

# Spatial Statistical Data Fusion for Remote Sensing Applications

Hai NGUYEN, Noel CRESSIE, and Amy BRAVERMAN

Aerosols are tiny solid or liquid particles suspended in the atmosphere; examples of aerosols include windblown dust, sea salts, volcanic ash, smoke from wildfires, and pollution from factories. The global distribution of aerosols is a topic of great interest in climate studies since aerosols can either cool or warm the atmosphere depending on their location, type, and interaction with clouds. Aerosol concentrations are important input components of global climate models, and it is crucial to accurately estimate aerosol concentrations from remote sensing instruments so as to minimize errors “downstream” in climate models. Currently, space-based observations of aerosols are available from two remote sensing instruments on board NASA’s Terra spacecraft: the Multiangle Imaging SpectroRadiometer (MISR), and the MODerate-resolution Imaging Spectrometer (MODIS). These two instruments have complementary coverage, spatial support, and retrieval characteristics, making it advantageous to combine information from both sources to make optimal inferences about global aerosol distributions.

In this article, we predict the true aerosol process from two noisy and possibly biased datasets, and we also estimate the uncertainties of these estimates. Our data-fusion methodology scales linearly and bears some resemblance to Fixed Rank Kriging (FRK), a variant of kriging that is designed for spatial interpolation of a single, massive dataset. Our spatial statistical approach does not require assumptions of stationarity or isotropy and, crucially, allows for change of spatial support. We compare our methodology to FRK and Bayesian melding, and we show that ours has superior prediction standard errors compared to FRK and much faster computational speed compared to Bayesian melding.

KEY WORDS: Aerosol optical depth; Fixed rank kriging; Geostatistics; Prediction standard error.

## 1. INTRODUCTION

Scientific advances in the study of climate are accelerating at a fast pace thanks in part to new technologies for collecting data about Earth’s atmosphere, oceans, ice sheets, and land ecosystems, and for computing estimates from these data. A prime example is the emergence of remote sensing observations from satellites as a vast, largely untapped resource for improving our understanding of these complex, interacting geophysical processes. One of the most important uses of these data is as input to climate models to represent processes that are not understood well enough to describe dynamically. In these cases, empirical characterizations called “climatologies” are inputted into models instead of letting the models internally generate their own estimates. The accuracies of climate predictions depend on the availability of good climatologies describing behavior of key variables and their uncertainties.

Among the variables that are important for accurate climate simulations are atmospheric aerosol properties. An aerosol is a suspension of fine solid particles or of liquid droplets in a gas. In the atmosphere, they may arise from natural or anthropogenic causes. Natural aerosols include sea salt, windblown dust, and volcanic ash. Anthropogenic aerosols include smoke and ash from land-clearing fires and pollutants from manufacturing. Aerosol climatologies provide information on where

and when aerosols are injected into (sources) and removed from (sinks) the atmosphere, and this is important because aerosols affect the climate system in at least two ways. First, they can create strong backscattering of solar radiation and absorb light, cooling the planet. This is known as the aerosol direct effect and, according to the Intergovernmental Panel on Climate Change (IPCC), this cooling is about  $-0.7 \text{ W/m}^2$ , compared to the total global warming effect of  $2.5 \text{ W/m}^2$  (IPCC 2001). Second, aerosols can catalyze the formation of clouds by providing condensation nuclei, thus warming the planet because clouds form a blanket that holds heat in (Lohmann and Feichter 2005). This is known as the aerosol indirect effect. The way these effects are represented in climate models is a large source of uncertainty in model simulations and predictions. It is crucial that information on sources and sinks be accurate so that modelers can concentrate on the representation of transport and effects of aerosols on other components of the climate system.

Two instruments on board NASA’s Terra satellite have been collecting information about aerosols in the atmosphere since it was launched into polar orbit in December 1999. The Multiangle Imaging SpectroRadiometer (MISR; Diner et al. 1998) and the MODerate-resolution Imaging Spectrometer (MODIS; King et al. 1992) both observe reflected solar radiation and use this information to estimate the global distribution of aerosols. The total amount of aerosol at any location and time is given by a quantity called aerosol optical depth (AOD), which measures the attenuation of solar energy passing through a column of atmosphere from the top of atmosphere to the ground. AODs usually range between 0 and 1; an AOD of less than 0.1 corresponds to a crystal-clear scene with maximum visibility, while an AOD near 1 indicates very hazy conditions. AOD is measured at different

Hai Nguyen is Statistician, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91125 (E-mail: [hai.nguyen@jpl.nasa.gov](mailto:hai.nguyen@jpl.nasa.gov)). Noel Cressie is Director of the Program in Spatial Statistics and Environmental Statistics and Distinguished Professor, Department of Statistics, The Ohio State University, Columbus, OH 43210 (E-mail: [ncressie@stat.osu.edu](mailto:ncressie@stat.osu.edu)). Amy Braverman is Senior Statistician, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91125 (E-mail: [Amy.J.Braverman@jpl.nasa.gov](mailto:Amy.J.Braverman@jpl.nasa.gov)). The research described in this article began as part of Nguyen’s PhD dissertation in the Department of Statistics at UCLA, and was continued at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA. It is supported by NASA’s Earth Science Technology Office through its Advanced Information Systems Technology program. We would like to express our appreciation for all comments received during the review process.

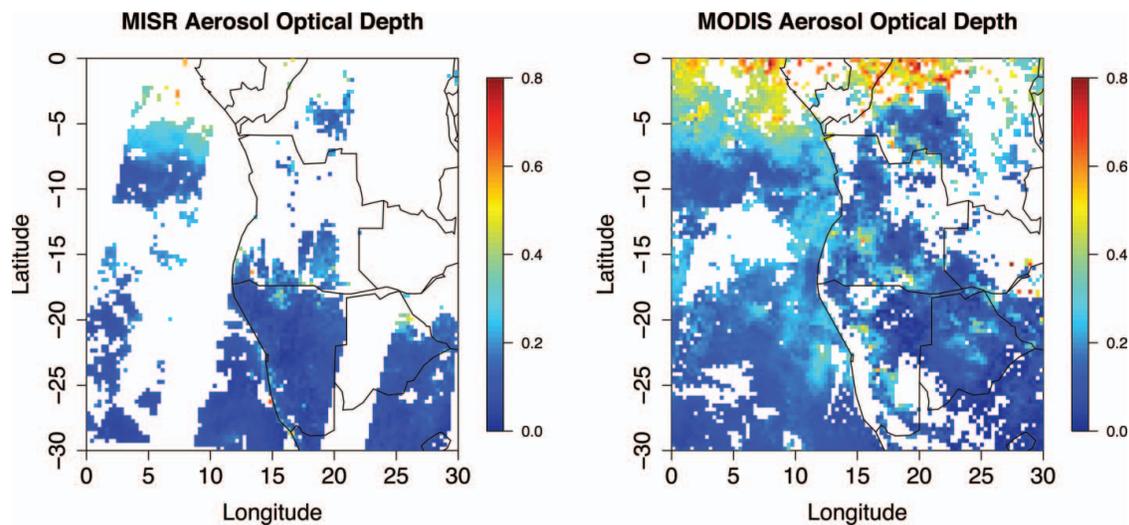


Figure 1. MISR and MODIS AOD, January 1–16, 2001, over southern Africa. MISR footprints are at 17.6 km<sup>2</sup> resolution, and MODIS footprints are at 10 km<sup>2</sup> resolution; for display purposes in the figure, both datasets are averaged to one-eighth degree resolution. White pixels are those for which no reading is obtained from the instrument.

wavelengths, and here we use AOD at 0.55 μm (green band), which is commonly used for aerosol climatologies.

Both instruments provide relatively high-resolution data on aerosol amount and type, but they do so using different technologies and data-processing algorithms. MISR collects data in four visible bands (blue, green, red, and near infrared) and from nine different viewing angles along Terra’s ground track. With its 300 km cross-track field of view, MISR provides complete coverage of the Earth every 9 days on 17.6 km<sup>2</sup> spatial regions called footprints. MODIS provides measurements in 36 spectral bands from a single-view angle at 10 km<sup>2</sup> resolution and, with its 1500 km cross-track field of view, MODIS has a global coverage rate of about 2 days. Since they use different technologies and make different assumptions, the instruments have different strengths and weaknesses that can be exploited. For instance, MODIS generally has more complete coverage, but it cannot see over bright surfaces such as deserts or glint over large bodies of water. MISR generally has better measurement-error characteristics (see Paradise et al. 2007). We seek to produce better aerosol climatologies by combining data from the two instruments in a way that capitalizes on their strengths; we call our approach *Spatial Statistical Data Fusion* (SSDF).

### 1.1 Aerosols in Southern Africa

In this article, we examine MISR and MODIS AOD data in a 30° × 30° region of southern Africa, from 30°S latitude to the equator (0° latitude) and from the prime meridian (0° longitude) to 30°E longitude, observed from January 1–16, 2001 (see Figure 1). We choose this area because it produces a lot of aerosols due to agricultural land-clearing and its proximity to the Sahara Desert. During this time of year, these aerosols are transported southward and westward over the Atlantic ocean, eventually reaching South America. Because these sources and transport are to be injected into climate models, it is very important to accurately characterize AOD over this region.

Within this chosen spatio-temporal domain, Terra made three overpasses; MISR acquired 9308 observations at 17.6 km<sup>2</sup> resolution in the 30° × 30° region, and MODIS acquired 47,695 observations at 10 km<sup>2</sup> resolution in the same region. The MISR

and MODIS footprints are approximately square areas oriented along the direction of flight.

The maps of the raw data in Figure 1 illustrate the complementary coverage of the two instruments. MISR observations are clearly grouped into three distinct stripes, which correspond to the three orbital swaths MISR had over the domain between January 1 and January 16, 2001. MODIS had much more complete coverage of the domain due to its wider field of view, but there are also significant gaps in the data. These are most likely due to unfavorable atmospheric conditions causing radiance data to be too unstable to be of use. Both maps in Figure 1 indicate that there is a general trend of high visibility in the southern areas and low visibility near the northern edge of the domain, where the nations Gabon, Congo, and the Democratic Republic of Congo are found. Hazy conditions over this region may result from windblown dust coming from the Sahara Desert or land-clearing for agriculture.

### 1.2 Importance of Data Fusion

Differences in MISR and MODIS technologies, fields of view, and retrieval algorithms lead to coverage differences that can both be complementary and reinforcing. For instance, the MODIS aerosol-retrieval algorithm does not operate over desert or other bright land surfaces, while MISR’s does. MODIS also has difficulty computing AOD in sun glint over dark water, while MISR is able to do so because at least some of its view angles will not be in glint at any given time. On the other hand, MODIS has greater coverage due to its wider field of view. These complementary characteristics produce a joint dataset that has more complete coverage than either of the contributing datasets, and a data-fusion methodology can take advantage of this to provide more complete and accurate inferences over the domain. Where the two instruments do observe simultaneously, their data can reinforce each other’s if their measurement-error characteristics are known and if differences in their spatial supports are taken into account. The complementary strengths of MISR and MODIS, coupled with substantial areas of simultaneous observations, make these instruments’ datasets ideal for data fusion. In fact, they are frequently used in conjunction with one another

(Kahn et al. 2009), but no formal statistical data fusion has been done.

To exploit this synergy, any data-fusion methodology must overcome two difficulties, namely, the massiveness of the remote sensing data and their different footprints (spatial supports). Remote sensing instruments typically collect tens of thousands of observations per day. Considering that complete coverage of the Earth may take several days or even weeks, these remote sensing datasets are massive relative to the computational limitations of existing spatial-smoothing methodologies, such as splines or kriging. The second difficulty is that the datasets to be fused have different spatial supports. Remote sensing instruments capture and record the relevant process as an image, where each pixel, also called a footprint, corresponds to some area in the domain. These footprints differ in shape, size, orientation, and alignment, making it a challenge to estimate the underlying process. If not dealt with properly, this change-of-support problem can make inferences susceptible to the so called “ecological fallacy”: erroneous conclusions can occur when inferences drawn from aggregated data are assumed to apply to individual units (e.g., Cressie 1996).

Gotway and Young (2002) gave a review of the change-of-support problem, namely inferring a spatial process at one resolution from data at another resolution. This formulation is closer to addressing our problem, but their discussion and review of existing methodologies focus on upscaling and downscaling from a *single* dataset. Other contributions in the statistics literature include those by Wikle and Berliner (2005; single dataset, Bayesian change of resolution), Fuentes and Raftery (2005; combine observations and numerical-model output, Bayesian melding), and Berrocal, Gelfand, and Holland (2010; areal-level and point-level data, downscaling through Bayesian regression). However, these approaches to data fusion are not implementable for massive datasets, as their computational complexities are quadratic or higher with respect to their data size.

