kitagawa, 1996; Ripley, 1996; Ronchetti, 1997; Shao, 1997; Burnham and Anderson, 1998).

Here, we make a distinction between criteria for predictor selection and criteria for model selection. A broad view of *model selection* includes not only choosing from a set of competing probability measures, but more specifically could involve choosing from parameters in a parameter space that defines the set of probability measures (Burnham and Anderson, 1998). This is similar to, but different from, the problem of *predictor selection*, which refers to selecting estimators of an unobserved hidden random process. This distinction between model selection and predictor selection was formally pursued by Vaida and Blanchard (2005) and developed further by Huang and Chen (2007), Liang et al. (2008), Greven and Kneib (2010), and Muller et al. (2013). In this paper, we specifically consider predictor selection for spatial data.

Although the literature on predictor selection is expansive for non-spatially referenced data, it is rather limited in the *spatial-data setting*. In a recent paper, Zhu et al. (2010) has called for more research on the topic, and since then several papers have appeared (Chen et al., 2010; Bradley et al., 2011; Chen and Huang, 2011a,b; Lai et al., 2012).

The traditional approach to predictor (or model) selection is to select a single predictor (or model) that minimizes an estimate of a criterion (Burnham and Anderson, 1998). The *Akaike information criterion* (AIC) (Akaike, 1974) and the *Bayesian information criterion* (BIC) (Schwarz, 1978) are common choices for model selection criteria. Both of these criteria represent estimates of the expected value of the Kullback-Leibler divergence (Kullback and Leibler, 1951). Efron (1983, 1986), and Hastie et al. (2009) consider using training and validation data to select predictors assuming only that measurement errors are additive; here, the criterion involves no parametric assumptions. Other nonparametric selection criteria are also available, including cross-validation and covariance penalized squared error (see, Efron, 2004, and the references therein). “Big Data” are desirable in this setting since the dataset is divided into pieces: a training dataset (to compute the predictors), a validation dataset (to select each predictor based on an empirical criterion),

and perhaps a testing dataset (to evaluate the selected predictor) (Hastie et al., 2009).

Hence, minimizing an empirical criterion is a viable approach to selecting a spatial predictor, since large spatial datasets are common. Furthermore, the use of an empirical criterion avoids having to rely on Gaussian data-model and Gaussian process-model assumptions for a spatial predictor's optimality. Here, we consider the use of a validation dataset to define empirical criteria for spatial-predictor selection.

In this paper, we address the problem of selecting a predictor from a finite collection of spatial predictors of a process Y defined on D, a subset of d-dimensional Euclidean space. In what follows, we shall call the use of predictor selection criteria to select a single predictor for all of D, *global predictor selection*, and the resulting chosen predictor is called a *globally selected predictor* (GSP). A natural extension of GSP, in the spatial setting, is to select a possibly different spatial predictor at *each* location in D. We call this approach *local predictor selection*, which was introduced by Bradley et al. (2012). Clearly, there might be regions of the spatial domain where different predictors are selected; the combination of these selections is a spatial predictor we call a *locally selected predictor* (LSP).

An immediate application of LSP is to *compare* a given set of spatial predictors. Specifically, maps of the selected predictor can be used to describe *where* a predictor is performing well in comparison to the other predictors in the set of possible predictors. Thus, for a dataset under consideration, one can determine whether or not an individual predictor dominates its competitors over the entire spatial domain.

Another application of LSP is to *obtain a more precise spatial predictor* than GSP. Obviously, it not necessarily true that the predictor with smallest total squared prediction error (i.e., a GSP) has minimum squared prediction error at each individual spatial location in the domain of interest D. Bradley et al. (2012) show empirically that LSP can improve total squared prediction error. In this paper, we give a thorough development of LSP from both theoretical and practical standpoints.

The remainder of the paper is organized as follows. In Section 2, we define the LSP methodology, including the statistical model used in this paper, the formal definition of LSP, and theoretical justifications. Then, in Section 3, we provide practical details on the implementation of LSP, including the steps of our local-spatial-predictor-selection algorithm. In Section 4, a simulation study is used to demonstrate that LSP outperforms GSP in terms of total squared prediction error. In Section 5, we implement LSP for a very large spatial dataset of mid-tropospheric CO_2. We conclude with a discussion in Section 6, and we provide technical material in the Appendix.

2 Locally Selected Predictor Methodology

In this section, we define the LSP methodology. Specifically, we present the statistical model (Section 2.1), a generic definition of the LSP (Section 2.2), and a theoretical justification of the LSP (Section 2.3).

2.1 Statistical Model for Spatial Predictor Selection

We assume that observed and potential data have the following additive structure:

$$Z(\mathbf{u}) = Y(\mathbf{u}) + \epsilon(\mathbf{u});\ \mathbf{u} \in D, \tag{1}$$

where $Z(\mathbf{u})$ represents a datum at a spatial location $\mathbf{u} \in D$, $Y(\cdot)$ is a *hidden* process, $\epsilon(\cdot)$ is a measurement-error process independent of $Y(\cdot)$, and $D \equiv \{\mathbf{u}_j : j = 1, ..., N\} \subset \mathbb{R}^d$ is a spatial lattice in d-dimensional Euclidean space. In practice, D is a fine-resolution regular lattice or a possibly irregular lattice of small areas, with a finite number (N) of locations. The process $Y(\cdot)$ is the source of spatial dependence in the data, but its probability distribution is left unspecified. We only require that $E\left(|Y(\mathbf{u})|^2\right) < \infty$, for all $\mathbf{u} \in D$. We assume that $\epsilon(\cdot)$ is a white-noise process with mean zero and variance, $\mathrm{var}\left(\epsilon(\cdot)\right) \equiv \sigma^2_\epsilon v(\cdot)$; unless specified otherwise, σ^2_ϵ and $v(\cdot)$ are assumed unknown. That is, apart from the additive structure, (1) makes very few assumptions and specifically no parametric assumptions.

We observe realizations from the process $Z(\cdot)$ at certain spatial locations in D and divide them into training data and validation data. The *training data* are contained in the n-dimensional vector $\mathbf{Z}^{\mathrm{trn}} \equiv (Z(\mathbf{s}^{\mathrm{trn}}_1), ..., Z(\mathbf{s}^{\mathrm{trn}}_n))'$ observed at locations $D^{\mathrm{trn}} \equiv \{\mathbf{s}^{\mathrm{trn}}_1, ..., \mathbf{s}^{\mathrm{trn}}_n\} \subset D$. These n data are used to predict the process $Y(\cdot)$ at all locations $D \equiv \{\mathbf{u}_1, ..., \mathbf{u}_N\}$, where $n < N$.

The *validation data* are contained in the m-dimensional vector, $\mathbf{Z}^{\mathrm{val}} \equiv (Z(\mathbf{s}^{\mathrm{val}}_1), ..., Z(\mathbf{s}^{\mathrm{val}}_m))'$, observed at locations $D^{\mathrm{val}} \equiv \{\mathbf{s}^{\mathrm{val}}_1, ..., \mathbf{s}^{\mathrm{val}}_m\} \subset D$, where $D^{\mathrm{val}} \cap D^{\mathrm{trn}} = \emptyset$. These validation data are used to evaluate the "performance" of spatial predictors at locations D^{val}, where $m < N$ and $m + n \leq N$. For more on the use of validation and training datasets to compare and select predictors, see Hastie et al. (2009).

Using the training data $\mathbf{Z}^{\mathrm{trn}}$, we compute a set of spatial predictors of the unobserved hidden process $Y(\cdot)$ denoted as $\{\hat{Y}^{(k)}(\cdot, \mathbf{Z}^{\mathrm{trn}}) : k = 1, ..., K\}$, where $K \geq 2$ and $\hat{Y}^{(k)}$ is any real-valued function with domain $D \times \mathbb{R}^n$. These spatial predictors are often motivated by statistical models that involve more assumptions on $\mathbf{Z}^{\mathrm{trn}}$ and $\mathbf{Z}^{\mathrm{val}}$ than those given in (1), but that is inconsequential here. Our approach to comparing and selecting spatial predictors allows us to compare predictors derived from different assumptions on the

stochastic process or, possibly, predictors that are purely deterministic in nature. In Section 2.2, we provide the technical details behind our approach to comparing and selecting spatial predictors.

2.2 Locally Selected Predictor (LSP)

We first consider an ideal setting where the predictor-selection criterion is based on an unobservable quantity, an oracle (Donoho and Johnstone, 1994). The squared prediction error of the k-th predictor, $\hat{Y}^{(k)}$, at generic prediction location $\mathbf{u}$ is

$$\mathrm{SPE}^{(k)}(\mathbf{u}, \mathbf{Z}^{\mathrm{trn}}) \equiv (Y(\mathbf{u}) - \hat{Y}^{(k)}(\mathbf{u}, \mathbf{Z}^{\mathrm{trn}}))^2; \ k = 1, ..., K. \tag{2}$$

Minimizing (2) results in an oracle predictor since it depends on the unobserved $Y(\mathbf{u})$. From (2), we define local spatial-predictor selection at $\mathbf{u} \in D$, as follows. For a given $Y(\cdot)$, define

$$\tilde{k}(\mathbf{u}, \mathbf{Z}^{\mathrm{trn}}) \equiv \arg\min \left\{ \mathrm{SPE}^{(k)}(\mathbf{u}, \mathbf{Z}^{\mathrm{trn}}) : k = 1, ..., K \right\}. \tag{3}$$

Then for any $k = 1, ..., K$, it is immediate that $\sum_{i=1}^{N} \mathrm{SPE}^{(\tilde{k})}(\mathbf{u}_i, \mathbf{Z}^{\mathrm{trn}}) \leq \sum_{i=1}^{N} \mathrm{SPE}^{(k)}(\mathbf{u}_i, \mathbf{Z}^{\mathrm{trn}})$. Hence, the locally selected *oracle predictor*,

$$\tilde{\mathbf{Y}} \equiv (\hat{Y}^{(\tilde{k})}(\mathbf{u}_i, \mathbf{Z}^{\mathrm{trn}}) : i = 1, ..., N)', \tag{4}$$

is optimal since it performs as well or better than any of the $k = 1, ..., K$ spatial predictors. However, since the unobserved $\{Y(\mathbf{u}_1), ..., Y(\mathbf{u}_N)\}$ are used to compute $\tilde{k}$ and $\tilde{\mathbf{Y}}$, (4) is unavailable in practice.

Thus, we consider estimating SPE in (2) with a local average of the squared validation error,

$$\mathrm{LSVE}^{(k)}(\mathbf{u}_i, \mathbf{Z}^{\mathrm{trn}}; H^{\mathrm{val}}) \equiv \frac{1}{|H^{\mathrm{val}}(\mathbf{u}_i)|} \sum_{\mathbf{s} \in H^{\mathrm{val}}(\mathbf{u}_i)} (Z(\mathbf{s}) - \hat{Y}^{(k)}(\mathbf{s}, \mathbf{Z}^{\mathrm{trn}}))^2, \tag{5}$$

where $H^{\mathrm{val}}(\mathbf{u}) \subset D^{\mathrm{val}}$ is a generic set of validation locations that are "close to" $\mathbf{u}$, and $|H^{\mathrm{val}}(\mathbf{u}_i)|$ denotes the number of elements contained in the set $H^{\mathrm{val}}(\mathbf{u}_i)$; $i = 1, ..., N$.