Recent spatial inferential methodologies that are scalable include those by Nychka, Wikle, and Royle (2002; modeling nonstationary covariance functions with multiresolutional wavelet models), Banerjee et al. (2008; approximate optimal prediction with dimension reduction through a small set of space-filling locations), Stein and Jun (2008; modeling nonstationary covariance models using the discrete Fourier transform), and Cressie and Johannesson (2008; Fixed Rank Kriging (FRK) based on the Spatial Random Effects (SRE) model). Other approaches that are not strictly scalable, but can still handle large spatial datasets, are approximations obtained by covariance tapering (e.g., Furrer, Genton, and Nychka 2006) and from using Gaussian Markov random fields (e.g., Lindgren, Rue, and Lindström 2011).

In what follows, we leverage the computational efficiency provided by FRK to solve the data-fusion problem when the data sources are massive, as is often the case in remote sensing. In Section 2, we briefly review the basic spatial-statistical framework and the SRE model. In Section 3, we outline a data-fusion methodology that builds upon FRK to resolve the issues of massiveness and heterogeneous spatial support. We call this Spatial Statistical Data Fusion (SSDF). In Section 4, we dis-

cuss the results of SSDF and compare its performance to FRK and Bayesian melding. In Section 5, we discuss our findings and possible extensions of SSDF to related problems. Finally, parameter estimation is described in the Appendices.

## 2. THE SPATIAL STATISTICAL MODEL

Our data-fusion methodology exists within a geostatistical framework, which is a part of the broader area of spatial statistics. Here, we briefly review that framework, give some necessary notation, and present basic derivations for estimation in a spatial context. In particular, we discuss a parameterization of the spatial covariance matrix through the SRE model (Cressie and Johannesson 2008), which permits fast inversion of very large covariance matrices.

Let  $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$  be a hidden, real-valued spatial process on a discretized domain. The domain of interest is  $\cup\{A_i \subset \mathfrak{R}^d : i = 1, \dots, N_D\}$ , which is made up of  $N_D$  fine-scale, nonoverlapping, areal regions  $\{A_i\}$  with locations  $D \equiv \{\mathbf{p}_i \in A_i : i = 1, \dots, N_D\}$ . We call these fine-scale regions the Basic Areal Units (BAUs), and they represent the smallest resolution at which we will make predictions with our model. In the remote sensing context, the choice of BAUs is natural and often determined by the smallest resolution required by the scientific problems for which the predicted field will be used. Moreover, the SRE model that we shall use below has a *form* that is invariant to the choice of BAUs.

For the moment, consider just one instrument. Let  $\mathbf{Z}$  be a vector of observations at  $N$  footprints  $\{B_1, \dots, B_N\} \subset \mathfrak{R}^d$ , where a generic footprint  $B$  is made up of BAUs with locations indexed by  $D \cap B$ . We assume that each element of  $\mathbf{Z} \equiv (Z(B_1), \dots, Z(B_N))'$  is generated as an average of  $Y(\cdot)$  plus an independent error term,  $\epsilon(\cdot)$ ; that is, at both, where  $Y(\cdot)$  is observed and at potential footprints,

$$Z(B) = \frac{1}{|D \cap B|} \left\{ \sum_{\mathbf{s} \in D \cap B} Y(\mathbf{s}) \right\} + \epsilon(B); \quad B \subset \mathfrak{R}^d. \quad (1)$$

The measurement-error term  $\epsilon(B)$  may have nonzero mean that captures the instrument bias, and it has measurement-error variance  $\sigma_\epsilon^2 v(B) > 0$ , where  $v(\cdot)$  is assumed known and allows for the possibility of nonconstant variance over the domain  $D$ . For instance, if we know quantitatively how certain observations have higher measurement errors relative to surrounding regions due to unfavorable atmospheric conditions, we can construct the appropriate function  $v(\cdot)$  to reflect the varying measurement errors. Finally,  $\epsilon(B_i)$  and  $\epsilon(B_j)$  are assumed independent, for  $i \neq j$ .

The true process,  $Y(\cdot)$ , is assumed to have a linear mean structure:

$$Y(\mathbf{s}) = \mathbf{t}(\mathbf{s})' \boldsymbol{\alpha} + v(\mathbf{s}) + \xi(\mathbf{s}); \quad \mathbf{s} \in D, \quad (2)$$

where notice that it is defined on the BAUs, which are indexed by  $D$ . The first term on the right-hand side of Equation (2) accounts for an assumed linear model in the trend, where  $\mathbf{t}(\cdot) \equiv (t_1(\cdot), \dots, t_p(\cdot))'$  is a vector of  $p$  known covariates, such as geographical coordinates or other physical variables. The vector of linear coefficients,  $\boldsymbol{\alpha}$ , is unknown. The middle term,  $v(\cdot)$ ,

captures the spatial covariance and is assumed to have mean zero and finite, possibly heteroscedastic variance. The last term,  $\xi(\cdot)$ , describes the variability of the process at the scale of the BAU. We assume that  $\xi(\cdot)$  is an independent Gaussian process with mean zero and variance  $\sigma_\xi^2$ . The spatial covariance function for Equation (2) is,

$$\begin{aligned} \text{cov}(Y(\mathbf{u}), Y(\mathbf{v})) &= \text{cov}(v(\mathbf{u}) + \xi(\mathbf{u}), v(\mathbf{v}) + \xi(\mathbf{v})) \\ &= C(\mathbf{u}, \mathbf{v}) + \sigma_\xi^2 I(\mathbf{u} = \mathbf{v}); \quad \mathbf{u}, \mathbf{v} \in D. \end{aligned} \quad (3)$$

Combining Equations (1) and (2), we have the following general linear model:

$$\begin{aligned} \mathbf{Z} &= \mathbf{Y} + \boldsymbol{\epsilon}, \\ \mathbf{Y} &= \mathbf{T}\boldsymbol{\alpha} + \mathbf{v} + \boldsymbol{\xi}, \end{aligned}$$

where  $\boldsymbol{\epsilon}$ ,  $\mathbf{v}$ , and  $\boldsymbol{\xi}$  are random vectors of length  $N$  representing the corresponding processes evaluated at the observation footprints,  $B_1, \dots, B_N$ . Further,  $\mathbf{T}$  is an  $N \times p$  matrix of the covariates,  $(\mathbf{t}(B_1), \dots, \mathbf{t}(B_N))'$ , where  $\mathbf{t}(B) \equiv \sum_{\mathbf{s} \in D \cap B} \mathbf{t}(\mathbf{s}) / |D \cap B|$  is a  $p$ -dimensional vector; and  $\boldsymbol{\epsilon}$ ,  $\mathbf{v}$ , and  $\boldsymbol{\xi}$  are statistically independent. It is shown below that the covariance matrix  $\boldsymbol{\Sigma} \equiv \text{var}(\mathbf{Z})$  is given by,

$$\boldsymbol{\Sigma} = [\text{cov}(Z(B_i), Z(B_j))] = \boldsymbol{\Sigma}_Y + \sigma_\xi^2 \mathbf{E} + \sigma_\epsilon^2 \mathbf{V}. \quad (4)$$

In Equation (4), the  $N \times N$  covariance matrices on the right-hand side are defined by  $\boldsymbol{\Sigma}_Y \equiv [C(B_i, B_j)]$ ,  $\mathbf{E} \equiv \left[ \frac{|D \cap B_i \cap B_j|}{|D \cap B_i| |D \cap B_j|} \right]$ , and  $\mathbf{V} \equiv \text{diag}(v(B_1), \dots, v(B_N))$ , where the latter is a known diagonal matrix. The covariance function  $C(B_i, B_j)$  can be expanded in terms of the BAU covariances,

$$C(B_i, B_j) = \frac{1}{|D \cap B_i| |D \cap B_j|} \sum_{\mathbf{u} \in D \cap B_i} \sum_{\mathbf{v} \in D \cap B_j} C(\mathbf{u}, \mathbf{v}).$$

We are interested in inferring the ‘‘true’’ process,  $Y(\cdot)$ , at all locations  $\mathbf{s}_0 \in D$  (i.e., at the BAU level of resolution).

Kriging yields the best linear unbiased predictor,  $\hat{Y}(\mathbf{s}_0) = \mathbf{a}'\mathbf{Z}$ , but the methodology requires inversion of the  $N \times N$  covariance matrix,  $\boldsymbol{\Sigma}$ . This inversion can be a computational bottleneck for large-to-massive datasets. Cressie and Johannesson (2006, 2008) developed a flexible, nonstationary spatial statistical model that resolves the computational bottleneck (see also Shi and Cressie 2007). They express the spatial covariance term,  $v(\mathbf{s})$ , in Equation (2) as the product of an  $r$ -dimensional vector of known spatial basis functions,  $\mathbf{S}(\mathbf{s})$ , and an  $r$ -dimensional Gaussian random variable,  $\boldsymbol{\eta}$ , with  $\text{var}(\boldsymbol{\eta}) = \mathbf{K}$ . That is,  $v(\mathbf{s}) = \mathbf{S}(\mathbf{s})'\boldsymbol{\eta}$ . Consequently, the process model in Equation (2) can be written as the linear mixed model,

$$Y(\mathbf{s}) = \mathbf{t}(\mathbf{s})'\boldsymbol{\alpha} + \mathbf{S}(\mathbf{s})'\boldsymbol{\eta} + \xi(\mathbf{s}). \quad (5)$$

Cressie and Johannesson (2008) call the model for the stochastic part,

$$\mathbf{S}(\cdot)'\boldsymbol{\eta} + \xi(\cdot), \quad (6)$$

the SRE model; and they call the model given by Equation (5) the Spatial Mixed Effects (SME) model.

The model given by Equation (5), substituted into Equation (1), has remarkable change-of-support properties; observe

that the covariance between two footprints is,

$$\begin{aligned} \text{cov}(Z(B_i), Z(B_j)) &= \text{cov} \left( \frac{1}{|D \cap B_i|} \left\{ \sum_{\mathbf{u} \in D \cap B_i} Y(\mathbf{u}) \right\} + \epsilon(B_i), \right. \\ &\quad \left. \times \frac{1}{|D \cap B_j|} \left\{ \sum_{\mathbf{v} \in D \cap B_j} Y(\mathbf{v}) \right\} + \epsilon(B_j) \right). \end{aligned}$$

Because  $\xi(\cdot)$  and  $\epsilon(\cdot)$  are independent of each other and of  $\boldsymbol{\eta}$ , we obtain,

$$\begin{aligned} \text{cov}(Z(B_i), Z(B_j)) &= \frac{1}{|D \cap B_i|} \frac{1}{|D \cap B_j|} \sum_{\mathbf{u} \in D \cap B_i} \sum_{\mathbf{v} \in D \cap B_j} \text{cov}(Y(\mathbf{u}), Y(\mathbf{v})) \\ &= \frac{1}{|D \cap B_i|} \frac{1}{|D \cap B_j|} \sum_{\mathbf{u} \in D \cap B_i} \sum_{\mathbf{v} \in D \cap B_j} \{ \mathbf{S}(\mathbf{u})'\mathbf{K}\mathbf{S}(\mathbf{v}) \\ &\quad + \sigma_\xi^2 I(\mathbf{u} = \mathbf{v}) \} + \sigma_\epsilon^2 I(i = j)v(B_i) \\ &= \mathbf{S}(B_i)'\mathbf{K}\mathbf{S}(B_j) + \sigma_\xi^2 \frac{|D \cap B_i \cap B_j|}{|D \cap B_i| |D \cap B_j|} \\ &\quad + \sigma_\epsilon^2 I(i = j)v(B_i), \end{aligned} \quad (7)$$

where the last two terms in Equation (7) have been defined in Equation (4), and

$$\mathbf{S}(B_i) \equiv \frac{1}{|D \cap B_i|} \sum_{\mathbf{u} \in D \cap B_i} \mathbf{S}(\mathbf{u}).$$

In a similar manner, we can show that,

$$\text{cov}(Z(B_i), Y(\mathbf{s}_0)) = \mathbf{S}(B_i)'\mathbf{K}\mathbf{S}(\mathbf{s}_0) + \text{cov}(\xi(B_i), \xi(\mathbf{s}_0)), \quad (8)$$

where  $\text{cov}(\xi(B_i), \xi(\mathbf{s}_0)) = \sigma_\xi^2 \frac{I(\mathbf{s}_0 \in B_i)}{|D \cap B_i|}$ . Notice that Equation (7) allows us to express the covariance between spatial averages explicitly in terms of the covariance parameter  $\mathbf{K}$ , which is defined at the BAU level through Equation (6). When  $\mathbf{K}$  is unknown, its estimation proceeds straightforwardly from aggregated data (Cressie and Johannesson 2008).