Since the right-hand side of (5) can be calculated from the data (i.e., it is a statistic), we use the spatial predictor that minimizes $\{\mathrm{LSVE}^{(k)}(\mathbf{u}_i, \mathbf{Z}^{\mathrm{trn}}; H^{\mathrm{val}}) : k = 1, ..., K\}$. That is, the index of the selected

predictor is, for $i = 1, ..., N$,

$$\hat{k}(\mathbf{u}_i, \mathbf{Z}^{\mathrm{trn}}; H^{\mathrm{val}}) \equiv \arg\min \left\{ \mathrm{LSVE}^{(k)}(\mathbf{u}_i, \mathbf{Z}^{\mathrm{trn}}; H^{\mathrm{val}}) : k = 1, ..., K \right\}. \tag{6}$$

This defines a *locally selected predictor* (LSP),

$$\hat{\mathbf{Y}}^{\mathrm{LSP}} \equiv (\hat{Y}^{(\hat{k}(\mathbf{u}_i, \mathbf{Z}^{\mathrm{trn}}; H^{\mathrm{val}}))}(\mathbf{u}_i, \mathbf{Z}^{\mathrm{trn}}) : i = 1, ..., N)'. \tag{7}$$

Notice that different choices for $\{H^{\mathrm{val}}(\mathbf{u}_i) : i = 1, ..., N\}$ yield different LSPs.

For the special case of $H^{\mathrm{val}}(\cdot) \equiv D^{\mathrm{val}}$ we obtain the *globally selected predictor*,

$$\hat{\mathbf{Y}}^{\mathrm{GSP}} \equiv (\hat{Y}^{(\hat{k}(\mathbf{u}_i, \mathbf{Z}^{\mathrm{trn}}; D^{\mathrm{val}}))}(\mathbf{u}_i, \mathbf{Z}^{\mathrm{trn}}) : i = 1, ..., N)'. \tag{8}$$

In this paper, we compare $\hat{\mathbf{Y}}^{\mathrm{GSP}}$ to $\hat{\mathbf{Y}}^{\mathrm{LSP}}$ based on various choices of H^{val} to show that local spatial-predictor selection outperforms global spatial-predictor selection. Note that this is automatically true when $\hat{k}$ is replaced with $\tilde{k}$ in (7) and (8), however we have already discussed how the oracle predictor based on $\tilde{k}$ is generally unavailable in practice.

Of course, one can replace the squared prediction error in (2) with other data-dependent criteria to create new versions of (6) and (7). Bradley et al. (2012) considered a local empirical deviance information criterion, which does not require the separation of observations into training and validation datasets. However, it is defined most naturally for linear spatial predictors, and in this paper we wish to assume as little as possible and do not distinguish between linear and non-linear spatial predictors.

In what follows, we shall use (5), (6), and (7) to define our LSPs. As a practical matter, we must specify H^{val}, and in Section 2.3 we present several choices for H^{val}.

2.3 Specification of LSPs

In the Appendix, we prove that for H^{val} given by

$$H^{\mathrm{val}}(\mathbf{s}_j^{\mathrm{val}}) = \{\mathbf{s}_j^{\mathrm{val}}\}, \tag{9}$$

the following result holds,

$$E\left(\mathrm{SPE}^{(\hat{k})}(\mathbf{s}_j^{\mathrm{val}}, \mathbf{Z}^{\mathrm{trn}})\right) = E\left(\mathrm{SPE}^{(\tilde{k})}(\mathbf{s}_j^{\mathrm{val}}, \mathbf{Z}^{\mathrm{trn}})\right);\ j = 1, ..., m. \tag{10}$$

Thus, $\hat{\mathbf{Y}}^{\mathrm{LSP}}$ can be considered optimal at elements of the vector defined by the validation locations $D^{\mathrm{val}} = \{\mathbf{s}_j^{\mathrm{val}} : j = 1, ..., m\}$, since for all $\mathbf{u} \in D^{\mathrm{val}}$,

$$E\left(\mathrm{SPE}^{(\hat{k})}(\mathbf{u}, \mathbf{Z}^{\mathrm{trn}})\right) = E\left(\mathrm{SPE}^{(\tilde{k})}(\mathbf{u}, \mathbf{Z}^{\mathrm{trn}})\right) \leq E\left(\mathrm{SPE}^{(k)}(\mathbf{u}, \mathbf{Z}^{\mathrm{trn}})\right);\ k = 1, ..., K. \tag{11}$$

However, notice that (11) does not necessarily hold for $\mathbf{u} \notin D^{\mathrm{val}}$. One of our strategies for selecting spatial predictors is to define $H^{\mathrm{val}}(\mathbf{u})$ at *all* locations $\mathbf{u} \in D$.

Since the stochastic nature of the spatial process $Y(\cdot)$ is left unspecified in (1), it is difficult to ascertain the relationship between $E\left(\mathrm{SPE}^{(\hat{k})}(\mathbf{u}, \mathbf{Z}^{\mathrm{trn}})\right)$ and $E\left(\mathrm{SPE}^{(\tilde{k})}(\mathbf{u}, \mathbf{Z}^{\mathrm{trn}})\right)$ for $\mathbf{u} \notin D^{\mathrm{val}}$. However, we argue that choices for H^{val} other than $H^{\mathrm{val}}(\mathbf{u}) = D^{\mathrm{val}}$, for all $\mathbf{u} \in D$, which yields the GSP, can lead to spatial-predictor selection with improved mean squared prediction error.

For example, consider *simple local predictor selection* (SLS) defined by,

$$H_1^{\mathrm{val}}(\mathbf{u}) \equiv \begin{cases} \{\mathbf{u}\} & \text{if } \mathbf{u} \in D^{\mathrm{val}} \\ D^{\mathrm{val}} & \text{if } \mathbf{u} \notin D^{\mathrm{val}}. \end{cases} \tag{12}$$

Notice that we introduce a subscript in (12) to distinguish H_1^{val} from a generic H^{val}. Here, LSP based on (6) and (12) is optimal at $\mathbf{u} \in D^{\mathrm{val}}$ and equals the relevant component of GSP in (8) at all $\mathbf{u} \notin D^{\mathrm{val}}$. Thus, the total mean squared prediction error of LSP based on (12) is smaller than or equal to the total mean squared prediction error of the GSP. That is,

$$\sum_{\mathbf{s} \in D^{\mathrm{val}}} E\left(\mathrm{SPE}^{(\hat{k}(\mathbf{s}, \mathbf{Z}^{\mathrm{trn}}; H_1^{\mathrm{val}}))}(\mathbf{s}, \mathbf{Z}^{\mathrm{trn}})\right) \leq \sum_{\mathbf{s} \in D^{\mathrm{val}}} E\left(\mathrm{SPE}^{(\hat{k}(\mathbf{s}, \mathbf{Z}^{\mathrm{trn}}; D^{\mathrm{val}}))}(\mathbf{s}, \mathbf{Z}^{\mathrm{trn}})\right). \tag{13}$$

Adding the term, $\sum_{\mathbf{u} \notin D^{\mathrm{val}}} E\left(\mathrm{SPE}^{(\hat{k}(\mathbf{u}, \mathbf{Z}^{\mathrm{trn}}; D^{\mathrm{val}}))}(\mathbf{u}, \mathbf{Z}^{\mathrm{trn}})\right)$, to both sides of (13) results in

$$\sum_{\mathbf{u} \in D} E\left(\mathrm{SPE}^{(\hat{k}(\mathbf{u}, \mathbf{Z}^{\mathrm{trn}}; H_1^{\mathrm{val}}))}(\mathbf{u}, \mathbf{Z}^{\mathrm{trn}})\right) \leq \sum_{\mathbf{u} \in D} E\left(\mathrm{SPE}^{(\hat{k}(\mathbf{u}, \mathbf{Z}^{\mathrm{trn}}; D^{\mathrm{val}}))}(\mathbf{u}, \mathbf{Z}^{\mathrm{trn}})\right). \tag{14}$$

Consequently, LSP based on H_1^{val} in (12) outperforms GSP in terms of total mean squared prediction error. However, one can conceive of many different choices of H^{val}, in addition to (12), that might lead to further improvement in the total mean squared prediction error of the LSP.

Because of the form of H_1^{val} in (12), we expect that its associated LSP surface will have discontinuities. Then a reasonable alternative would be to select a predictor at $\mathbf{u} \notin D^{\text{val}}$ based on a *local* average of the squared validation error. Thus, we consider an approach that we call *moving-window predictor selection* (MWS). Define

$$H_2^{\text{val}}(\mathbf{u}) \equiv \begin{cases} \{\mathbf{u}\} & \text{if } \mathbf{u} \in D^{\text{val}} \\ \{\mathbf{s} : \mathbf{s} \in D^{\text{val}} \text{ and } ||\mathbf{s} - \mathbf{u}|| \leq w\} & \text{if } \mathbf{u} \notin D^{\text{val}}, \end{cases} \tag{15}$$

where $w > 0$ is prespecified. According to (6) and (15), at $\mathbf{u} \notin D^{\text{val}}$ the selected predictor performs best in terms of average squared validation error, averaged over validation locations within a d-dimensional ball centered at $\mathbf{u}$ with radius w. If $\text{LSVE}^{(k)}(\mathbf{u}, \mathbf{Z}^{\text{trn}})$ in (5) is similar in value to $\text{SPE}^{(k)}(\mathbf{u}, \mathbf{Z}^{\text{trn}})$ in (2), for $\mathbf{u} \notin D^{\text{val}}$, then H_2^{val} is a viable alternative.

We also consider a related approach to (15) that we call *nearest-neighborhood predictor selection* (NNS). Instead of selecting a spatial predictor based on the average squared validation error over validation locations within a ball of fixed radius centered at $\mathbf{u}$, we select based on the average squared validation error over the g locations in D^{val} closest to $\mathbf{u}$, where g is a prespecified positive integer less than or equal to $m - 1$. Define

$$H_3^{\text{val}}(\mathbf{u}) \equiv \begin{cases} \{\mathbf{u}\} & \text{if } \mathbf{u} \in D^{\text{val}} \\ \{\mathbf{s} : \mathbf{s} \in D^{\text{val}} \text{ and } ||\mathbf{s} - \mathbf{u}|| \leq w_g\} & \text{if } \mathbf{u} \notin D^{\text{val}}, \end{cases} \tag{16}$$

where w_g is the g-th largest value from among the values, $||\mathbf{s}_1^{\text{val}} - \mathbf{u}||, ..., ||\mathbf{s}_m^{\text{val}} - \mathbf{u}||$. According to (6) and (16), the model is selected that performs best in terms of average squared validation error, averaged over validation locations within a d-dimensional ball centered at $\mathbf{u}$, but now with an adaptive radius that contains exactly g nearest neighbors.