The result in Equation (7) leads to the specialization of Equation (4) to:

$$\boldsymbol{\Sigma} = \mathbf{S}'\mathbf{K}\mathbf{S} + \sigma_\xi^2 \mathbf{E} + \sigma_\epsilon^2 \mathbf{V}, \quad (9)$$

where  $\mathbf{S} \equiv (\mathbf{S}(B_1), \dots, \mathbf{S}(B_N))'$ . Using the Sherman-Morrison-Woodbury formula (e.g., Henderson and Searle 1981), the matrix inverse is given by,

$$\boldsymbol{\Sigma}^{-1} = \mathbf{U}^{-1} - \mathbf{U}^{-1}\mathbf{S}'(\mathbf{K}^{-1} + \mathbf{S}\mathbf{U}^{-1}\mathbf{S}')^{-1}\mathbf{S}\mathbf{U}^{-1}, \quad (10)$$

where  $\mathbf{U} \equiv \sigma_\xi^2 \mathbf{E} + \sigma_\epsilon^2 \mathbf{V}$ . The inversion is exact for any covariance function  $C(\cdot, \cdot)$  obtained from the class of SRE models given by Equation (6). The procedure requires inversion of the  $N \times N$  matrix  $\mathbf{U}$ , which is typically very sparse, and inversion of  $\mathbf{K}$  and  $(\mathbf{K}^{-1} + \mathbf{S}'\mathbf{U}^{-1}\mathbf{S})$ , both of which are  $r \times r$  matrices ( $r \ll N$ ).

For diagonal  $\mathbf{U}$ , the computational burden of inverting  $\boldsymbol{\Sigma}$  is dominated by inversions of  $r \times r$  matrices ( $O(r^3)$ ) and dense matrix multiplications ( $O(Nr^2)$ ). Since  $N \gg r$ , the number of computations required to invert  $\boldsymbol{\Sigma}$  is  $O(Nr^2)$  (Cressie and Johannesson 2008), and hence for fixed  $r$  the number of computations required to invert  $\boldsymbol{\Sigma}$  grows only linearly with the data size  $N$ . In our case,  $\mathbf{U}$  is not diagonal but it is sparse. If there

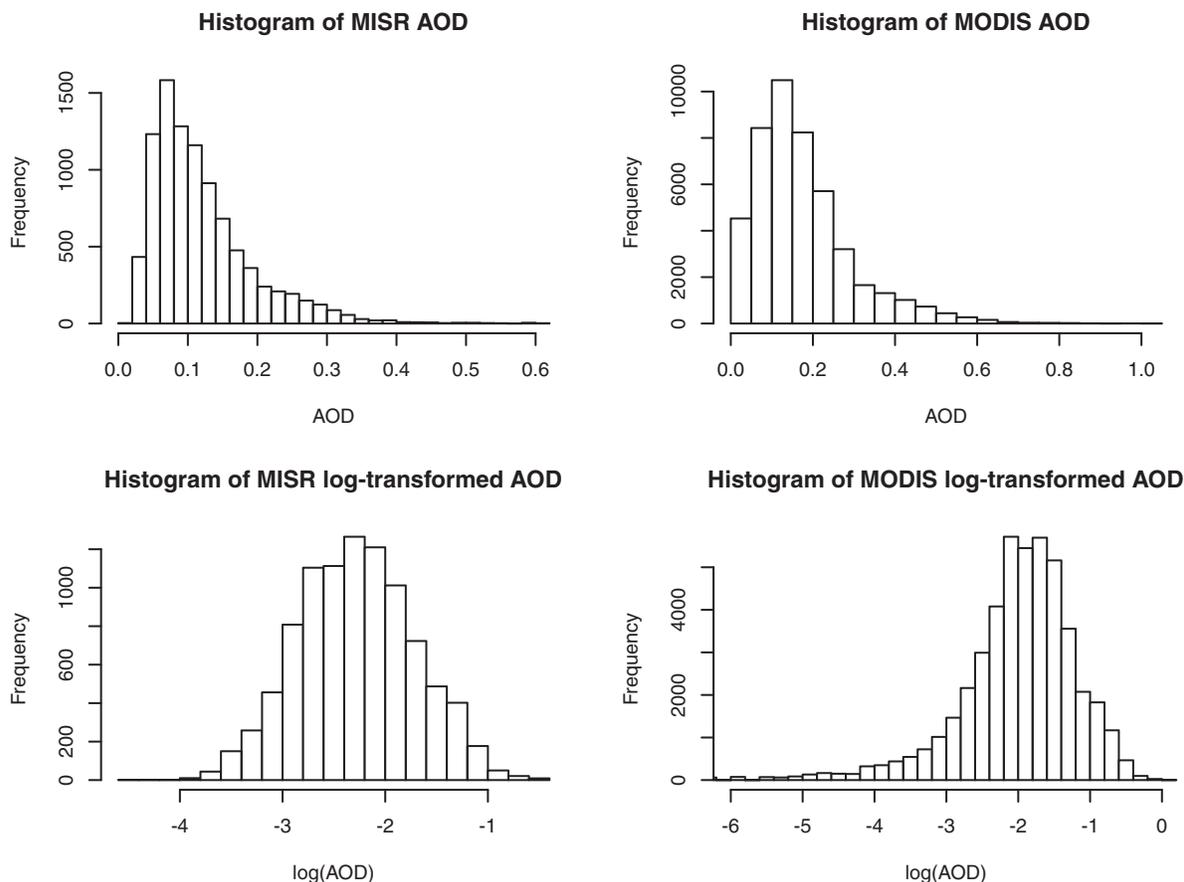


Figure 2. Top row: histograms of MISR and MODIS AOD. Bottom row: histograms of log-transformed MISR and MODIS AOD.

are very few overlaps between the footprints, the matrix  $\mathbf{U}$  is very sparse and can be inverted with sparse-matrix techniques. If  $\mathbf{U}$  is too dense for inversion with sparse-matrix techniques, the matrix  $\mathbf{E}$  may be approximated with an SRE model, and the computational efficiency of the Sherman-Morrison-Woodbury formula may again be leveraged. In summary, the combination of a fixed-rank, positive-definite matrix  $\mathbf{K}$ , and a set of spatial basis functions,  $\{S_j(\cdot) : j = 1, \dots, r\}$ , produces a flexible, nonstationary family of covariance functions that allows rapid computation of the covariance matrix of a single spatial dataset  $\mathbf{Z}$  and hence rapid computation of  $\{\hat{Y}(\mathbf{s}_0) : \mathbf{s}_0 \in D\}$ .

In the remainder of this section, we apply the parameterization described above to aerosol data in southern Africa. We first discretized the  $30^\circ \times 30^\circ$  domain into a fine-scale grid of regular hexagons using the Discrete Global Grid software (Carr et al. 1998). Specifically, we used resolution 17 of the ISEA Aperture 3 Hexagon (ISEA3H) global grid, with intercell distance of 0.675 km and cell area of 0.395 km<sup>2</sup>. These hexagons are defined to be the BAUs,  $\{A_i : i = 1, \dots, N_D\}$ , whose centers define the index set  $D$ . The covariates  $\mathbf{t}(\cdot)$  were constructed using the three-dimensional vector made up of unity, latitude, and longitude.

Histograms of MISR and MODIS AOD are displayed in the top row of panels in Figure 2. Both histograms exhibit long tails on the right, indicating that visibility in the domain is mostly clear and that hazy conditions are relatively rare. Since we assumed in the models given above that the data are normally distributed, we took the log of both MISR and MODIS AOD to

ensure that the normality assumption is (at least approximately) met. The histograms of the log-transformed AOD for MISR and MODIS are shown in the bottom row of panels in Figure 2. An important feature of the log-transform is that it ensures our predictions will be positive. AODs typically range between 0 and 1, and applying our data-fusion methodology to nontransformed data may lead to negative AOD, which would be inappropriate. In what follows, we apply our methodology on all the log-transformed AOD values (essentially assuming that the AOD data is lognormal), and then we transform the predictions back into AOD for interpretation.

We used  $\mathbf{t}(\cdot)$  to detrend the raw data as described in Appendix A. The covariance parameters were estimated from the combined dataset consisting of  $\mathbf{Z}_1$  and  $\mathbf{Z}_2$  (see the Appendices). To construct  $\mathbf{S}(\cdot)$ , we used local bisquare functions at  $b_0$  resolutions:

$$f_{a(b)}(\mathbf{u}) = \begin{cases} \left(1 - \frac{\|\mathbf{u} - \mathbf{m}_{a(b)}\|^2}{d_b^2}\right)^2 & \text{for } \|\mathbf{u} - \mathbf{m}_{a(b)}\| \leq d_b, \\ 0 & \text{otherwise,} \end{cases}$$

where  $\mathbf{u}, \mathbf{m}_{a(b)} \in \mathfrak{R}^d$ ;  $\mathbf{m}_{a(b)}$  is the  $a$ th center point at the  $b$ th resolution for  $b = 1, 2, \dots, b_0$ ; and  $\|\cdot\|$  denotes Euclidean distance. Note that  $\mathbf{m}_{a(b)}$  denotes a center location for a set of basis functions in the  $b$ th resolution, and  $d_b$  is defined as 1.5 times the shortest distance between two such center points (Cressie and Johannesson 2008).

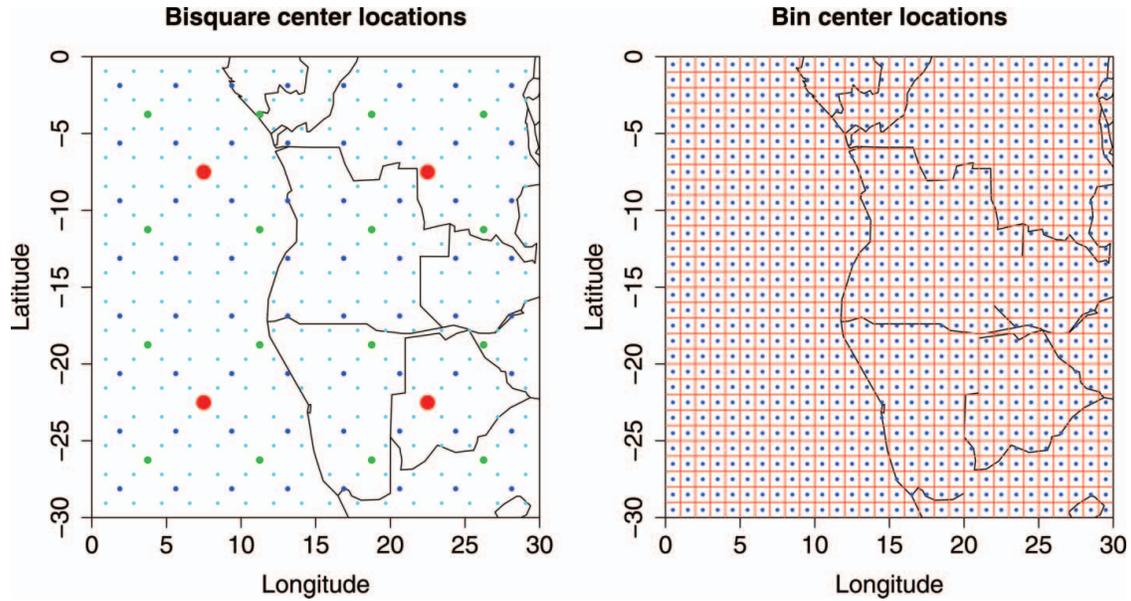


Figure 3. Left panel: bisquare basis function centers at four resolutions (Resolution-1 centers are in red, Resolution-2 centers are in green, Resolution-3 centers are in blue, and Resolution-4 centers are in teal). Right panel: Bin centers (blue dots) of the  $1^\circ \times 1^\circ$  bins. The online version of this figure is in color.

We employed a quadtree design with  $b_0 = 4$  resolutions for construction of  $S(\cdot)$ . The four spatial basis functions at resolution 1 are defined relative to the four coarsest centers, which are spread evenly over the domain. The next 16 basis functions are defined relative to each of 16 evenly spaced centers at Resolution 2. We repeated the procedure twice more, for a total of  $4 + 16 + 64 + 256 = 340$  basis functions. The four resolutions of bisquare-basis function centers are displayed in the left panel of Figure 3, and the  $900 1^\circ \times 1^\circ$  bins used for calculating the empirical covariance functions are displayed in the right panel of Figure 3. Having introduced and parameterized the geostatistical estimation framework in the single-dataset case, we describe the data-fusion methodology in the next section.

### 3. SPATIAL STATISTICAL DATA FUSION (SSDF)

In data fusion, we are interested in optimally estimating an underlying process,  $Y(\cdot)$ , from two (or more) realizations,  $\mathbf{Z}_1$  and  $\mathbf{Z}_2$ . In this section, we derive the best linear unbiased predictor for the process  $Y(\cdot)$ . We start with model definitions and properties (Section 3.1), and we derive the general optimal predictor without assuming any specific model for the covariance structure (Section 3.2). We then assume that the covariance structure is determined by an SRE model, and we show how the resulting optimal predictor is suitable for massive and spatially incompatible data (Section 3.3). Finally, in Section 3.4, we apply our data-fusion methodology to the AOD data from southern Africa, introduced in Section 2.