A different approach to (15) and (16) would be to choose a predictor at location $\mathbf{u}$ based on a single, well chosen neighbor. This leads us to an approach that we call *Voronoi-polygon predictor selection* (VPS). Voronoi polygons (or Delaunay tesselations) are a partition of a spatial domain that are formed by nearest-distance locus points of a finite set of locations in the domain. In our case, the locations in domain D are $\{\mathbf{s}_j^{\text{val}} : j = 1, ..., m\}$. Hence, one and only one validation location is contained within each Voronoi polygon

V_j, for $j = 1, ..., m$. Further, for any $\mathbf{u} \in V_j$, $||\mathbf{u} - \mathbf{s}_j^{\text{val}}|| \leq ||\mathbf{u} - \mathbf{s}_{j'}||; j' \neq j$ (e.g., Cressie, 1993, p. 374).

Then VPS is defined by,

$$H_4^{\text{val}}(\mathbf{u}) \equiv \{\mathbf{s}_j^{\text{val}} : \mathbf{u} \in V_j, j = 1, ..., m\}; \mathbf{u} \in D. \tag{17}$$

Notice that for each $i = 1, ..., N$, $H_4^{\text{val}}(\mathbf{u}_i)$ in (17) is *equal* to a single validation location, and $H_4^{\text{val}}(\mathbf{u}) = \{\mathbf{u}\}$, for $\mathbf{u} \in D^{\text{val}} \subset D$.

Finally, we consider a maximal choice for H^{val} based on averaging *all* squared validation errors. This leads to the aforementioned GSP in (8). Define,

$$H_5^{\text{val}}(\mathbf{u}) \equiv D^{\text{val}}; \mathbf{u} \in D. \tag{18}$$

In subsequent sections, we investigate the performance of LSP based on $H_1^{\text{val}}, ..., H_4^{\text{val}}$ and GSP based on H_5^{val} in terms of mean squared prediction error.

3 Locally Selected Predictor in Practice

In this section, we describe the LSP methodology from a practical perspective. Specifically, we provide a set of competing predictors to select from (Section 3.1), and we set out the steps of our LSP algorithm (Section 3.2).

3.1 Spatial Predictors under Consideration

To implement the LSP methodology, we need to specify $\{\hat{Y}^{(k)}(\cdot, \mathbf{Z}^{\text{trn}}) : k = 1, ..., K\}$. In this paper, we consider local selection of $K = 7$ well known spatial predictors, which are discussed below. One of the most well known methods of spatial prediction is *kriging*, which is often associated with (but not restricted to) classes of stationary covariance functions for $Y(\cdot)$ (e.g., Cressie, 1993, Section 3.4). Spatial predictions using kriging are obtained by minimizing the mean squared prediction error among the class of linear unbiased predictors. The term kriging was first coined by Matheron (1963), although other disciplines have used different terminology, such as optimal interpolation (e.g., Cressie, 1990). Since then, kriging with stationary covariance functions 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).

In this paper, we call this approach *traditional stationary kriging* (TSK), and we computed it using the R-package "geoR" version 1.7-4 (Ribeiro, Jr. and Diggle, 2012).

Another popular method of spatial prediction is known as *smoothing splines* (SSP). Spatial predictions using smoothing splines are obtained by minimizing a penalized-least-squares criterion (Wahba, 1990; Nychka, 2001). In subsequent sections, we computed SSP using the Matlab (Version 8.0) function "griddata."

There are spatial prediction methods not based on optimizing a criterion, but they express, in an *ad hoc* way, that nearby observations receive more weight when predicting at a given location. For example, *negative-exponential-distance weighting* (EDW) is a simple spatial interpolation method where a datum's negative log weight is proportional to the Euclidean distance from the prediction location to the datum's location. Let $d_j(\mathbf{u}) \equiv ||\mathbf{u} - \mathbf{s}_j^{\mathrm{trn}}||$ be the Euclidean distance between $\mathbf{u}$ and $\mathbf{s}_j^{\mathrm{trn}}$. The negative-exponential-distance-weighting predictor is,

$$\hat{Y}^{\mathrm{EDW}}(\mathbf{u}, \mathbf{Z}^{\mathrm{trn}}) \equiv \frac{\sum_{j=1}^{n} \exp\{-d_j(\mathbf{u})\} Z(\mathbf{s}_j^{\mathrm{trn}})}{\sum_{j=1}^{n} \exp\{-d_j(\mathbf{u})\}}.$$

In this paper, we wrote our own Matlab script to compute EDW.

Recently, Cressie and Johannesson (2006, 2008) have defined a method of prediction associated with the spatial mixed effects (SME) model (Cressie and Johannesson, 2006, 2008) called *Fixed Rank Kriging* (FRK), which seeks efficient calculation of the kriging predictor in the setting where n is very-large-to-massive. The spatial mixed effects model incorporates spatial dependence using a random linear combination of spatial basis functions. The number of terms in this linear combination is specified to be much smaller than the number of data points n, which is often called a "reduced-rank" approach to spatial prediction (Cressie and Johannesson, 2008; Banerjee et al., 2008; Finley et al., 2009; Wikle, 2010). An advantage of a reduced-rank approach is that the resulting FRK predictor can be computed very quickly. To compute the FRK predictor, we used Matlab code provided by The Ohio State University's *Spatial Statistics and Environmental Statistics* (SSES) website (Katzfuss and Cressie, 2011b).

Banerjee et al. (2008) and Finley et al. (2009) also use a reduced-rank approach to define a spatial-predictor called the *modified predictive process* (MPP). Their approach is to first predict random effects, which they call the predictive process. Then, predictions of $Y(\cdot)$ are found by multiplying the prediction of the random effects by a set of basis functions. To compute the MPP predictor, we used the R-package "spBayes" (Finley et al., 2012).

In contrast to FRK and MPP, Lindgren et al. (2011) do not use a reduced-rank approach. Instead, the random effects are specified to be a Gaussian Markov random field with a precision (i.e., inverse covariance) matrix chosen in a such way that $\mathrm{cov}(Y(\mathbf{u}+\mathbf{h}), Y(\mathbf{u}))$ is close to a Matérn covariance function. Lindgren et al. (2011) call their method the "SPDE approach," where SPDE is shorthand for stochastic partial differential equation. In this paper, we refer to their SPDE approach as simply SPD. By parameterizing the precision matrix directly, they avoid (a difficult) inversion of $\mathrm{cov}\left(\mathbf{Z}^{\mathrm{trn}}\right)$. To compute the SPD predictor, we used the R-package "inla" (Rue et al., 2009; H. Rue, 2012).

A similar approach is called *lattice kriging* (LTK) (Nychka, 2001). Again, a low-rank approach is avoided here, by assuming that the random-effects component of the spatial model follows a first-order vector autoregressive process. Specifically, the random-effects components are ordered according to the index of the knots used to compute spatial basis functions, and the model is not invariant to changes in the ordering. Computational efficiency is obtained for large n by using sparse-matrix techniques. To compute the LTK predictor, we used the R package "LatticeKrig" (Nychka et al., 2012).

Consider the generic predictor, PRD. Then to predict $Y(\mathbf{u})$, we notate the predictor as $\hat{Y}^{\mathrm{PRD}}(\mathbf{u}, \mathbf{Z}^{\mathrm{trn}})$. For the seven predictors used in Sections 3-5, PRD = TSK, SSP, EDW, FRK, MPP, SPD, and LTK.

3.2 A Summary of the Spatial-Predictor Selection Methodology

We now set out the steps we take to compute the LSP based on the seven spatial predictors (Section 3.1) and for a given H^{val} (Section 2.3). We give generic steps, however we acknowledge that in any particular application, some modifications may be needed.

1. We start by computing the K (here, $K = 7$) spatial predictors. As was mentioned in Section 3.1, there are many methods of spatial prediction available, both stochastic and non-stochastic, and the set of predictors under consideration here was chosen from among the classical and the more recent.

2. For each of the seven predictors given in Section 3.1, we estimate the squared prediction error at each location $\mathbf{u}_i$, for $i = 1, ..., N$, using a local average of the squared validation error given by (5). That is, we calculate the values in the set, $\{\mathrm{LSVE}^{(k)}(\mathbf{u}_i, \mathbf{Z}^{\mathrm{trn}}; H_b^{\mathrm{val}}) : k = 1, ..., 7, i = 1, ..., N, b = 1, ..., 5\}$.

3. We construct the LSP given by (7), for the five choices of H^{val} given by (12), (15), (16), (17), and (18): $(\hat{Y}^{(\hat{k}(\mathbf{u}_i, \mathbf{Z}^{\mathrm{trn}}; H_b^{\mathrm{val}}))}(\mathbf{u}_i, \mathbf{Z}^{\mathrm{trn}}) : i = 1, ..., N)'; \;\; b = 1, ..., 5$, where recall that $b = 5$ in fact corresponds to the GSP.

4. We create maps of the spatial surfaces obtained from the following vectors, $(\hat{k}(\mathbf{u}_i, \mathbf{Z}^{\mathrm{trn}}; H_b^{\mathrm{val}}) : i = 1, ..., N)'$ and $(\hat{Y}^{(\hat{k})}(\mathbf{u}_i, \mathbf{Z}^{\mathrm{trn}}) : i = 1, ..., N)'$. For each $b = 1, ..., 5$, the map of $\hat{k}(\cdot)$ can be used to compare the K predictors under consideration. This map indicates regions of D where a particular predictor is preferred according to LSVE given by (5). For each $b = 1, ..., 5$, the map of $\hat{Y}^{(\hat{k})}(\cdot, \mathbf{Z}^{\mathrm{trn}})$ can be used to visualize the LSP of the process $Y(\cdot)$.

In Sections 4 and 5, we provide a simulation study and a real data analysis to exemplify the LSP defined by Steps 1 through 4. Further, in the simulation study, we show that an LSP generally leads to an improvement in total mean squared prediction error over that for the GSP.

4 A Simulation Study

In this section, we investigate statistical properties of our local approach to predictor selection using a simulation study. We demonstrate the gains in predictive performance that can be achieved by using an LSP as opposed to the GSP.

4.1 Simulation Set-Up

In what follows, a two-dimensional spatial domain $D \equiv \{\mathbf{u} = (u_1, u_2)' : u_1 = -25, ..., 25, u_2 = -25, ..., 25\}$ is considered. This gives a total of $N = 51 \times 51 = 2,601$ spatial locations, at which approximately 50% (1,241 locations) have observations. For a given value of $Y(\cdot)$, we simulate $Z(\cdot)$ using the model in (1) and the following assumptions: $\epsilon(\cdot)$ is a Gaussian white-noise process, $v(\cdot) \equiv 1$, and σ_ϵ^2 is obtained below using (20). Then the spatial random process $Y(\cdot)$ is simulated according to a zero-mean Gaussian process as follows:

$$Y(\mathbf{u}) = \mathbf{S}(\mathbf{u})'\boldsymbol{\eta} + \xi(\mathbf{u}); \ \ \mathbf{u} \in D.$$

The vector $\mathbf{S}(\cdot)$ is defined as an N-dimensional vector of Fourier basis functions (e.g., Royle and Wikle, 2005). The N-dimensional random vector $\boldsymbol{\eta}$ is specified as a Gaussian process with mean zero and covariance matrix $\mathbf{K}$. The random process $\xi(\cdot)$ is assumed to be a Gaussian white-noise process with mean zero and variance σ_ξ^2. The random processes $\nu(\cdot)$ and $\xi(\cdot)$ are assumed to be independent.