#### 3.1 Data Model and Properties

Assume an underlying true process  $Y(\cdot)$  defined at the BAU level and two datasets (e.g., from two instruments),  $\mathbf{Z}_1$  and  $\mathbf{Z}_2$ , observed at different spatial supports, both larger than the BAUs that define  $D$ . As in Equation (1), assume that the data vectors,  $\{\mathbf{Z}_k : k = 1, 2\}$ , are generated as spatial averages over a set of

footprints according to the following model:

$$Z_k(B_{km}) = \frac{1}{|D \cap B_{km}|} \left\{ \sum_{\mathbf{u} \in D \cap B_{km}} Y(\mathbf{u}) \right\} + \epsilon_k(B_{km});$$

$$B_{km} \subset \mathfrak{R}^d, \quad (11)$$

where  $B_{km}$  represents the  $m$ th footprint in dataset  $k$ , the common process  $Y(\cdot)$  is given by Equation (2), and  $\epsilon_k(B_{km})$  is a Gaussian error process associated with the measurement for footprint  $B_{km}$ ;  $m = 1, \dots, N_k$ ,  $k = 1, 2$ .

In vector notation,  $\mathbf{Z}_k$ , the data vector for dataset  $k$  may be written as,

$$\mathbf{Z}_k = \mathbf{Y}_k + \boldsymbol{\epsilon}_k, \quad (12)$$

$$\mathbf{Y}_k = \mathbf{T}_k \boldsymbol{\alpha} + \mathbf{v}_k + \boldsymbol{\xi}_k; \quad k = 1, 2,$$

where  $\boldsymbol{\epsilon}_k$ ,  $\mathbf{v}_k$ , and  $\boldsymbol{\xi}_k$  are, respectively, the measurement-error, the smooth-spatial variation, and the fine-scale variation processes evaluated at the footprints in dataset  $k$ . The covariance between the  $k$ th and  $l$ th data vectors,  $\boldsymbol{\Sigma}_{kl}$ , is given as

$$\boldsymbol{\Sigma}_{kl} \equiv \text{cov}(\mathbf{Z}_k, \mathbf{Z}_l) = \text{cov}(\mathbf{v}_k, \mathbf{v}_l) + \sigma_\xi^2 \mathbf{E}_{kl} + \sigma_\epsilon^2 I(k=l) \mathbf{V}_k;$$

$$k, l = 1, 2,$$

where  $\mathbf{E}_{kl} = \left[ \frac{|D \cap B_{ka} \cap B_{lb}|}{|D \cap B_{ka}| |D \cap B_{lb}|} \right]$  is a matrix where the  $(a, b)$  element is the number of BAUs in the intersection between the two footprints  $B_{ka}$  and  $B_{lb}$ , divided by the product of the number of BAUs in the two footprints, and  $\mathbf{V}_k$  is a known diagonal matrix associated with the measurement-error process for dataset  $k$ . Similarly, the covariance between dataset  $\mathbf{Z}_k$  and the process  $Y(\cdot)$  at location  $\mathbf{s}$  is,

$$\text{cov}(\mathbf{Z}_k, Y(\mathbf{s})) \equiv \text{cov}(\mathbf{v}_k, \nu(\mathbf{s})) + \mathbf{b}_k(\mathbf{s}),$$

where

$$\mathbf{b}_k(\mathbf{s}) \equiv \text{cov}(\boldsymbol{\xi}_k, \xi(\mathbf{s})) = \sigma_\xi^2 \left( \frac{I(\mathbf{s} \in B_{k1})}{|D \cap B_{k1}|}, \dots, \frac{I(\mathbf{s} \in B_{kN_k})}{|D \cap B_{kN_k}|} \right)';$$

$$k = 1, 2. \quad (13)$$

As we mentioned for the model in Equation (1), we allow for the possibility of systematic instrument bias. Suppose that the trend of  $Y(\cdot)$  is linear in spatial covariates, that is,  $E(Y(\mathbf{s})) = \mathbf{t}(\mathbf{s})'\boldsymbol{\alpha}$ , for  $\mathbf{s} \in D$ . Then  $Y(B)$  has expectation given by  $\mathbf{t}(B)'\boldsymbol{\alpha}$ , where  $B$  is a generic footprint and the definition of  $\mathbf{t}(B)$  is given below Equation (3). Now suppose that the bias is captured multiplicatively, as follows:

$$E(\epsilon_k(B_{km})) = c_k E(Y(B_{km})) = c_k \mathbf{t}(B_{km})'\boldsymbol{\alpha}.$$

Consequently,

$$E(Z(B_{km})) = (1 + c_k) \mathbf{t}(B_{km})'\boldsymbol{\alpha}; \quad m = 1, \dots, N_k, k = 1, 2. \tag{14}$$

That is, the data are biased, and we call the parameters  $\{c_k : k = 1, 2\}$  the *bias coefficients*. These multiplicative bias coefficients will be obtained off-line, either from instrument specifications or from comparison with independent, unbiased data sources. Finally, the measurement-error variances are given by

$$\text{var}(\epsilon_k(B_{km})) = \sigma_{\epsilon,k}^2 v_k(B_{km}); \quad m = 1, \dots, N_k,$$

where for  $k = 1, 2$ ,  $v_k(B_{km})$  is known.

### 3.2 SSDF Predictors

The data-fusion predictor proposed below is a linear combination of  $\mathbf{Z}_1$  and  $\mathbf{Z}_2$ . We solve for the optimal set of data-fusion coefficients,  $\mathbf{a}_{1s}$  and  $\mathbf{a}_{2s}$ , that minimize the mean squared prediction error (MSPE) of the estimator at location  $\mathbf{s}$ . Without assuming a specific model for the covariance structure  $C(\cdot, \cdot)$ , let the estimator be written as

$$\hat{Y}(\mathbf{s}) \equiv \mathbf{a}'_{1s} \mathbf{Z}_1 + \mathbf{a}'_{2s} \mathbf{Z}_2. \tag{15}$$

Then, subject to an unbiasedness constraint, we minimize

$$\begin{aligned} E(Y(\mathbf{s}) - \hat{Y}(\mathbf{s}))^2 &= \text{var}(\mathbf{a}'_{1s} \mathbf{Z}_1 + \mathbf{a}'_{2s} \mathbf{Z}_2 - Y(\mathbf{s})), \\ &= \mathbf{a}'_{1s} \text{var}(\mathbf{Z}_1) \mathbf{a}_{1s} + \mathbf{a}'_{2s} \text{var}(\mathbf{Z}_2) \mathbf{a}_{2s} + \text{var}(Y(\mathbf{s})) \\ &\quad + 2\mathbf{a}'_{1s} \text{cov}(\mathbf{Z}_1, \mathbf{Z}_2) \mathbf{a}_{2s} - 2\mathbf{a}'_{1s} \text{cov}(\mathbf{Z}_1, Y(\mathbf{s})) \\ &\quad - 2\mathbf{a}'_{2s} \text{cov}(\mathbf{Z}_2, Y(\mathbf{s})). \end{aligned} \tag{16}$$

Formally, the unbiasedness constraint is,

$$E(Y(\mathbf{s})) = \mathbf{t}(\mathbf{s})'\boldsymbol{\alpha} = E(\mathbf{a}'_{1s} \mathbf{Z}_1 + \mathbf{a}'_{2s} \mathbf{Z}_2) = E(\hat{Y}(\mathbf{s})),$$

uniformly in  $\boldsymbol{\alpha}$ . Under the multiplicative-bias assumption given by Equation (14), this becomes,

$$\mathbf{t}(\mathbf{s})' = (1 + c_1) \mathbf{a}'_{1s} \mathbf{T}_1 + (1 + c_2) \mathbf{a}'_{2s} \mathbf{T}_2,$$

where  $\mathbf{T}_k = (\mathbf{t}(B_{k1}), \dots, \mathbf{t}(B_{kN_k}))'$ ;  $k = 1, 2$ .

The solution to the minimization problem in Equation (16) can be simplified by considering an alternative formulation of our problem. Given the data vectors  $\mathbf{Z}_1$  and  $\mathbf{Z}_2$ , we can stack them to form the following model,

$$\begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{T}_1 \\ \mathbf{T}_2 \end{pmatrix} \boldsymbol{\alpha} + \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \end{pmatrix}, \tag{17}$$

or equivalently,

$$\mathbf{Z}_F = \mathbf{T}_F \boldsymbol{\alpha} + \mathbf{v}_F + \boldsymbol{\xi}_F + \boldsymbol{\epsilon}_F, \tag{18}$$

where  $\mathbf{Z}_F$ ,  $\mathbf{T}_F$ ,  $\mathbf{v}_F$ ,  $\boldsymbol{\xi}_F$ , and  $\boldsymbol{\epsilon}_F$  each denotes the stacked version of the corresponding vector. From Equation (17), we can see

that  $\text{var}(\mathbf{Z}_F)$  and  $\text{cov}(\mathbf{Z}_F, Y(\mathbf{s}))$  are

$$\text{var}(\mathbf{Z}_F) \equiv \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \text{cov}(\mathbf{Z}_F, Y(\mathbf{s})) \equiv \begin{pmatrix} \text{cov}(\mathbf{Z}_1, Y(\mathbf{s})) \\ \text{cov}(\mathbf{Z}_2, Y(\mathbf{s})) \end{pmatrix}. \tag{19}$$

Under this formulation, the linear interpolator in Equation (15) can be written as

$$\hat{Y}(\mathbf{s}) = \mathbf{a}'_F \mathbf{Z}_F, \tag{20}$$

where  $\mathbf{a}'_F \equiv (\mathbf{a}'_{1s}, \mathbf{a}'_{2s})$  is an  $(N_1 + N_2)$ -dimensional vector of data-fusion coefficients. Then, subject to an unbiasedness constraint, we minimize,

$$\begin{aligned} E(Y(\mathbf{s}) - \hat{Y}(\mathbf{s}))^2 &= \text{var}(Y(\mathbf{s}) - \mathbf{a}'_F \mathbf{Z}_F), \\ &= \text{var}(Y(\mathbf{s})) - 2\mathbf{a}'_F \text{cov}(\mathbf{Z}_F, Y(\mathbf{s})) \\ &\quad + \mathbf{a}'_F \text{var}(\mathbf{Z}_F) \mathbf{a}_F, \end{aligned} \tag{21}$$

with respect to  $\mathbf{a}_F$ , where the unbiasedness constraint is  $p$ -dimensional:

$$\mathbf{0} = \mathbf{a}'_F \mathbf{C} \mathbf{T}_F - \mathbf{t}(\mathbf{s})', \tag{22}$$

and  $\mathbf{C}$  is the following  $(N_1 + N_2) \times (N_1 + N_2)$  diagonal matrix,

$$\mathbf{C} \equiv \begin{pmatrix} (1 + c_1) \mathbf{I}_{N_1} & \mathbf{0} \\ \mathbf{0} & (1 + c_2) \mathbf{I}_{N_2} \end{pmatrix}.$$

To simplify notation, let  $\boldsymbol{\Sigma}_F \equiv \text{var}(\mathbf{Z}_F)$  and  $\mathbf{c}_F \equiv \text{cov}(\mathbf{Z}_F, Y(\mathbf{s}))$ . Then, using Lagrange multipliers to minimize Equation (21) subject to Equation (22), the objective function to minimize with respect to the coefficients  $\mathbf{a}_F$  and the  $p$ -dimensional vector of Lagrange multipliers  $\mathbf{m}$  is

$$\begin{aligned} M &\equiv C(\mathbf{s}, \mathbf{s}) + \sigma_{\xi}^2 - 2\mathbf{a}'_F \mathbf{c}_F + \mathbf{a}'_F \text{var}(\mathbf{Z}_F) \mathbf{a}_F \\ &\quad + (\mathbf{a}'_F \mathbf{C} \mathbf{T}_F - \mathbf{t}(\mathbf{s})') \mathbf{m}. \end{aligned}$$

Upon differentiating  $M$  with respect to  $\mathbf{a}_F$  and  $\mathbf{m}$ , and setting the results equal to zero, we have

$$(2\mathbf{c}'_F \quad \mathbf{t}(\mathbf{s})') = (\mathbf{a}'_F \quad \mathbf{m}') \begin{pmatrix} 2\boldsymbol{\Sigma}_F & \mathbf{C} \mathbf{T} \\ \mathbf{T}' \mathbf{C} & \mathbf{0} \end{pmatrix}. \tag{23}$$

We can solve for  $\mathbf{a}'_F$  and  $\mathbf{m}'$  by inverting the matrix on the right-hand side of Equation (23) using block inversion. The solutions are

$$\mathbf{a}'_F = (\mathbf{c}'_F + (\mathbf{t}(\mathbf{s})' - \mathbf{c}'_F \boldsymbol{\Sigma}_F^{-1} \mathbf{C} \mathbf{T}_F) (\mathbf{T}'_F \mathbf{C} \boldsymbol{\Sigma}_F^{-1} \mathbf{C} \mathbf{T}_F)^{-1} \mathbf{T}'_F \mathbf{C}) \boldsymbol{\Sigma}_F^{-1}, \tag{24}$$

and

$$\mathbf{m}' = (\mathbf{t}(\mathbf{s})' - \mathbf{c}'_F \boldsymbol{\Sigma}_F^{-1} \mathbf{C} \mathbf{T}_F) (\mathbf{T}'_F \mathbf{C} \boldsymbol{\Sigma}_F^{-1} \mathbf{C} \mathbf{T}_F)^{-1}. \tag{25}$$

Having derived the data-fusion coefficients  $\mathbf{a}_F$ , we can produce the SSDF prediction and its prediction standard error at  $\mathbf{s} \in D$ , as follows,

$$\begin{aligned} Y(\mathbf{s})^{\text{SSDF}} &\equiv \mathbf{a}'_F \mathbf{Z}_F \\ \sigma(\mathbf{s})^{\text{SSDF}} &\equiv (C(\mathbf{s}, \mathbf{s}) + \sigma_{\xi}^2 - 2\mathbf{a}'_F \mathbf{c}_F + \mathbf{a}'_F \boldsymbol{\Sigma}_F \mathbf{a}_F)^{\frac{1}{2}}, \end{aligned} \tag{26}$$

where  $\mathbf{a}_F$  is given by Equation (24). Note that while Equation (26) produces predictions at the BAU level, it can be easily modified to produce predictions over larger areal regions (e.g., Cressie 1993, p. 124).