The $N \times N$ matrix $\mathbf{K} \equiv \mathrm{cov}(\boldsymbol{\eta})$ is calibrated against the stationary exponential covariance function

as follows. We choose positive-definite $\mathbf{K}$ such that $||\mathbf{SKS}' - \mathbf{\Sigma}^{(0)}||_F^2$ is minimized, where $||\cdot||_F$ is the Frobenius norm, $\mathbf{S} \equiv (\mathbf{S}(\mathbf{u}_j) : j = 1, ..., N)'$, and the (i,j)-th element of $\mathbf{\Sigma}^{(0)}$ is $\exp\{-||\mathbf{u}_i - \mathbf{u}_j||/\kappa\}$, for $\mathbf{u}_i, \mathbf{u}_j \in D$. The Frobenius norm is given by $||\mathbf{M}||_F \equiv \{\mathrm{trace}(\mathbf{M}'\mathbf{M})\}^{1/2}$ for a square matrix $\mathbf{M}$. The parameter $\kappa > 0$ of the exponential covariance function represents the strength of spatial dependence.

In this simulation study, we choose to fix the value of two parameters of our simulation model: the amount of spatial dependence κ used to calibrate the matrix $\mathbf{K} \equiv \mathrm{cov}(\boldsymbol{\eta})$, and the amount of fine-scale variation $\sigma_\xi^2 \equiv \mathrm{var}\,(\xi(\cdot))$. Following Cressie et al. (2010), Bradley et al. (2011), and Katzfuss and Cressie (2011a), we set $\kappa = 5$, which corresponds to an effective range of 15; and we calibrate σ_ξ^2 by specifying the fine-scale-variation proportion, which is defined as:

$$\mathrm{FVP} \equiv \frac{N\sigma_\xi^2}{N\sigma_\xi^2 + \mathrm{trace}(\mathbf{SKS}')}. \tag{19}$$

The matrices $\mathbf{K}$ and $\mathbf{S}$ are specified above; consequently, $\mathrm{trace}(\mathbf{SKS}')$ is known. Here, we consider $\mathrm{FVP} = 0.05$ and solve (19) to obtain $\sigma_\xi^2 = 0.0526$.

4.2 Factors of the Simulation Study

We consider four different factors in this simulation study: the signal-to-noise ratio (**SNR**), the proportion of training data (**PTD**), the missing-data structure (**MDS**), and the LSPs (**LSP**). We now present the details of these factors.

Signal-to-Noise Ratio (SNR)

Notice that if $\epsilon(\mathbf{u}) = 0$, then $\hat{k}(\mathbf{u}) = \tilde{k}(\mathbf{u})$ by their definitions in (6) and (3). Thus, we expect σ_ϵ^2 to have a role in determining the difference between $\hat{k}$ and the oracle choice $\tilde{k}$. We calibrate the value of σ_ϵ^2 using the signal-to-noise ratio,

$$\mathrm{SNR} \equiv \frac{N\sigma_\xi^2 + \mathrm{trace}(\mathbf{SKS}')}{N\sigma_\epsilon^2}. \tag{20}$$

Since σ_ξ^2 and the matrices $\mathbf{K}$ and $\mathbf{S}$ are specified above, we can solve (20) for σ_ϵ^2. We consider SNR = 5 (and SNR = 10), which yields $\sigma_\epsilon^2 = 0.2105$ (and $\sigma_\epsilon^2 = 0.1053$). We consider SNR = 5 (labeled as **SNR**= 1) and SNR = 10 (labeled as **SNR**= 2) as moderate and large values of SNR, respectively.

In what follows, we simulate 240 replicates of $\{Z(\mathbf{u}) : \mathbf{u} \in D\}$ and $\{Y(\mathbf{u}) : \mathbf{u} \in D\}$ with **SNR** = 1 and

another 240 replicates of $\{Z(\mathbf{u}) : \mathbf{u} \in D\}$ and $\{Y(\mathbf{u}) : \mathbf{u} \in D\}$ with **SNR** = 2. We give an example of the surfaces of simulated values of $Z(\cdot)$ and $Y(\cdot)$ for both **SNR** = 1 and **SNR** = 2 in the top and bottom rows of Figure 1, respectively.

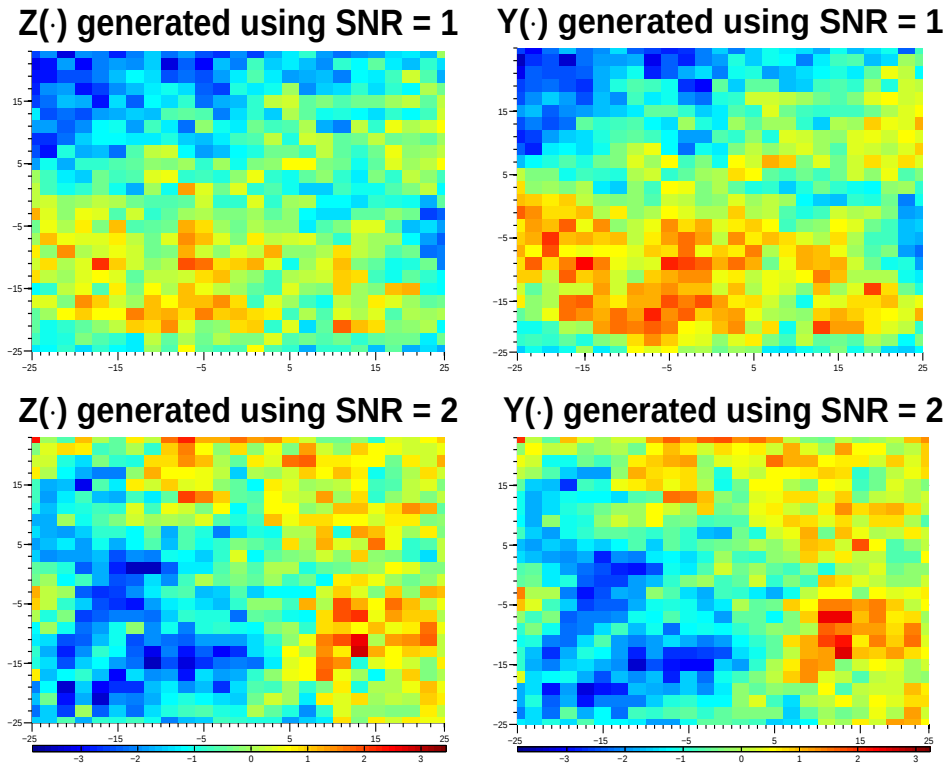

Figure 1: Two realizations of the simulated spatial process $Z(\cdot)$ and $Y(\cdot)$, at two signal-to-noise ratios, respectively. The top row is simulated using **SNR** $= 1$ (moderate SNR) and the bottom row is simulated using **SNR** $= 2$ (high SNR).

Proportion of Training Data (PTD)

Recall that there are n training data $\{Z(\mathbf{u}) : \mathbf{u} \in D^{\text{trn}}\}$ and m validation data $\{Z(\mathbf{u}) : \mathbf{u} \in D^{\text{val}}\}$, where $D^{\text{trn}} \cap D^{\text{val}} = \emptyset$. Thus, there are a total of $m + n$ observations that need to be separated into training and validation datasets. The separation of observed data into training and validation datasets has been looked at by many (see, e.g., Amari et al., 1997; Guyon, 1997; Larsen and Goutte, 1999).

Let the proportion of training data be $\text{PTD} \equiv n/(n + m)$, where in the simulation $n + m = 1,241$ is held fixed. We consider two levels: $\text{PTD} = 0.4$ (labeled as **PTD** = 1) and $\text{PTD} = 0.6$ (labeled as **PTD** = 2). The incorporation of this factor does not address the problem of how to obtain D^{trn} and D^{val} for local spatial predictor selection; in the spatial setting, we randomly select the training and validation locations

from the set of observed data locations.

Missing-Data Structure (MDS)

The number of elements contained in H^{val} partially depends on the positioning of the validation locations in the spatial domain D. Thus, in our simulation study we consider two types of missing data over the spatial lattice D: missing in blocks (labeled as **MDS** = 1) and missing at random (labeled as **MDS** = 2), respectively. These correspond to two levels of the factor, "missing-data structure."

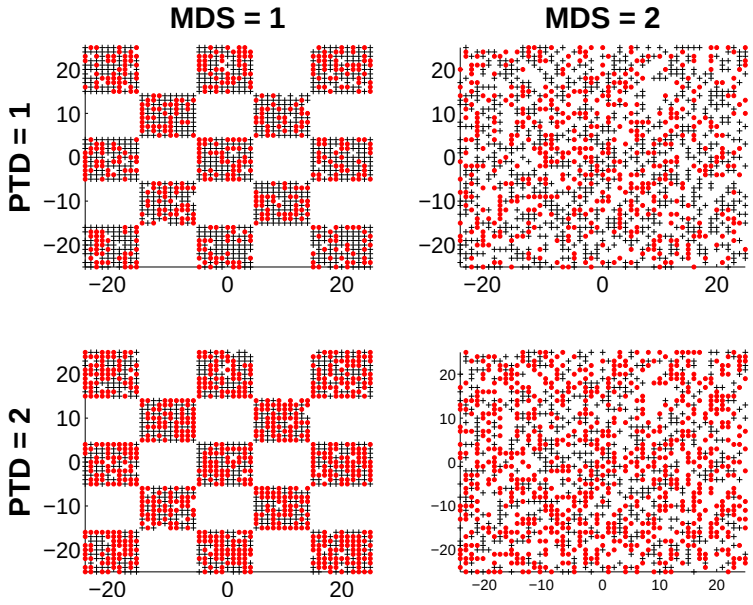

Figure 2: The training data locations D^{trn} are given by a black plus sign. The validation data locations D^{val} are given by a red circle. White indicates the spatial locations where no data were observed. The top row represents the case where $n/(n+m) = 0.6$ (**PTD** = 1) and the bottom row represents the case where $n/(n+m) = 0.4$ (**PTD** = 2). The two panels display two types of **MDS** to be considered in our simulation study, given by missing in blocks in the left column (**MDS** = 1) and missing at random in the right column (**MDS** = 2).

In Figure 2, spatial locations in D^{trn} are indicated by a black "+" and spatial locations in D^{val} are indicated by a red circle for each of the levels of **PTD**. When **PTD** = 1, we set $n = 720$ and $m = 521$ for both levels of **MDS**; and for **PTD** = 2, we set $n = 521$ and $m = 720$ for both levels of **MDS**.

The Locally Selected Spatial Predictor (LSP)

Recall that in Sections 2.2 and 2.3 we introduced the concept of local selection and four LSPs: LSP based on simple local predictor selection or SLS given by (12) (abbreviated as LSP-SLS), LSP based on moving

window predictor selection or MWS given by (15) (abbreviated as LSP-MWS), LSP based on nearest neighborhood predictor selection or NNS given by (16) (abbreviated as LSP-NNS), and LSP based on Voronoi polygon predictor selection or VPS given by (17) (abbreviated as LSP-VPS). In this section, our focus is on comparing the squared prediction errors of these four LSPs to the squared prediction errors of the GSP and that of the seven individual spatial predictors. Hence, we include **LSP** as a factor at four levels: LSP-SLS (labeled as **LSP** = 1), LSP-MWS (labeled as **LSP** = 2), LSP-NNS (labeled as **LSP** = 3), and LSP-VPS (labeled as **LSP** = 4). We compute the LSPs from the seven spatial predictors given in Section 3.1.

To compute LSP-MWS and LSP-NNS, we need to specify the window's radius w (see (15)) and the number of nearest neighbors g (see (16)), respectively. To calibrate these values for our simulation study, we consider the *average squared prediction error*,

$$\text{ASPE}\left(\hat{Y}^{\text{PRD}}\right) \equiv \frac{1}{N}\sum_{i=1}^{N}\left(Y(\mathbf{u}_i) - \hat{Y}^{\text{PRD}}(\mathbf{u}_i, \mathbf{Z}^{\text{trn}})\right)^2, \tag{21}$$

where recall that $\hat{Y}^{\text{PRD}}$ is a generic predictor. Notice that (21) depends on the hidden spatial process $Y(\cdot)$, which we want to predict, but in the simulation we know it.

We have 480 replications of $Y(\cdot)$. For each factor-level combination of **SNR**, **PTD**, and **MDS**, we compute the average (over the 480 replicates) $\text{ASPE}\left(\hat{Y}^{(\hat{k}(\cdot,\mathbf{Z}^{\text{trn}};H_2^{\text{val}}))}\right)$, for H_2^{val} that depends on the moving window's radius $w = 1, 2, ..., 75$. The calibrated value of w corresponds to the LSP-MWS that minimizes the average ASPE over w. Similarly, for each factor-level combination, we compute the average $\text{ASPE}\left(\hat{Y}^{(\hat{k}(\cdot,\mathbf{Z}^{\text{trn}};H_3^{\text{val}}))}\right)$, for H_3^{val} that depends on the number of nearest neighbors $g = 1, 2, ..., m$. The calibrated value of g corresponds to the LSP-NNS that minimizes the average ASPE over g. In Table 1, we report the calibrated values of the window's radius w and the calibrated values of the number of nearest neighbors g for each factor-level combination.

4.3 Assessment of LSPs

The ASPE is always positive and skewed, but a fourth-root transformation is a symmetrizing and variance-stabilizing transformation that makes the response suitable for a formal analysis of variance (ANOVA); see Cressie et al. (2010). We define the *paired squared error* (PSE) as,

$$\text{PSE}(\hat{Y}^{(\hat{k}(\cdot,\mathbf{Z}^{\text{trn}};H_b^{\text{val}}))}) \equiv \min\left\{\left(\text{ASPE}\left(\hat{Y}^{\text{PRD}}\right)\right)^{1/4}\right\} - \left(\text{ASPE}\left(\hat{Y}^{(\hat{k}(\cdot,\mathbf{Z}^{\text{trn}};H_b^{\text{val}}))}\right)\right)^{1/4}, \tag{22}$$

SNR	MDS	PTD	Calibrated w	Calibrated g
SNR = 1	**MDS** = 1	**PTD** = 1	48	699
	MDS = 1	**PTD** = 2	37	440
	MDS = 2	**PTD** = 1	2	6
	MDS = 2	**PTD** = 2	24	144
SNR = 2	**MDS** = 1	**PTD** = 1	41	645
	MDS = 1	**PTD** = 2	64	521
	MDS = 2	**PTD** = 1	2	1
	MDS = 2	**PTD** = 2	61	518

Table 1: The calibrated values of the window's radius w, used in MWS, and of the number of nearest neighbors g, used in NNS, as defined in Section 2.3. These calibrated values are used in Section 4.3.

where $b = 1, 2, 3,$ 4, and the minimum in (22) is taken over PRD = TSK, SSP, EDW, FRK, MPP, SPD, and LTK. Notice that if the average of PSE in (22) is positive over the 480 replications of $Y(\cdot)$, then the LSP based on H_b^{val} is considered "better than" any of the individual spatial predictors *and* the GSP.

Thus, PSE in (22), can be considered the response in a paired experiment between the LSPs and the global spatial predictors. Boxplots of $\text{PSE}(\hat{Y}^{(\hat{k}(\cdot,\mathbf{Z}^{\text{trn}};H_b^{\text{val}}))})$ in (22) over the 480 replicates are approximately symmetric (see Figure 3). Thus, a formal ANOVA is reasonable for this setting.

In Table 2, we present results of an ANOVA with up to two-way interactions between the factors defined in Section 4.2. Large F-statistics are highlighted in gray. We see that main effects for each factor in Section 4.2 corresponds to a large F-statistic, as do the **MDS** $\times$ **PTD**, **MDS** $\times$ **LSP**, and **PTD** $\times$ **LSP** interactions.

In Figure 4, we provide a main-effects plot associated with this ANOVA. Larger positive values of the paired squared error in (22) are more favorable since that implies that the LSP is better than each of the GSP and the individual spatial predictors. In the first panel of Figure 4, the PSE is larger when the SNR is larger, which is intuitively reasonable (see the discussion in Section 4.2). In the second panel of Figure 4, the PSE is larger when PTD = 0.4, which shows that local spatial prediction has an advantage when more resources are put into the validation data.

In the third panel of Figure 4, the PSE is larger when the data are missing at random, which has implications for spatial sampling design. Finally, in the fourth panel of Figure 4, LSP-MWS, LSP-NNS, and LSP-SLS appear to have similar values of positive PPEs, while LSP-VPS is a distant fourth. In fact, LSP-VPS has a main effect that in this simulation experiment is *negative*.

In Figure 5, we provide an interaction plot showing all the two-way interactions associated with this ANOVA. Recall that parallel lines imply no interaction; the interaction between PTD and MDS, MDS and

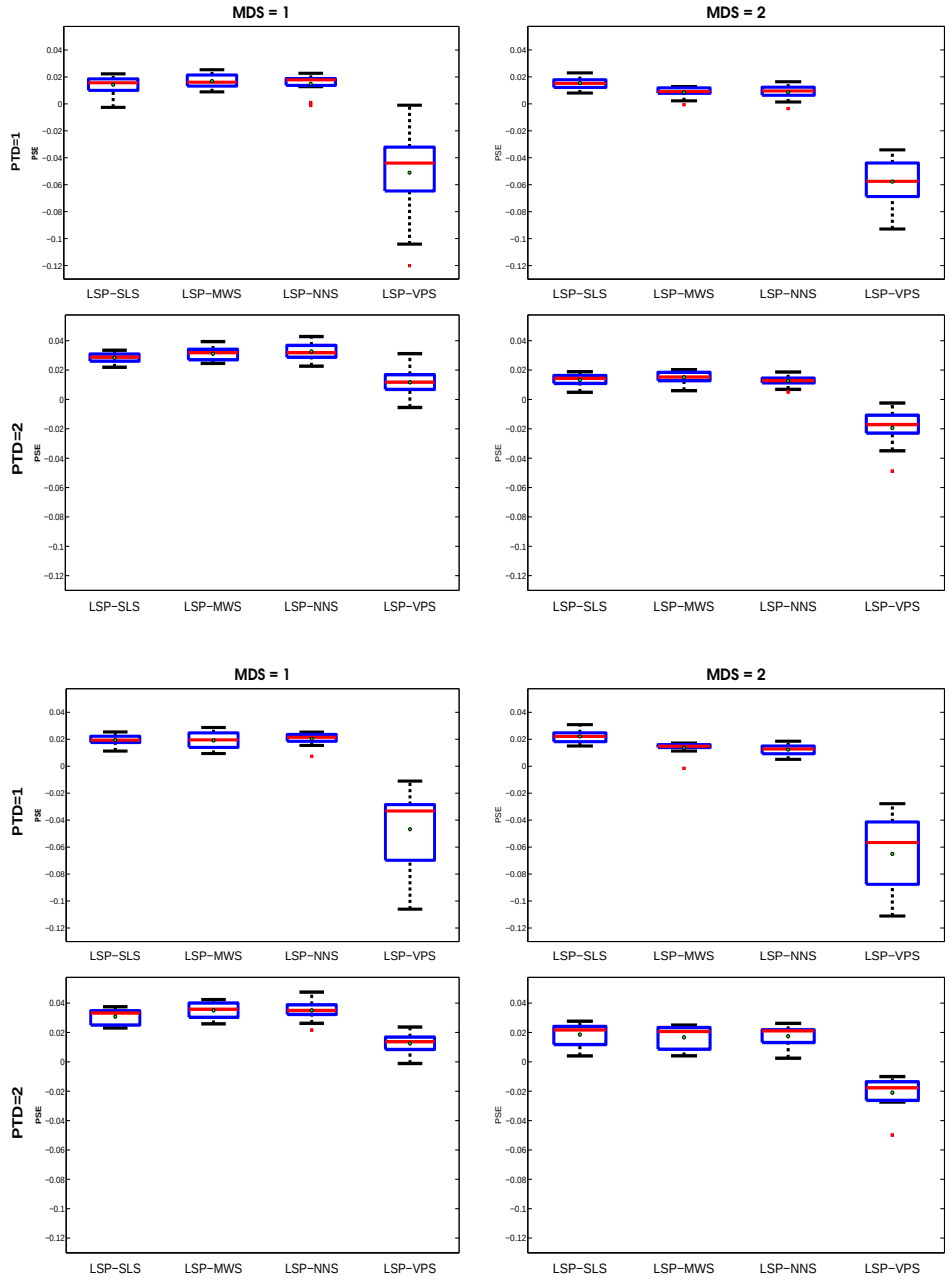

Figure 3: Boxplots of the PSE response variable in (22) by levels of **PTD** and **MDS**. The top (bottom) four boxplots have a signal-to-noise ratio fixed at SNR = 5 (SNR = 10). The left and right columns fix **MDS** $= 1$ and 2, respectively. The top and bottom rows fix **PTD** $= 1$ and 2, respectively. Sample averages are indicated by a green circle.

Source	DF	SS $\times 10^{-3}$	MS $\times 10^{-3}$	F
SNR	1	0.96	0.96	7.91
PTD	1	21.3	21.3	174.71
MDS	1	43.73	43.73	358.76
LSP	3	218.57	72.86	597.67
SNR $\times$ PTD	1	0.03	0.03	0.27
SNR $\times$ MDS	1	0.02	0.02	0.13
SNR $\times$ LSP	3	0.61	0.2	1.68
PTD $\times$ MDS	1	6.2	6.2	50.86
PTD $\times$ LSP	3	4.19	1.4	11.46
MDS $\times$ LSP	3	41.54	13.85	113.58
Residual	461	56.2	0.1219	
Total	479	393.35		

Table 2: Analysis of variance (ANOVA) on $\mathrm{PSE}(\hat{Y}^{(\hat{k}(\mathbf{u},\mathbf{Z}^{\mathrm{trn}};H_b^{\mathrm{val}}))})$, for up to two-way interactions. In the table, the column Source indicates the source of variability used in the ANOVA; DF denotes degrees of freedom; SS denotes sum of squared error; MS denotes mean squared error; and F denotes the F-statistic associated with the Source. Large F-statistics are highlighted in gray. There are a total of 32 factor-level combinations.