There are two practical problems with this general predictor. The first is that we need to invert the  $(N_1 + N_2) \times (N_1 + N_2)$  matrix  $\Sigma_F$  in Equation (24), which can cause a computational bottleneck for large  $(N_1 + N_2)$ . The second is that for most covariance models  $C(\cdot, \cdot)$ , estimating the model parameters from data with different spatial supports is a nontrivial task. In the next section, we use the SRE model to address these problems. Estimation of the SSDF parameters is discussed in Appendix A, and empirical estimates of the covariance and cross-covariance matrices are discussed in Appendix B.

### 3.3 Prediction Equations Under the SRE Model

The SRE model described in Section 2 allows us to express covariances between spatial averages explicitly in terms of its covariance parameter  $\mathbf{K}$ . By modeling the covariance function  $C(\cdot, \cdot)$  with that implied by the SRE model given by Equation (6), we not only solve the change-of-support problem, but also make the procedure computationally scalable. Under this model, the data can be written as

$$\begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{T}_1 \\ \mathbf{T}_2 \end{pmatrix} \boldsymbol{\alpha} + \begin{pmatrix} \mathbf{S}'_1 \\ \mathbf{S}'_2 \end{pmatrix} \boldsymbol{\eta} + \begin{pmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \end{pmatrix}, \quad (27)$$

where  $\mathbf{S}_k \equiv (\mathbf{S}(B_{k1}), \dots, \mathbf{S}(B_{kN_k}))'$ ,  $\boldsymbol{\eta}$  is an  $r$ -dimensional Gaussian random variable with  $\text{var}(\boldsymbol{\eta}) = \mathbf{K}$ , and  $r \ll \min(N_1, N_2)$  is fixed. From Equation (27), we can see that under the assumption of a multiplicative bias in the observed dataset,  $\text{E}(\mathbf{Z}_F) = \mathbf{C}\mathbf{T}_F\boldsymbol{\alpha}$  and  $\text{var}(\mathbf{Z}_F)$  can be written as

$$\text{var}(\mathbf{Z}_F) \equiv \Sigma_F = \mathbf{S}'_F \mathbf{K} \mathbf{S}_F + \sigma_{\xi}^2 \mathbf{E}_F + \mathbf{V}_F, \quad (28)$$

where  $\mathbf{S}'_F \equiv (\mathbf{S}_1, \mathbf{S}_2)'$ ,

$$\mathbf{E}_F \equiv \begin{pmatrix} \mathbf{E}_{11} & \mathbf{E}_{12} \\ \mathbf{E}_{21} & \mathbf{E}_{22} \end{pmatrix}, \text{ and } \mathbf{V}_F \equiv \begin{pmatrix} \sigma_{\epsilon,1}^2 \mathbf{V}_1 & 0 \\ 0 & \sigma_{\epsilon,2}^2 \mathbf{V}_2 \end{pmatrix}.$$

Notice that  $\mathbf{K}$ ,  $\sigma_{\xi}^2$ , and  $\{\sigma_{\epsilon,k}^2 : k = 1, 2\}$  are the spatial-covariance parameters to be estimated. Similarly, we can see that for  $\mathbf{s}, \mathbf{u} \in D$ ,

$$\text{var}(Y(\mathbf{s})) \equiv \mathbf{S}(\mathbf{s})' \mathbf{K} \mathbf{S}(\mathbf{s}) + \sigma_{\xi}^2, \quad (29)$$

$$\text{cov}(Y(\mathbf{s}), Y(\mathbf{u})) \equiv \mathbf{S}(\mathbf{s})' \mathbf{K} \mathbf{S}(\mathbf{u}); \quad \mathbf{s} \neq \mathbf{u}, \quad (30)$$

and

$$\mathbf{c}_F \equiv \text{cov}(\mathbf{Z}_F, Y(\mathbf{s})) = \mathbf{S}'_F \mathbf{K} \mathbf{S}(\mathbf{s}) + \mathbf{b}_F(\mathbf{s}), \quad (31)$$

where  $\mathbf{b}_F(\mathbf{s}) \equiv (\mathbf{b}_1(\mathbf{s})', \mathbf{b}_2(\mathbf{s})')'$ , and  $\{\mathbf{b}_k(\cdot) : k = 1, 2\}$  were defined in Equation (13).

We can now rewrite the SSDF predictor and prediction standard error given by Equation (26) in terms of the covariances implied by the SRE model given by Equation (6). For  $\mathbf{s} \in D$ ,

$$Y(\mathbf{s})^{\text{SSDF}} = \mathbf{a}'_F \mathbf{Z}_F, \quad (32)$$

$$\begin{aligned} \sigma(\mathbf{s})^{\text{SSDF}} &= (\mathbf{S}(\mathbf{s})' \mathbf{K} \mathbf{S}(\mathbf{s}) + \sigma_{\xi}^2 - 2\mathbf{a}'_F (\mathbf{S}'_F \mathbf{K} \mathbf{S}(\mathbf{s}) + \mathbf{b}_F(\mathbf{s})) \\ &\quad + \mathbf{a}'_F (\mathbf{S}'_F \mathbf{K} \mathbf{S}_F + \sigma_{\xi}^2 \mathbf{E}_F + \mathbf{V}_F) \mathbf{a}_F)^{\frac{1}{2}}. \end{aligned} \quad (33)$$

The expression for the data-fusion coefficients  $\mathbf{a}_F$  will be the same as Equation (24), with the exception that we can now compute  $\Sigma_F^{-1}$  using the Sherman-Morrison-Woodbury formula given by Equation (10), as described in Section 2. Specifically,

$$\Sigma_F^{-1} = \mathbf{U}_F^{-1} - \mathbf{U}_F^{-1} \mathbf{S}'_F (\mathbf{K}^{-1} + \mathbf{S}_F \mathbf{U}_F^{-1} \mathbf{S}'_F)^{-1} \mathbf{S}_F \mathbf{U}_F^{-1},$$

where  $\mathbf{U}_F \equiv \sigma_{\xi}^2 \mathbf{E}_F + \mathbf{V}_F$ . As for the single-instrument case, the inversion of the covariance matrix  $\Sigma_F$  only requires inversion of  $\mathbf{U}_F$ ,  $\mathbf{K}$ , and  $(\mathbf{K}^{-1} + \mathbf{S}_F \mathbf{U}_F^{-1} \mathbf{S}'_F)$ ; the latter two are fixed-rank  $r \times r$  matrices.

### 3.4 Application to Aerosol Data in Southern Africa

To apply SSDF, we need to estimate the bias coefficients,  $c_1$  and  $c_2$ , and the SSDF parameters  $\mathbf{K}$ ,  $\sigma_{\xi}^2$ ,  $\sigma_{\epsilon,1}^2$ , and  $\sigma_{\epsilon,2}^2$ . Paradise et al. (2007) estimated the multiplicative bias coefficients for MISR and MODIS AOD as  $c_1 = 0.08$  and  $c_2 = 0.22$ , respectively. Since these are multiplicative biases, they convert to simple additive biases of  $\log(1 + c_1)$  and  $\log(1 + c_2)$  for the log-transformed MISR and MODIS AOD. We assume that the log-transformed AOD data have these additive biases and no multiplicative bias. We then removed the additive biases by subtracting them from the log-transformed data before applying SSDF. Using the individual semivariograms at lags 10, 20, 30, and 40 km, and the cross-variogram at lag 10 km (see Appendix A), we obtained estimates for the variability parameters:  $\hat{\sigma}_{\xi}^2 = 8.25$ ,  $\hat{\sigma}_{\epsilon,1}^2 = 0.26$  for MISR, and  $\hat{\sigma}_{\epsilon,2}^2 = 0.43$  for MODIS (we assumed that all observations within a given dataset have the same measurement-error variability, so we let  $v_k(\cdot) = 1$ ).

We estimated  $\mathbf{K}$  using a binned method-of-moments estimator (see the Appendices). Following Cressie and Johannesson (2008), we computed a diagnostic summary of the SSDF parameter estimates by comparing theoretical semivariograms to empirical semivariograms as functions of spatial lag. Figure 4 shows comparisons of theoretical and empirical semivariograms at four locations in  $D$ . Specifically, we computed the empirical semivariogram, as a function of spatial lag, using all MODIS data within 300 km of a given location. We chose to use MODIS data because MODIS has much better coverage of the domain. The theoretical semivariograms at corresponding spatial lags were computed by averaging the values implied by Equation (8). The good fit of the theoretical semivariograms to the empirical semivariograms in Figure 4 indicates that the fitted SSDF model fits the data well.

## 4. RESULTS AND COMPARISONS

In this section, we apply SSDF to optimally estimate AOD in a  $30^\circ \times 30^\circ$  region of southern Africa. We compare the SSDF predictions to FRK predictions in Section 4.1, and we compare SSDF to Bayesian melding (Fuentes and Raftery 2005) in terms of computing time and predictive performance in Section 4.2.

### 4.1 SSDF and FRK Predictions

Having estimated the SSDF parameters in Section 3.3, we solved for the optimal SSDF predictors and their prediction standard errors at each prediction location, as given by Equation (26). We overlaid a  $300 \times 300$  regular grid over the domain, and we used the resulting 90,000 grid cells as our prediction locations. Each grid cell is a  $0.1^\circ \times 0.1^\circ$  region, and this is the spatial support of the AOD to be predicted.

For comparison, we did optimal prediction based on the MISR and MODIS data individually. Note that the SSDF formulas can be adapted to optimally predict the process  $Y(\cdot)$  from a single dataset,  $\mathbf{Z}$ , without any major modification. We can simply replace the stacked terms with the corresponding single-dataset

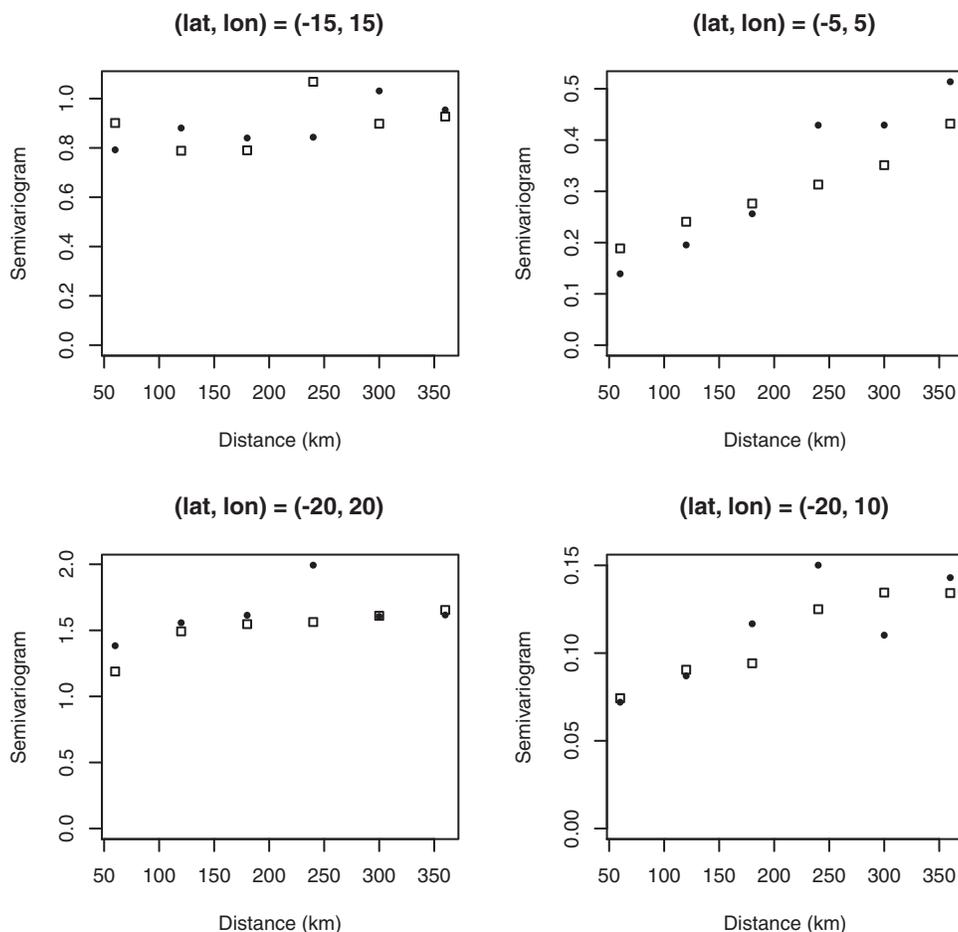


Figure 4. Comparison of empirical semivariograms versus estimated semivariograms as functions of distance at four locations in  $D$ . Solid circles indicate robust semivariogram estimates, and white squares denote semivariogram estimates derived from the SRE model.

elements (e.g.,  $\mathbf{Z}_F = \mathbf{Z}$ ,  $\mathbf{S}_F = \mathbf{S}$ , etc.), and the prediction and prediction standard error in Equations (32) and (33), respectively, remain the same. In its single-dataset form, SSDF is essentially FRK (Cressie and Johannesson 2008).