LSP, and PTD and LSP that we saw in Table 2 is visualized in the appropriate parts of Figure 5. Clearly, LSP-VPS does very poorly when the data are missing at random. Also, none of the line plots crosses over, so the conclusions from Figure 5 hold consistently at all levels of interacting factors.

In the next section we demonstrate use of the LSP on a very large spatial dataset of mid-tropospheric carbon dioxide.

5 Application: Global Mid-Tropospheric Carbon Dioxide

We demonstrate local predictor selection using a very large spatial dataset of 74,367 measurements of global mid-tropospheric CO_2 in parts per million (ppm). The data were obtained between -60° and 90° latitude from Februrary 1 through Februrary 9, 2010, by the Atmospheric InfraRed Sounder (AIRS) instrument on board the Aqua satellite (Chahine et al., 2006); no data were released below -60° latitude. The AIRS instrument collects measurements in the form of spectra and requires the use of retrieval algorithms to convert them into mid-tropospheric CO_2 values in ppm (e.g., Landgrebe, 2003; Chahine et al., 2006).

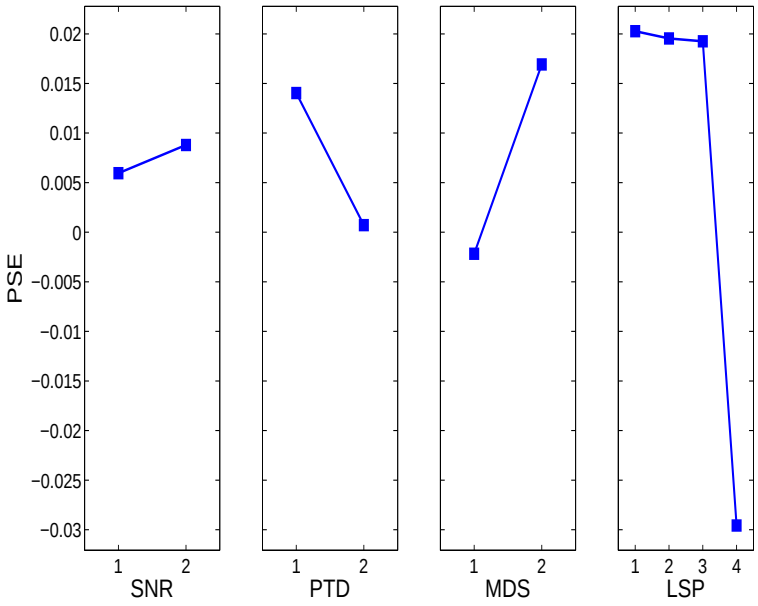

Figure 4: Main-effects plots for the ANOVA of Table 2, using PSE in (22) as the response and **SNR**, **PTD**, **MDS**, and **LSP** from Section 4.2 as the factors.

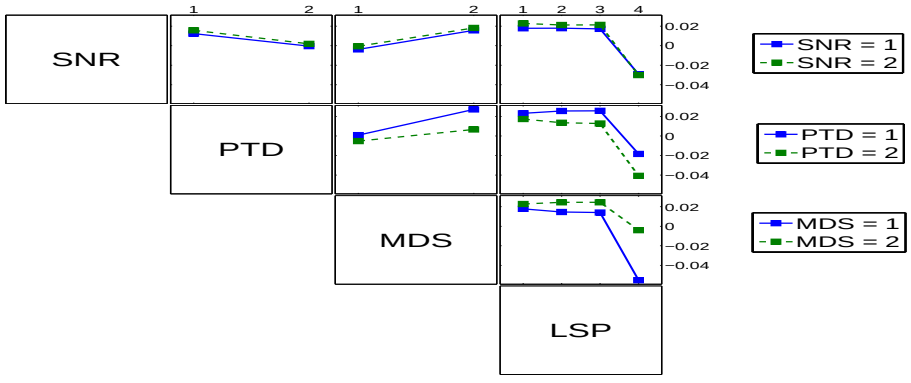

Figure 5: Two-way interaction plots for the ANOVA of Table 2, using PSE in (22) as the response and **SNR**, **PTD**, **MDS**, and **LSP** from Section 4.2 as the factors.

5.1 Testing Data

To determine the H^{val} we finally use, we set aside a further q observations to compare LSPs based on different choices for H^{val}. We refer to these as *testing data*, whose locations are denoted as $D^{\text{tst}} \equiv \{\mathbf{s}_j^{\text{tst}} : j = 1, ..., q\}$, where $D^{\text{tst}} \cap D^{\text{val}} = D^{\text{tst}} \cap D^{\text{trn}} = \emptyset$. Hence, the total size of the dataset is $n + m + q$, which in our case is 74,367. Define the root average squared testing error associated with the predictor $\hat{Y}^{\text{PRD}}$ as,

$$\text{RSTE}(\hat{Y}^{\text{PRD}}) \equiv \left(\frac{1}{q} \sum\nolimits_{j=1}^{q} (Z(\mathbf{s}_j^{\text{tst}}) - \hat{Y}^{\text{PRD}}(\mathbf{s}_j^{\text{tst}}, \mathbf{Z}^{\text{trn}}))^2 \right)^{1/2}.$$

In Section 5.2, our selection of H^{val} will be based on minimizing $\text{RSTE}(\hat{Y}^{\text{PRD}})$.

Recall that our dataset consists of 74,367 observations. The data were randomly split into training, validation, and testing datasets with $n = 44,621$ (60% of the data), $m = 27,746$ (38% of the data), and $q = 2,000$ (2% of the data), respectively. The training, validation, and testing datasets are displayed in Figure 6. We need to predict $Y(\cdot)$ at D^{val} and D^{tst}, and we impose a regular grid on the globe that is 1.25 degrees longitude by 1 degree latitude. The nodes of the regular grid are the prediction locations that will form our spatial map, and hence $D = D^{\text{trn}} \cup D^{\text{val}} \cup D^{\text{tst}} \cup \{\mathbf{u} = (i, j) : i = -60, -59, ..., 89, 90, j = -180, -178.75, ..., 178.75, 180\}$.

5.2 Local Spatial-Predictor Selection for Mid-Tropospheric CO_2

In this section, we perform local predictor selection based on the four spatial predictors from Section 3.1 that are computationally efficient for very large spatial datasets, namely EDW, FRK, SPD, and LTK. Notice that MPP with a comparable number of knots to FRK was not able to handle the computations required for this example, and it is well known that TSK and SSP are not scalable to larger data sizes.

Using the training data displayed in the top panel of Figure 6, we calculated the four spatial predictors, $\hat{Y}^{\text{EDW}}$, $\hat{Y}^{\text{FRK}}$, $\hat{Y}^{\text{SPD}}$, and $\hat{Y}^{\text{LTK}}$. All of our computations were performed on a dual quad core 2.8 GHz 2x Xeon X5560 processor, with 96 Gbytes of memory. In Figure 7, we see that spatial predictions of mid-tropospheric CO_2 based on SPD suggest that $Y(\cdot)$ is progressively smaller (larger) for negative (positive) latitudes. Predictions of mid-tropospheric CO_2 using EDW, FRK, and LTK show the same pattern, but the predicted surfaces are less smooth than that for SPD.

In the left column of Figure 8, maps are given of the optimal $\hat{k}(\cdot)$ given by (12), for simple local predictor selection (H_1^{val}), moving-window predictor selection with $w = 5^\circ$ (H_2^{val}), nearest-neighborhood predictor

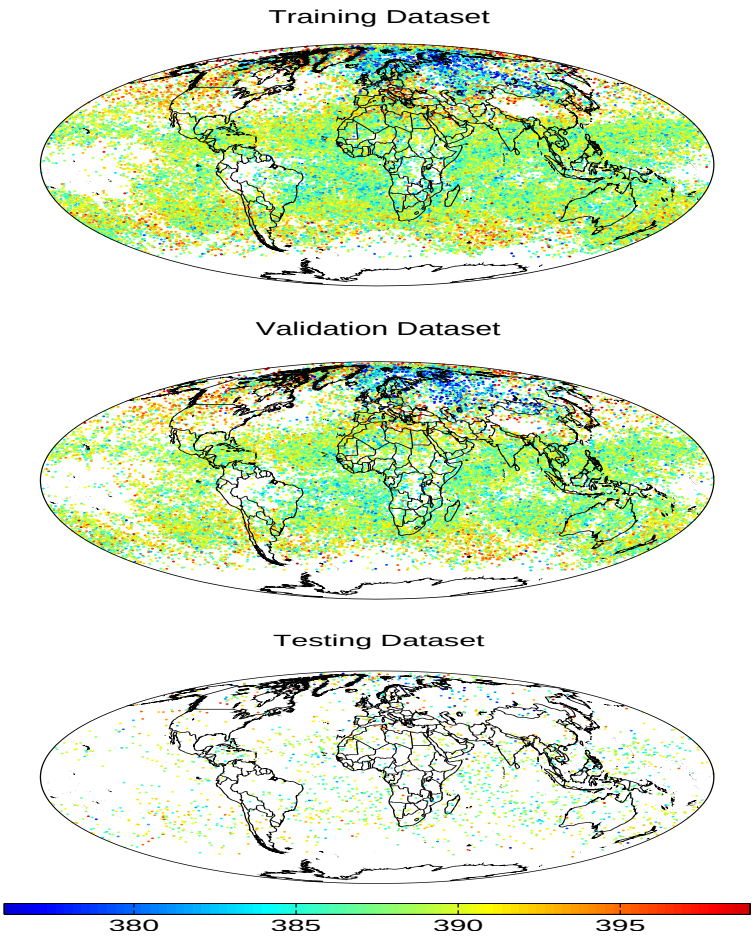

Figure 6: A spatial dataset made up of of 9 days of measurements of global mid-tropospheric CO_2 in parts per million (ppm). The data were collected between -60° degrees and 90° degrees latitude from Februrary 1 through Februrary 9, 2010. The data were then randomly split into training, validation, and testing datasets with $m = 27,746$ (38% of the data), $n = 44,621$ (60% of the data), and $q = 2,000$ (2% of the data), respectively.

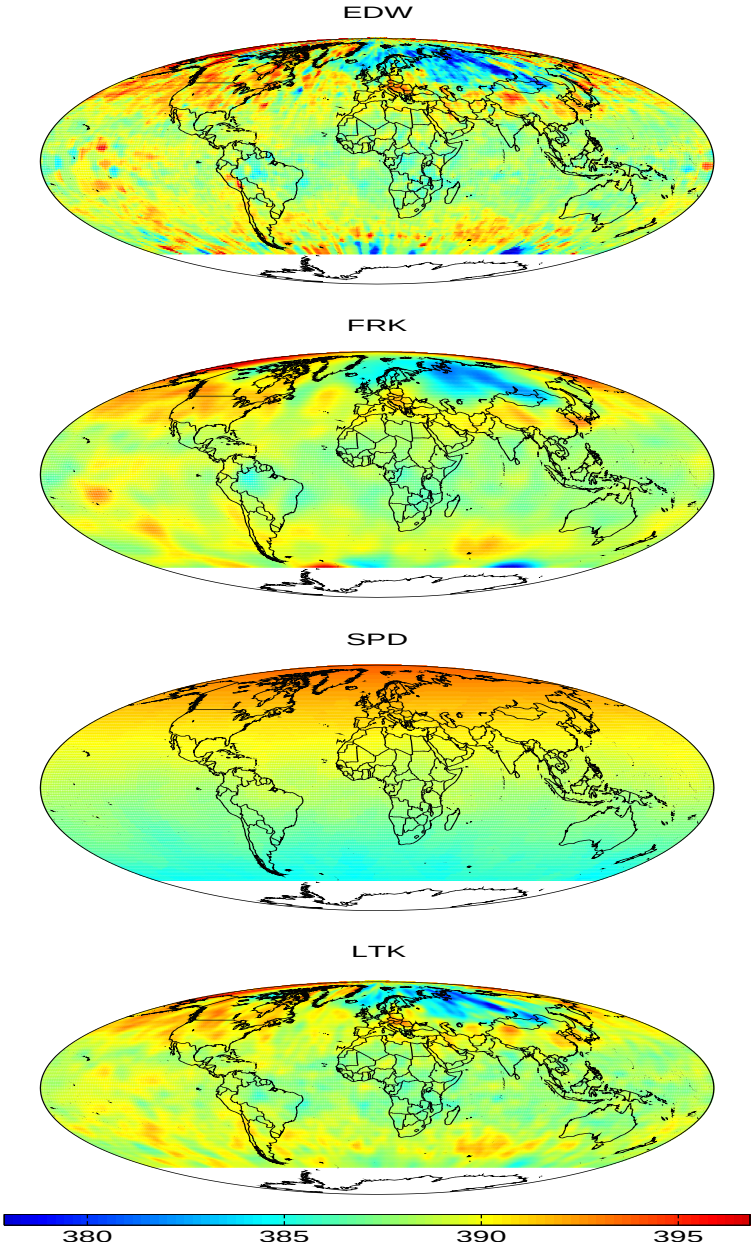

Figure 7: Spatial prediction of mid-tropospheric CO_2 concentrations using EDW, FRK, SPD, and LTK. Note that we do not predict below latitude $-60°$, since AIRS has not released any observations there.

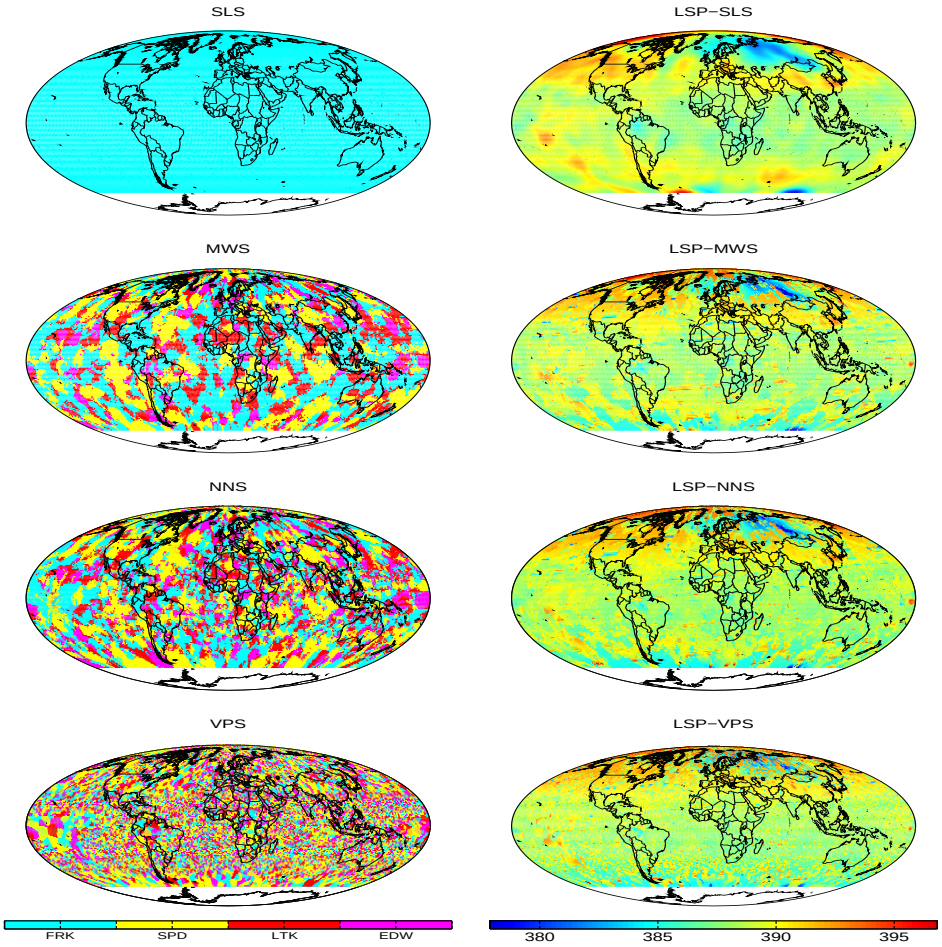

Figure 8: In the left column, the selected predictor is reported location-by-location. The colors represent the predictor that was selected at the nodes of the regular grid for the AIRS CO_2 dataset. Each panel represents a different choice of $H^{\text{val}}(\cdot)$, which is indicated by its title. Note that no predictions are made below latitude -60°, since AIRS has not released any observations there. In the right column, the LSPs are reported; each panel represents a different choice of $H^{\text{val}}(\cdot)$. From top to bottom: simple local (SLS), moving-window (MWS), nearest-neighborhood (NNS), and Voronoi-polygon (VPS) predictor selection.

selection with $q = 15$ (H_3^{val}), and Voronoi-polygon predictor selection (H_4^{val}). After inspecting the maps, there does not appear to be a single value of $k \in \{1, 2, 3, 4\}$ that dominates over any specific region on the globe. When using moving-window predictor selection, FRK is selected at 34.16% of the prediction locations, SPD is selected at 33.15% of the prediction locations, LTK is selected at 18.57% of the prediction locations, and EDW is selected at 14.12% of the prediction locations. When using nearest-neighborhood predictor selection SPD is selected at 35.89% of the prediction locations, FRK is selected at 29.74% of the prediction locations, LTK is selected at 17.68% of the prediction locations, and EDW is selected at 16.70% of the prediction locations. Compare this to the GSP, which selects FRK.

The maps of LSPs are given in the right column of Figure 8. The LSP (in each panel) displays a pattern closer to that of EDW, FRK, and LTK, than to SPD shown in Figure 7; however, the prediction map is less smooth when using LSP than when using any of the four individual spatial predictors. This suggests that each of the individual spatial predictors may be too smooth and LSP can be used to identify this.

In Table 3, the *root average squared testing error* (i.e., RSTE given in Section 5.1) is reported for EDW, FRK, SPD, LTK, and the four LSPs based on H^{val} given by SLS, MWS, NNS, and VPS. Each LSP (except VPS) appears to be performing as well or better than the individual spatial predictors at the testing locations (D^{tst}). The smallest value of RSTE corresponds to LSP-MWS, suggesting that moving-window selection is the most appropriate selection method. Based on the RSTE in Table 3, FRK appears to be performing the best among the four individual spatial predictors; however, the incorporation of EDW, SPD, and LTK with FRK using our LSP approach leads to an improvement in predictions.

Predictor	RSTE
EDW	4.2090
FRK	3.9920
SPD	4.2780
LTK	4.0250
GSP	3.9920
LSP-SLS	3.9920
LSP-MWS	3.9652
LSP-NNS	3.9887
LSP-VPS	4.1520

Table 3: The root average squared testing error (RSTE) defined in Section 5.1, for individual spatial predictors, GSP, and LSPs, applied to the global mid-tropospheric CO_2 dataset. Moving-window predictor selection (LSP-MWS) is performed with radius $w = 5$ degrees. Nearest-neighborhood predictor selection (LSP-NNS) is performed with $g = 15$.

6 Discussion

In this paper, we propose criteria to compare and select from a finite number of spatial predictors defined on a (possibly irregular) spatial lattice $D \equiv \{\mathbf{u}_1, ..., \mathbf{u}_N\} \subset \mathbb{R}^d$. The criteria are empirically based, reflecting the nonparametric assumptions in our model and the large datasets available to us. Further, our approach allows us to see at which locations an individual spatial predictor is performing well in comparison to the other predictors. This leads naturally to the idea of selecting predictors optimally, location-by-location; the result is a new, hybrid spatial predictor, which we call a locally selected predictor (LSP).

We have developed the notion of local prediction selection using local squared validation error. We have shown that an obvious LSP exists at validation locations, and we have provided a variety of ways to select predictors at locations that are not used for validation, which includes simple local predictor selection, moving-window predictor selection, nearest-neighborhood predictor selection, and Voronoi-polygon predictor selection.

Our simulations in Section 4 were based on a full-rank random-effects model using Fourier spatial basis functions, but the predictors from which optimal selections were made came from a variety of types. Using the average (over simulations) of the average (over all locations in D) squared prediction error as our criterion, we demonstrated that LSPs generally outperform the globally selected predictor (GSP).

We also demonstrated how to implement LSP in practice using a very large spatial dataset of mid-tropospheric CO_2. Furthermore, the root average squared testing error indicates that one can obtain improvements in spatial predictions when using moving-window predictor selection. In this example, the data suggest that fixed rank kriging (FRK) and the stochastic partial-differentiation approach (SPD) perform comparably and are better at more prediction locations than the negative-exponential-distance weighting approach (EDW) and the lattice-kriging approach (LTK). However, an LSP that incorporates FRK, EDW, SPD, and LTK results in an even better spatial predictor.

Estimates of the mean squared prediction error for the LSPs is an important problem that has not been developed in this paper. A simple statistic one might use to quantify the uncertainty of the LSPs is $\mathrm{LSVE}^{(\hat{k})}$ from (5). Burnham and Anderson (1998) suggest obtaining mean squared prediction errors using the statistical model behind the globally selected predictor, however this is likely to lead to overly optimistic values. Furthermore, this will not work if one of the spatial predictors (e.g., EDW) does not come with a measure of prediction uncertainty. In line with the methodology presented in this paper and the concept of model

averaging, we expect that an appropriate quantification of the uncertainty of the LSP-MWS (for example) predictor, would be based on

$$\sum_{k=1}^{K} w_k(\mathbf{u})\mathrm{LSVE}^{(k)}(\mathbf{u}, \mathbf{Z}^{\mathrm{trn}};\ H_2^{\mathrm{val}});\ \ \mathbf{u} \in D,$$

where $\{w_k(\mathbf{u}) : k = 1, ..., K\}$ are local non-negative weights that sum to 1, and $\mathrm{LSVE}^{(k)}$ is given by (5).

A more parametric approach to predictor selection is to use information criteria (e.g, Bradley et al., 2012), which avoids having to divide the data up for different purposes, but it involves making more assumptions. Spatial-predictor selection using this approach will be reported on elsewhere.