Since we are interested in the comparison of prediction properties, we used the same estimates of  $\mathbf{K}$ ,  $\sigma_\varepsilon^2$ , and  $\{\sigma_{\varepsilon,k}^2 : k = 1, 2\}$  in the three predictors, which are derived from the combined dataset of  $\mathbf{Z}_1$  and  $\mathbf{Z}_2$  (see Appendices). Having computed the predictions using log-transformed AOD data, we transformed them back onto the AOD scale (see, e.g., Cressie 1993, p. 135). The top row of panels in Figure 5 shows the individual-dataset FRK predictors (left and center), and the SSDF predictor (right). The corresponding root mean squared prediction errors (RMSPE; i.e., prediction standard errors) are shown in the bottom row of panels.

The patterns of prediction standard errors from individual-dataset FRKs look like the respective patterns of the data density. This makes sense intuitively, because we expect to have small prediction standard errors for locations with many nearby observations and large prediction standard errors for locations with very few nearby observations. The FRK map based on MISR has large parts of the region with large prediction standard errors due to its incomplete coverage. Since we assumed that AOD is log-normal, the prediction standard errors are proportional to the underlying trend, which translates into smaller prediction stan-

dard errors in the southern areas and larger prediction standard errors in the northern areas.

The SSDF prediction map (top, right panel) looks very similar to that of the MODIS map, because MODIS has much denser and more complete coverage. However, there are a few noticeable differences between the two in areas where SSDF is able to capitalize on complementary information from MISR to improve the prediction. One example of this is the region of low visibility off the coast of Angola between  $-12^\circ$  and  $-18^\circ$  latitude, and  $3^\circ$  and  $10^\circ$  longitude. These hazy conditions are due to aerosols blown from inland areas; however, MODIS has sparse coverage of this region due to unfavorable atmospheric conditions. MODIS FRK reports relatively high prediction standard errors within this region, compared to nearby observed locations. MISR has some coverage in this area and SSDF takes advantage of this complementary coverage to achieve lower prediction standard errors there.

Another example of improvement through SSDF is in southern Namibia, between latitudes  $-30^\circ$  and  $-20^\circ$ , and longitudes  $10^\circ$  and  $20^\circ$ . MODIS has missing data in this region, and the observational pattern in the right panel of Figure 1 seems to suggest that MODIS has difficulty making retrievals of AOD in the transition between land and ocean. MISR, on the other hand, has good coverage of coastal areas. Over the coast of Namibia, the retrieved MODIS AODs around the missing-data

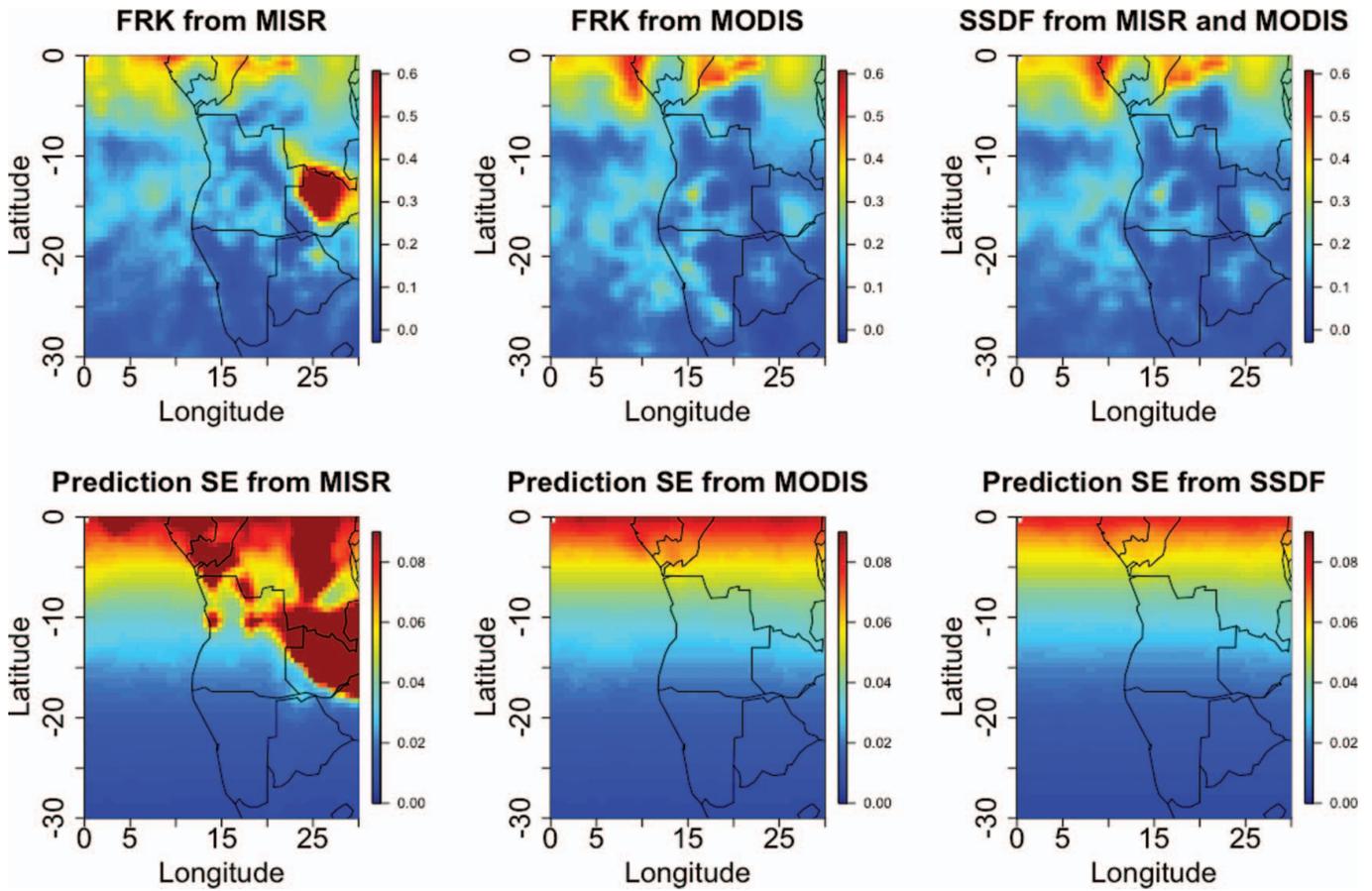


Figure 5. Predictions of AOD (top row of panels) and prediction standard errors of AOD (bottom row of panels) produced by applying FRK to MISR data (left panels) and to MODIS data (center panels), and by applying SSDF to MISR and MODIS data together (right panels).

region suggest medium values of AOD. However, the MISR data indicate that the Namibia coast has mostly good visibility with low aerosol content in the atmosphere. Looking at the MODIS FRK map alone, one might conclude that there are some medium aerosol plumes in the atmosphere above Namibia, but these plumes are likely spurious due to the sparse data coverage and instrument instability in the transition between land and ocean. SSDF is able to use the good coverage of MISR to show that the sky around Namibia has mostly low aerosol content. In summary, we would have missed certain aspects of AOD distribution over the domain if it were not for data fusion; aerosol climatologies constructed from SSDF benefit from the two instruments in a way that capitalizes on their strengths, and they are more accurate representations of the underlying AOD process.

One quantitative measure of improvement afforded by SSDF prediction is the ratio of SSDF prediction standard errors to individual FRK prediction standard errors. Figure 6 shows two such histograms: One is of  $\{\sigma^{\text{SSDF}}(s_j)/\sigma^{\text{FRK}}_{\text{MISR}}(s_j)\}$ , and the other is of  $\{\sigma^{\text{SSDF}}(s_j)/\sigma^{\text{FRK}}_{\text{MODIS}}(s_j)\}$ , where the index in the sequences runs from  $j = 1, \dots, N_p$ , the number of prediction locations. Both histograms have an upper limit of 1, indicating that SSDF prediction standard errors, in all instances, are smaller than prediction standard errors produced from both single-dataset optimal predictions. The improvement due to our optimal data-fusion methodology is often remarkable, particularly in comparison to MISR FRK (due to MISR’s lower sampling rate).

It is not surprising that SSDF produces more precise estimates than are produced by applying optimal spatial statistical prediction (i.e., FRK) to the input datasets separately. In effect, SSDF combines the two data sources into a single meta-dataset and uses more information. To see this, the SSDF and the single-dataset predictors can be written as an optimization of the estimates:

$$\begin{aligned}
 Y_{\text{SSDF}}(\mathbf{s}) &= \mathbf{a}'_1 \mathbf{Z}_1 + \mathbf{a}'_2 \mathbf{Z}_2, & \text{subject to } (1 + c_1) \mathbf{a}'_1 \mathbf{T}_1 \\
 & & + (1 + c_2) \mathbf{a}'_2 \mathbf{T}_2 = \mathbf{t}(\mathbf{s})', \\
 Y_{\text{MISR}}(\mathbf{s}) &= \mathbf{a}'_1 \mathbf{Z}_1 + \mathbf{0}' \mathbf{Z}_2, & \text{subject to } (1 + c_1) \mathbf{a}'_1 \mathbf{T}_1 = \mathbf{t}(\mathbf{s})', \\
 Y_{\text{MODIS}}(\mathbf{s}) &= \mathbf{0}' \mathbf{Z}_1 + \mathbf{a}'_2 \mathbf{Z}_2, & \text{subject to } (1 + c_2) \mathbf{a}'_2 \mathbf{T}_2 = \mathbf{t}(\mathbf{s})'.
 \end{aligned}$$

The MISR FRK predictor has optimal MSPE over the unbiased space spanned by  $(\mathbf{a}_1, \mathbf{0})$ , while the MODIS FRK predictor has optimal MSPE over the unbiased space spanned by  $(\mathbf{0}, \mathbf{a}_2)$ . The SSDF predictor, by definition, has minimum MSPE over the larger unbiased space spanned by  $(\mathbf{a}_1, \mathbf{a}_2)$ . Thus, SSDF offers superior performance with respect to MSPE when compared to nonfused individual unbiased spatial prediction.

#### 4.2 SSDF Versus Bayesian Melding

The analysis in the previous section shows that SSDF has lower prediction standard errors than the single-dataset FRKs. We now consider how the SSDF methodology compares to alternative data-fusion methodologies. To gain some insight into the relative performance and efficiency, we compared SSDF to

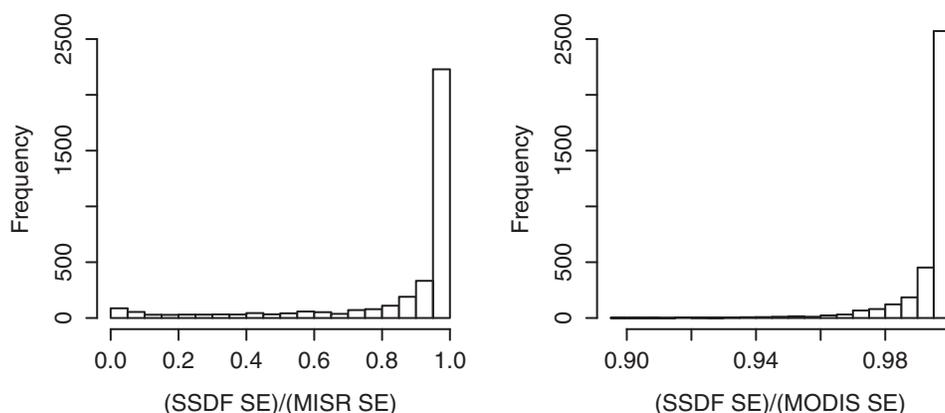


Figure 6. Histograms of prediction–standard error ratios. Left panel: SSDF compared to MISR. Right panel: SSDF compared to MODIS.

Bayesian melding (Fuentes and Raftery 2005). While Bayesian melding is presented in the context of combining pollution measurements from observation stations with outputs from numerical pollution models, the methodology can be adapted to make optimal prediction of a single process using outputs from two instruments.

Bayesian melding models the process locally, as a stationary isotropic random field with three parameters (smoothness, sill, and range) that describe the local spatial structure. These parameters are assumed to have known prior distributions, and they are allowed to vary across the domain. Spatial prediction at any given location is obtained by sampling from the process' posterior distribution. However, as will become apparent below, the methodology is not designed for very large datasets such as the one in this study, so we applied it to a suitably small random subset of the AOD data in southern Africa.

For our comparison exercise, we created random subsets of the MISR and MODIS AOD data, each of which is one-fortieth of the original data size. The subsets of MISR and MODIS AOD data have sizes of 1192 and 232, respectively, and they are shown in Figure 7. To divide the datasets into training and

test data, we designated a region of Namibia between  $-28^\circ$  and  $-20^\circ$  latitude and between  $14^\circ$  and  $19^\circ$  longitude as a *reserved region* (see Figure 7). We chose this region because it has good coverage from both MISR and MODIS. The data outside of this reserved region are used as training data for Bayesian melding, and the data inside the region (35 MISR observations and 43 MODIS observations) are withheld and used for comparison between SSDF and Bayesian-melding predictions. Since the SSDF methodology requires large datasets for stable parameter estimates, the SSDF training dataset consisted of the full AOD datasets *minus* any data from the reserved region. Hence, the two training datasets are not quite the same due to the unique requirements of the two methodologies, but a comparison of their predictive performance over the reserved region gives a direct comparison of their spatial-prediction efficiencies.

We applied both SSDF and Bayesian melding to log-transformed AOD data, and we compared the predictive performance on the log-transformed AOD scale, where errors and biases are additive. (Notice that we add MISR and MODIS biases back onto the predictions to allow a proper comparison to the testing data in the reserved region.)

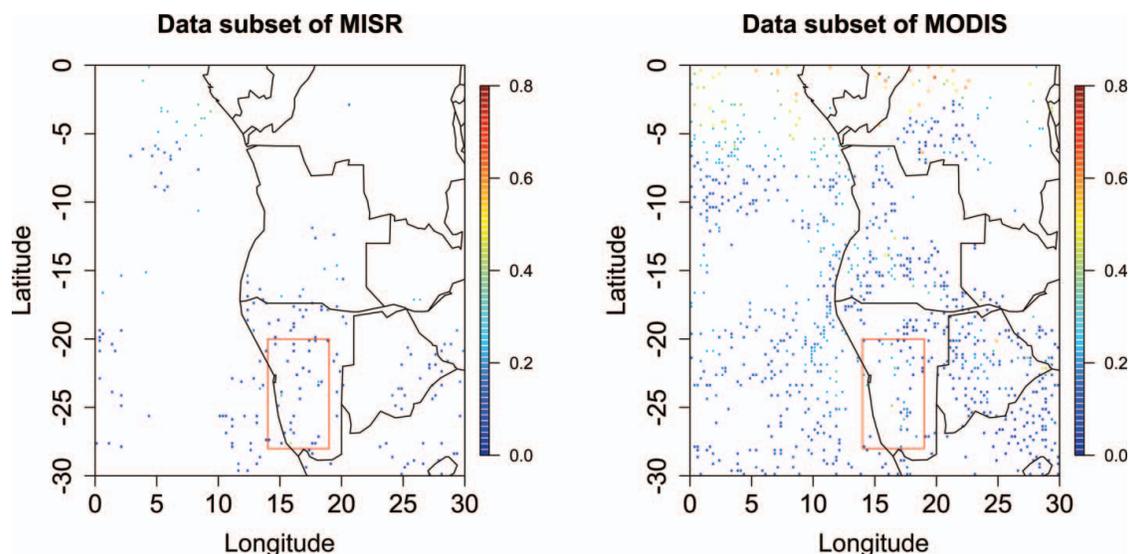


Figure 7. Subsets of MISR and MODIS AOD used for Bayesian melding. The red rectangles indicate the reserved region from which we withheld data for testing. The online version of this figure is in color.

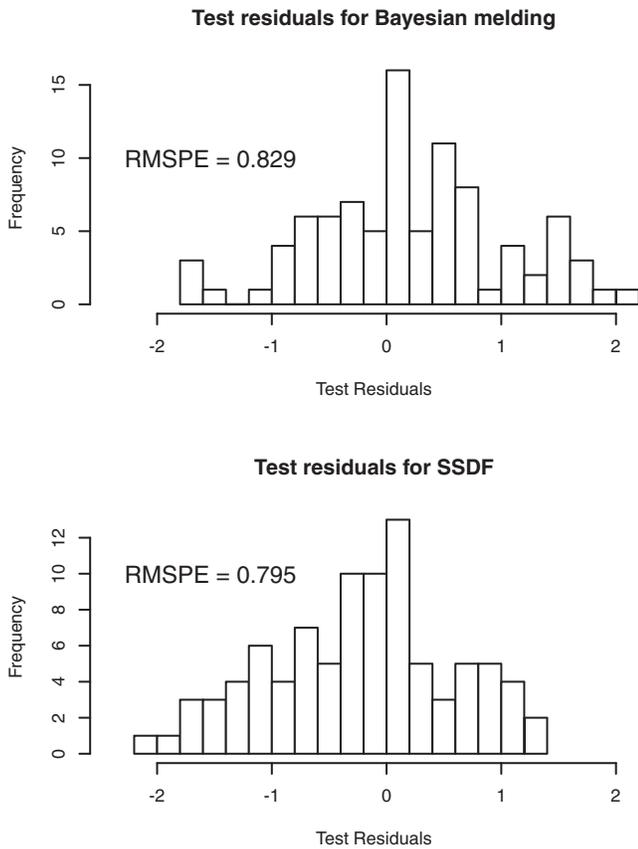


Figure 8. Top panel: test residuals for Bayesian melding. Bottom panel: test residuals for SSDF. The residuals are on the log-transformed AOD scale, and the empirical root mean squared prediction error (RMSPE) is shown on the upper left of each panel.

In applying Bayesian melding to the log-transformed AOD data, we removed the additive biases. Following Fuentes and Raftery (2005), we assumed that the process has a Matérn stationary covariance function. We used gamma priors for the range and smoothness parameter. For the smoothness parameter, the prior mean was 0.05 and the variance 1, and for the range parameter the prior mean was 400 km and the standard deviation was 50 km. For the sill parameter,  $\zeta$ , the prior was  $p(\zeta) \propto \zeta^{-1}$ , which is an improper uniform distribution on  $\log(\zeta)$ . As with SSDF, we added the MISR and MODIS log-transformed additive biases back onto the Bayesian-melding predictions before comparing to the testing data in the reserved region.

For each of the two methodologies, we grouped the test residuals for both MISR and MODIS into a single vector. Figure 8 displays the histograms of the combined test residuals for both SSDF and Bayesian melding. Bayesian melding has an empirical RMSPE of 0.829, and SSDF has a corresponding value of 0.795, which is about 4% smaller. Note that this exercise compares model outputs against withheld data ( $\mathbf{Z}$ ) instead of the truth ( $\mathbf{Y}$ ), so the test residuals reflect variability due to measurement errors as well. It seems that Bayesian melding is able to achieve almost the same predictive performance as that of SSDF. We now compare the computational complexity of the two methodologies.

As expected from the discussion in Section 2, the SSDF algorithm is very fast, taking about 48 sec to process a combined

dataset of about 50,000 observations on a 3.06 GHz machine with an Intel Dual Core processor. About 93% of that time (45 sec) was spent constructing the input matrices  $\{\mathbf{S}_k\}$  (about 9 sec),  $\mathbf{E}$  (about 30 sec), and estimating the parameters  $\mathbf{K}$ ,  $\sigma_\xi^2$ , and  $\{\sigma_{\epsilon,k}^2\}$  (about 5 sec, most of which was spent computing and lifting the eigenvalues of the binned empirical covariance and cross-covariance matrices). Computing the SSDF predictions and the MSPEs takes up the remaining 7% (about 3 sec). On the other hand, Bayesian melding takes about 80 min, since the methodology requires a very large number of stochastic integrals to be computed. The computational complexity of Bayesian melding is  $O((N_1 + N_2)^3)$ , so increasing the sample size beyond a few thousands would quickly produce an overwhelming computational burden. Clearly, doubling the input data sizes for Bayesian melding would lead to an eight-fold increase in computational burden. On the other hand, there would only be a two-fold increase in computational burden for SSDF.

We also consider algorithm storage requirements. Bayesian melding requires construction of a  $(N_1 + N_2) \times (N_1 + N_2)$  covariance matrix  $\Sigma$ , which requires  $8(N_1 + N_2)^2$  bytes of RAM assuming 64-bits double precision. Thus, its storage requirement grows quadratically with the data sizes. For instance, constructing the full covariance matrix  $\Sigma$  for a data vector of size 5000 would require 200 megabytes of RAM. On the other hand, the main computational limitation for SSDF is its storage of  $\{\mathbf{S}_k\}$ , which requires  $8(N_1 + N_2)r$  bytes of RAM. Use of multiresolutional basis functions that result in sparse  $\{\mathbf{S}_k\}$  can significantly improve use of the available memory. For instance, the bisquare basis functions used in this article typically result in zeros for about 320 of the 340 basis-function entries (about 94%) for a generic spatial location, and thus requires only  $(0.06) \cdot 8(N_1 + N_2)r$  bytes of RAM. For instance, construction of  $\mathbf{S}_F$  for a data vector of size 5000 and  $r = 340$  would require only 0.81 megabytes of RAM.

## 5. SUMMARY AND DISCUSSION

This article examines an application of estimating AOD over southern Africa from two remote sensing instruments—MISR and MODIS. We introduce SSDF, a best linear unbiased prediction methodology for estimating a true spatial process from two (conditionally) independent datasets. Our methodology is designed to handle massive datasets with incompatible supports, which are typical in remote sensing. In our application to fusing AOD from the MISR and MODIS instruments on board NASA's Terra satellite, we show that SSDF produces more precise estimates than are produced by applying optimal spatial statistical interpolation (i.e., FRK) to the input datasets separately. The predictions produced by data fusion are more complete and more representative of the underlying process than predictions made from either of the input datasets alone. This implies that aerosol climatologies resulting from SSDF are preferable as candidates for downstream input into climate models. We compare SSDF to Bayesian melding, and while we show that SSDF has slightly higher efficiency, we also show that it has much faster computational speed and much lower storage requirements. In fact, Bayesian melding is not scalable, and so it is not useful for applications involving large-to-massive datasets.

It is important to note that SSDF relies on very large data sizes to derive estimates of its parameters. In other words, there needs to be enough information in  $\mathbf{Z}$  to obtain stable estimates of the  $r \times r$  matrix  $\mathbf{K}$ . A larger number of basis functions,  $r$ , will allow for more flexibility in the SSDF fit, but it also requires more data values in  $\mathbf{Z}$  for stable estimates. An active area of research is selecting the appropriate class of basis functions for  $\mathbf{S}(\cdot)$  and the optimal number of basis function  $r$  (see Bradley, Cressie, and Shi 2011). In practice, we recommend using diagnostic plots such as those in Figure 4 to assess the fit between the empirical and theoretical semivariograms when applying SSDF.

In this article, we have chosen to estimate the SSDF parameters (using the binned method-of-moments). Hence, the prediction standard errors given in Section 3.3 do not include variability due to parameter estimation. A Bayesian approach is to put priors on  $\mathbf{K}$ ,  $\sigma_\xi^2$ , and  $\{\sigma_{\epsilon,k}^2\}$  and use a Markov chain Monte Carlo (MCMC) algorithm to produce optimal predictions and associated posterior uncertainties (e.g., Kang and Cressie 2011; Katzfuss and Cressie 2012). While the Bayesian approach can provide better prediction performance and more accurate measures of uncertainty, its overall computational burden using MCMC is usually too onerous for use with large remote sensing datasets.

Reduced-rank spatial models can be generalized to the spatio-temporal setting (Wikle et al. 2001). The SRE model,  $v(\mathbf{s}) = \mathbf{S}(\mathbf{s})'\boldsymbol{\eta} + \xi(\mathbf{s})$ , can be generalized to a Spatio-Temporal Random Effects model,  $v(\mathbf{s}, t) = \mathbf{S}_t(\mathbf{s})'\boldsymbol{\eta}(t) + \xi(\mathbf{s}, t)$ , where  $\{\boldsymbol{\eta}(t) : t = 0, 1, 2, \dots\}$  is an  $r$ -dimensional time series with mean  $\mathbf{0}$  and  $\text{cov}(\boldsymbol{\eta}(t_1), \boldsymbol{\eta}(t_2)) = \mathbf{K}(t_1, t_2); t_1, t_2 = 1, 2, 3, \dots$  (Cressie, Shi, and Kang 2010). This will naturally lead to a spatio-temporal statistical data-fusion methodology that we expect to have better prediction properties than the spatial-only version, but at the cost of building a more complex statistical model (Nguyen et al. 2011). Before this could become a practical technology that the remote sensing community might use for a given set of instruments, the trade-off between prediction efficiency and model/computational complexity would need to be quantified.