Acknowledgments

Jonathan Bradley's and Noel Cressie's research was supported by NASA's Earth Science Technology Office through its Advanced Information Systems Technology program Grant NNH08ZDA001N. Tao Shi's research is partially supported by NSF grants DMS-1007060 and DMS-1308458.

7 Appendix: Technical Result

In this appendix, we prove the technical result stated in (10).

Proposition: Assume (1) and (9). Then (10) holds, namely,

$$E\left(\mathrm{SPE}^{(\hat{k})}(\mathbf{s}_j^{\mathrm{val}}, \mathbf{Z}^{\mathrm{trn}})\right) = E\left(\mathrm{SPE}^{(\tilde{k})}(\mathbf{s}_j^{\mathrm{val}}, \mathbf{Z}^{\mathrm{trn}})\right);\ j = 1, ..., m.$$

Proof: By the definition of $\tilde{k}(\mathbf{s}_j^{\mathrm{val}}, \mathbf{Z}^{\mathrm{trn}}; H^{\mathrm{val}})$ in (3), we have

$$\mathrm{SPE}^{(\tilde{k})}\left(\mathbf{s}_j^{\mathrm{val}}, \mathbf{Z}^{\mathrm{trn}}\right) \leq \mathrm{SPE}^{(\hat{k})}\left(\mathbf{s}_j^{\mathrm{val}}, \mathbf{Z}^{\mathrm{trn}}\right); j = 1, ..., m.$$

Taking the expected value, we have

$$E\left(\mathrm{SPE}^{(\tilde{k})}\left(\mathbf{s}_j^{\mathrm{val}}, \mathbf{Z}^{\mathrm{trn}}\right)\right) \leq E\left(\mathrm{SPE}^{(\hat{k})}\left(\mathbf{s}_j^{\mathrm{val}}, \mathbf{Z}^{\mathrm{trn}}\right)\right); j = 1, ..., m. \tag{23}$$

Now H^{val} is given in (9) and hence by the definition of $\hat{k}(\mathbf{s}_j^{\text{val}}, \mathbf{Z}^{\text{trn}}; H^{\text{val}})$ in (6), we have

$$(Z(\mathbf{s}_j^{\text{val}}) - \hat{Y}^{(\hat{k})}(\mathbf{s}_j^{\text{val}}, \mathbf{Z}^{\text{trn}}))^2 \leq (Z(\mathbf{s}_j^{\text{val}}) - \hat{Y}^{(\tilde{k})}(\mathbf{s}_j^{\text{val}}, \mathbf{Z}^{\text{trn}}))^2; j = 1, ..., m. \tag{24}$$

Substituting the equation, $Z(\mathbf{s}_j^{\text{val}}) = Y(\mathbf{s}_j^{\text{val}}) + \epsilon(\mathbf{s}_j^{\text{val}})$, into (24) and canceling terms, we have

$$\begin{aligned}&(Y(\mathbf{s}_j^{\text{val}}) - \hat{Y}^{(\hat{k})}(\mathbf{s}_j^{\text{val}}, \mathbf{Z}^{\text{trn}}))^2 + 2\epsilon(\mathbf{s}_j^{\text{val}})(Y(\mathbf{s}_j^{\text{val}}) - \hat{Y}^{(\hat{k})}(\mathbf{s}_j^{\text{val}}, \mathbf{Z}^{\text{trn}})) \\ &\leq (Y(\mathbf{s}_j^{\text{val}}) - \hat{Y}^{(\tilde{k})}(\mathbf{s}_j^{\text{val}}, \mathbf{Z}^{\text{trn}}))^2 + 2\epsilon(\mathbf{s}_j^{\text{val}})(Y(\mathbf{s}_j^{\text{val}}) - \hat{Y}^{(\tilde{k})}(\mathbf{s}_j^{\text{val}}, \mathbf{Z}^{\text{trn}})); j = 1, ..., m.\end{aligned} \tag{25}$$

Now,

$$E\left(\epsilon(\mathbf{s}_j^{\text{val}})(Y(\mathbf{s}_j^{\text{val}}) - \hat{Y}^{(\hat{k})}(\mathbf{s}_j^{\text{val}}, \mathbf{Z}^{\text{trn}}))\right) = E\left(\epsilon(\mathbf{s}_j^{\text{val}})\right) E\left((Y(\mathbf{s}_j^{\text{val}}) - \hat{Y}^{(\hat{k})}(\mathbf{s}_j^{\text{val}}, \mathbf{Z}^{\text{trn}}))\right) = 0,$$

since $(Y(\mathbf{s}_j^{\text{val}}) - \hat{Y}^{(\hat{k})}(\mathbf{s}_j^{\text{val}}, \mathbf{Z}^{\text{trn}}))$ is a function of $\{Y(\mathbf{s}_j^{\text{val}}), \mathbf{Z}^{\text{trn}}\}$ and $\epsilon(\mathbf{s}_j^{\text{val}})$ is independent of $\{Y(\mathbf{s}_j^{\text{val}}), \mathbf{Z}^{\text{trn}}\}$. Similarly, $E\left(\epsilon(\mathbf{s}_j^{\text{val}})(Y(\mathbf{s}_j^{\text{val}}) - \hat{Y}^{(\tilde{k})}(\mathbf{s}_j^{\text{val}}, \mathbf{Z}^{\text{trn}}))\right) = 0$. Thus, when taking the expected value of (25), we have for $j = 1, ..., m$,

$$E\left((Y(\mathbf{s}_j^{\text{val}}) - \hat{Y}^{(\hat{k})}(\mathbf{s}_j^{\text{val}}, \mathbf{Z}^{\text{trn}}))^2\right) \leq E\left((Y(\mathbf{s}_j^{\text{val}}) - \hat{Y}^{(\tilde{k})}(\mathbf{s}_j^{\text{val}}, \mathbf{Z}^{\text{trn}}))^2\right). \tag{26}$$

Combining the inequalities in (23) and (26), we obtain (10), which is the desired result.

8 Bibliography

Akaike, H. (1974). "A new look at the statistical model identification." *IEEE Transactions on Automatic Control*, 19, 716–723.

Amari, S., Murata, N., Finke, N. M. K. R., and Yang, H. H. (1997). "Asymptotic statistical theory of overtraining and cross-validation." *IEEE Transactions on Neural Networks*, 8, 985–996.

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. 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.

— (2012). "Local spatial-predictor selection." In *Proceedings of the 2012 Joint Statistical Meetings*, 3098 – 3110. Alexandria, VA: American Statistical Association.

Burnham, K. P. and Anderson, D. R. (1998). *Model Selection and Multimodel Inference*, 2nd edn. New York, NY: Springer.

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.

Chen, C. S. and Huang, H. C. (2011a). "Geostatistical model averaging based on conditional information criteria." *Environmental and Ecological Statistics*, 19, 23–35.

Chen, C.-S. and Huang, H.-C. (2011b). "An improved Cp criterion for spline smoothing." *Journal of Statistical Planning and Inference*, 141, 445–452.

Chen, Y.-P., Huang, H.-C., and Tu, I.-P. (2010). "A new approach for selecting the number of factors." *Computational Statistics and Data Analysis*, 54, 2990–2998.

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. (2010). " 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.

Donoho, D. and Johnstone, I. (1994). "Ideal spatial adaptation by wavelet shrinkage." *Biometrika*, 81, 425–455.

Efron, B. (1983). "Estimating the error rate of a prediction rule: Improvement on cross-validation." *Journal of the Americal Statistical Association*, 78, 316–331.

— (1986). "How biased is the apparent error rate of a prediction rule?" *Journal of the Americal Statistical Association*, 81, 461–470.

— (2004). "The estimation of prediction error: Covariance penalties and cross-validation." *Journal of the Americal Statistical Association*, 99, 619–642.

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.

Greven, S. and Kneib, T. (2010). " On the behaviour of marginal and conditional AIC in linear mixed models." *Biometrika*, 97, 773–789.

Guyon, I. (1997). "A scaling law for the validation-set training-set size ratio." Tech. rep., Bell Laboratories, Berkeley, CA.

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

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

Huang, H. C. and Chen, C. S. (2007). " Optimal geostatistical model selection." *Journal of the American Statistical Association*, 102, 1009–1024.

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

— (2011b). "Tutorial on Fixed Rank Kriging (FRK) of CO2 data." Tech. rep., Report No. 858, Department of Statistics, The Ohio State University, Columbus, OH, http://www.stat.osu.edu/∼sses/papers.html.

Konishi, S. and Kitagawa, G. (1996). "Generalised information criteria in model selection." *Biometrika*, 83, 875–890.

Kullback, S. and Leibler, R. A. (1951). "On information and sufficiency." *Annals of Mathematical Statistics*, 22, 79 – 86.

Lai, R., Huang, H.-C., and Lee, T. (2012). "Fixed and random effects selection in nonparametric additive mixed models." *Electronic Journal of Statistics*, 6, 810–842.

Landgrebe, D. A. (2003). *Signal Theory Methods in Multispectral Remote Sensing*. Hoboken, NJ: Wiley.

Larsen, J. and Goutte, C. (1999). "On optimal data split for generalization estimation and model selection." In *Proceedings IEEE Workshop on Neural Networks for Signal Processing*, 225–234. New York, NY: IEEE Press.

Liang, H., Wu, H., and Zou, G. (2008). "A note on conditional AIC for linear mixed-effects models." *Biometrika*, 95, 773–778.

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.

Mallows, C. L. (1973). "Some comments on Cp." *Technometrics*, 15, 661–675.

Matheron (1963). "Principles of geostatistics." *Economic Geology*, 58, 1246–1266.

Muller, Scealy, and Welsh (2013). "Model selection in linear mixed models." *Statistical Science*, 28, 135 – 167.

Nychka, D., Hammerling, D., Sain, S., and Lerud, T. (2012). "Package 'LatticeKrig'." http://cran.r-project.org/web/packages/LatticeKrig/LatticeKrig.pdf. Retrieved November, 2012.

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.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. New York, NY: Press Syndicate of the University of Cambridge.

Ronchetti, E. (1997). "Robustness aspects of model choice." *Statistica Sinica*, 7, 327–338.

Ronchetti, E. and Staudte, R. (1994). "A robust version of Mallow's Cp." *Journal of the American Statistical Association*, 89, 550–559.

Royle, J. A. and Wikle, C. K. (2005). "Efficient statistical mapping of avian count data." *Environmental and Ecological Statistics*, 12, 225–243.

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.

Schwarz, G. E. (1978). "Estimating the dimension of a model." *Annals of Statistics*, 6, 461 – 464.

Shao, J. (1997). "An asymptotic theory for linear model selection." *Statistica Sinica*, 7, 221–264.

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, C. (1981). "Estimation of the mean of the multivariate normal distribution." *Annals of Statistics*, 9, 1135–1151.

Vaida, F. and Blanchard, S. (2005). “Conditional Akaike information for mixed-effects models.” *Biometrika*, 92, 351 – 370.

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

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

Zhu, J., Huang, H.-C., and Reyes, P. (2010). “On selection of spatial linear models for lattice data.” *Journal of the Royal Statistical Society Series B*, 72, 389–402.