### APPENDIX A: ESTIMATING THE SSDF PARAMETERS

To apply SSDF, we need to specify a set of spatial basis functions for  $\mathbf{S}(\cdot)$ , and we need to estimate the variance-covariance parameters  $\mathbf{K}$ ,  $\sigma_\xi^2$ , and  $\{\sigma_{\epsilon,k}^2 : k = 1, 2\}$  from detrended data. Recall from Section 3.3 that  $\mathbf{E}(\mathbf{Z}_F) = \mathbf{C}\mathbf{T}_F\boldsymbol{\alpha}$ . In the absence of prior knowledge about spatial-covariance structure, we detrend using the ordinary least squares estimator,

$$\hat{\boldsymbol{\alpha}} = (\mathbf{T}'_F \mathbf{C}\mathbf{C}\mathbf{T}'_F)^{-1} \mathbf{C}\mathbf{T}'_F \mathbf{Z}_F.$$

With this estimate of  $\boldsymbol{\alpha}$ , we can calculate the detail residuals,

$$\mathbf{D}_k \equiv \mathbf{Z}_k - (1 + c_k)\mathbf{T}_k\hat{\boldsymbol{\alpha}}; \quad k = 1, 2,$$

which reflect only the SRE process, the fine-scale variation component, and the measurement errors. Then  $\mathbf{D}_k \equiv (D_k(B_{k1}), \dots, D_k(B_{kN_k}))'$ , where  $D_k(B_{km})$  denotes the detail residual for footprint  $B_{km}$  from instrument  $k; m = 1, \dots, N_k$ .

The parameters  $\sigma_\xi^2$  and  $\{\sigma_{\epsilon,k}^2\}$  are estimated by computing the variogram and cross-variogram for small lag distance,  $h$ . We first compute a multiple-datasets generalization of the robust variogram estimator discussed by Cressie and Hawkins (1980) and Cressie (1993, sec. 2.4),

as follows. For  $h > 0$ , define

$$2\bar{\gamma}_{kl}(h) \equiv \frac{\left\{ \frac{1}{|N_{kl}(h)|} \sum_{N_{kl}(h)} |D_k(B_{km}) - D_l(B_{ln})|^2 \right\}^{\frac{1}{2}}}{0.457 + \frac{.494}{|N_{kl}(h)|}}, \quad (\text{A.1})$$

where  $2\bar{\gamma}_{kl}(\|\mathbf{h}\|) \equiv \text{var}(Z_k(\mathbf{s} + \mathbf{h}) - Z_l(\mathbf{s}))$  is the theoretical cross-variogram (Ver Hoef and Cressie 1993), and

$$N_{kl}(h) \equiv \{(B_{km}, B_{ln}) : \|\mathbf{p}(B_{km}) - \mathbf{p}(B_{ln})\| = h; m = 1, \dots, N_k; n = 1, \dots, N_l; k \leq l = 1, 2\}. \quad (\text{A.2})$$

In Equation (A.2),  $\mathbf{p}(B)$  is the centroid of a footprint  $B$ , and  $h$  represents the spatial lag. In practice, the set  $N_{kl}(h)$  is defined using a small tolerance interval around  $h$ , since it may not be possible to find pairs of locations that are exactly distance  $h$  apart (Cressie 1993, p. 70). The term  $|N_{kl}(h)|$  denotes the number of unique elements in  $N_{kl}(h)$ . Since these formulas will be used here for small lags, the assumption of a locally isotropic cross-variogram is not strong.

To estimate the measurement-error variance for instrument  $k$ , we plot the estimated semivariogram  $\bar{\gamma}_{kk}(h)$  against  $h$ , for  $h$  near the origin. Then, using weighted least squares (Cressie 1993, p. 97), we fit a linear semivariogram,  $\tilde{\gamma}_{kk}(h)$ , since the semivariogram function may be assumed to be linear at the small lags concerned. Kang, Liu, and Cressie (2009) showed that, for areal data,  $\sigma_{\epsilon,k}^2$  can be estimated unbiasedly from the fitted line's intercept:

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

To estimate  $\sigma_\xi^2$ , we use the method-of-moments and equate the cross-variogram,  $2\bar{\gamma}_{12}(h)$ , for the smallest possible lag  $h > 0$ , to the average of the expected lag- $h$  squared differences:

$$2\bar{\gamma}_{12}(h) = \frac{1}{|N_{12}(h)|} \left\{ \sum_{N_{12}(h)} (v_1(B_{1m})\sigma_{\epsilon,1}^2 + v_2(B_{2n})\sigma_{\epsilon,2}^2) + \sigma_\xi^2 \sum_{N_{12}(h)} \left( \frac{1}{|B_{1m} \cap D|} + \frac{1}{|B_{2n} \cap D|} - \frac{2|B_{1m} \cap B_{2n}|}{|B_{1m} \cap D||B_{2n} \cap D|} \right) \right\}.$$

Substituting in  $\{\hat{\sigma}_{\epsilon,k}^2\}$ , and fixing  $h$  as small as possible, we obtain the method-of-moments estimator,

$$\hat{\sigma}_\xi^2 = \left( 2|N_{12}(h)| \cdot \bar{\gamma}_{12}(h) - \sum_{N_{12}(h)} (v_1(B_{1m})\hat{\sigma}_{\epsilon,1}^2 + v_2(B_{2n})\hat{\sigma}_{\epsilon,2}^2) \right) \times \left( \sum_{N_{12}(h)} \left( \frac{1}{|B_{1m} \cap D|} + \frac{1}{|B_{2n} \cap D|} - \frac{2|B_{1m} \cap B_{2n}|}{|B_{1m} \cap D||B_{2n} \cap D|} \right) \right)^{-1}.$$

We use the cross-variogram because it allows us to fix  $h$  to be very small.

Next we compute the binned empirical covariance matrix estimator,  $\boldsymbol{\Sigma}_F^*$  (see Appendix B). With  $\hat{\sigma}_\xi^2$ ,  $\{\hat{\sigma}_{\epsilon,k}^2\}$ , and  $\boldsymbol{\Sigma}_F^*$ , we can estimate the covariance parameter,  $\mathbf{K}$ , as follows. We wish to find a  $\mathbf{K}$  for which the following matrix difference is "small":

$$\boldsymbol{\Sigma}_F^* - \bar{\mathbf{S}}_F \mathbf{K} \bar{\mathbf{S}}_F - \bar{\mathbf{U}}_F, \quad (\text{A.3})$$

where  $\bar{\mathbf{S}}_F$  and  $\bar{\mathbf{U}}_F$  are the binned versions of  $\mathbf{S}_F$  and  $\mathbf{U}_F$ , respectively (see Appendix B). We estimate  $\mathbf{K}$  by minimizing the Frobenius norm (e.g., Hastie 1996) of Equation (A.3). Let  $\mathbf{S}_F = \mathbf{Q}\mathbf{R}$ , where  $\mathbf{Q}$  and  $\mathbf{R}$  are the result of a QR decomposition of  $\mathbf{S}_F$  (Press et al. 1986, sec. 2.10). Then the estimate of  $\mathbf{K}$  is,

$$\hat{\mathbf{K}} = \mathbf{Q}(\mathbf{R}')^{-1}(\boldsymbol{\Sigma}_F^* - \bar{\mathbf{U}}_F)\mathbf{R}^{-1}\mathbf{Q}' \quad (\text{A.4})$$

(see Cressie and Johannesson 2008). Positive-definiteness of  $\hat{\mathbf{K}}$  is discussed in Appendix B. In summary, through Equations (A.1) and (A.4),

we have obtained method-of-moments estimators of  $\mathbf{K}$ ,  $\sigma_{\xi}^2$ , and  $\{\sigma_{\epsilon,k}^2\}$ , which are substituted into the SSDF Equations (32) and (33).

### APPENDIX B: EMPIRICAL COVARIANCE AND EMPIRICAL CROSS-COVARIANCE MATRICES

In this section, we describe the method of binning for computing empirical covariance and empirical cross-covariance matrices. Following Cressie and Johannesson (2008), we bin the data for instrument  $k$  by choosing from the domain a set of  $M_k$  bins ( $M_k \ll N_k$ ) with bin centers  $\{\mathbf{u}_{kg} : g = 1, \dots, M_k\}$ . The choices of bin sizes and centers do not have to be the same for both datasets, but the bin centers,  $\{\mathbf{u}_{kg}\}$ , should provide good coverage of  $D$  in both cases. Around bin center  $\mathbf{u}_{kg}$ , we define a neighborhood,  $\mathbf{N}(\mathbf{u}_{kg})$ , and a set of 0-1 weights,

$$w_{kgm} = \begin{cases} 1 & \text{if } \mathbf{p}(B_{km}) \in \mathbf{N}(\mathbf{u}_{kg}) \\ 0 & \text{otherwise,} \end{cases}$$

where  $m = 1, \dots, N_k$ , and  $\mathbf{p}(B)$  is the centroid of  $B$ . Let  $\mathbf{W}_k$  be an  $N_k \times M_k$  matrix defined by  $\mathbf{W}_k \equiv (\mathbf{w}_{k1}, \dots, \mathbf{w}_{kM_k})$ , where  $\mathbf{w}_{kg} \equiv (w_{kg1}, \dots, w_{kgN_k})'$  is an  $N_k$ -dimensional binary vector indicating which footprint among the  $N_k$  observations from dataset  $k$  is in the neighborhood of  $\mathbf{u}_{kg}$ . The matrix  $\mathbf{W}_k$  is composed of all such vectors for the bin centers  $\{\mathbf{u}_{kg} : g = 1, \dots, M_k\}$  in dataset  $k$ .

Now, define the  $M_k \times M_k$  empirical covariance matrix,  $\hat{\Sigma}_{kk}$ , with elements,

$$\hat{\Sigma}_{kk}(\mathbf{u}_{kg}, \mathbf{u}_{kh}) \equiv \begin{cases} V_D(\mathbf{u}_{kg}) & g = h \\ C_D(\mathbf{u}_{kg}, \mathbf{u}_{kh}) & g \neq h, \end{cases}$$

where

$$C_D(\mathbf{u}_{kg}, \mathbf{u}_{kh}) = \sum_{a=1}^{N_k} \sum_{b=1}^{N_k} w_{kga} w_{khb} D_k(B_{ka}) D_k(B_{kb}) / (\mathbf{w}'_{kg} \mathbf{1}_{N_k}) (\mathbf{w}'_{kh} \mathbf{1}_{N_k}),$$

$$V_D(\mathbf{u}_{kg}) = \sum_{a=1}^{N_k} w_{kga} D_k(B_{ka})^2 / (\mathbf{w}'_{kg} \mathbf{1}_{N_k}).$$

We also define the  $M_1 \times M_2$  empirical cross-covariance matrix,

$$\hat{\Sigma}_{12} \equiv [C_D(\mathbf{u}_{1g}, \mathbf{u}_{2h})],$$

where

$$C_D(\mathbf{u}_{1g}, \mathbf{u}_{2h}) = \sum_{a=1}^{N_1} \sum_{b=1}^{N_2} w_{1ga} w_{2hb} D_1(B_{1a}) D_2(B_{2b}) / ((\mathbf{w}'_{1g} \mathbf{1}_{N_1}) (\mathbf{w}'_{2h} \mathbf{1}_{N_2})).$$

More generally, the binned empirical covariance matrix,  $\hat{\Sigma}_{kl}$ , has dimension  $M_k \times M_l$ . To estimate  $\mathbf{K}$  from these empirical covariance and cross-covariance matrices, we need to concordantly bin the  $r \times N_k$  matrix  $\mathbf{S}_k$ , the  $N_k \times N_l$  matrix  $\mathbf{E}_{kl}$ , and the  $N_k \times N_k$  matrix  $\mathbf{V}_k$ , as follows. The binned versions are  $\tilde{\mathbf{S}}_k = \mathbf{S}_k \mathbf{W}_k \mathbf{B}_k^{-1}$ ,  $\tilde{\mathbf{E}}_{kl} = \mathbf{W}'_k \mathbf{E}_{kl} \mathbf{W}_l \div (\mathbf{W}'_k \mathbf{1}_{N_k} \mathbf{1}'_{N_l} \mathbf{W}_l)$ , and  $\tilde{\mathbf{V}}_k = \mathbf{W}'_k \mathbf{V}_k \mathbf{W}_k \div (\mathbf{W}'_k \mathbf{1}_{N_k} \mathbf{1}'_{N_k} \mathbf{W}_k)$ , where  $\mathbf{B}_k \equiv \text{diag}(\mathbf{1}'_{N_k} \mathbf{W}_k)$ , and  $\div$  represents element-wise division. Notice that the term  $\mathbf{B}_k$  in the expression above was mistakenly missing from the corresponding expression in the article by Cressie and Johannesson (2008); the equivalent expression in the article by Shi and Cressie (2007) has the proper term.

Recall from (28) that the relationship between covariance matrices and  $\mathbf{K}$  is,

$$\Sigma_F - \sigma_{\xi}^2 \mathbf{E}_F - \mathbf{V}_F = \mathbf{S}'_F \mathbf{K} \mathbf{S}_F. \quad (\text{B.1})$$

An empirical, binned version of the left-hand side of Equation (B.1) may not be positive-definite, resulting in an estimate of  $\mathbf{K}$  that is not positive-definite. Following Kang, Cressie, and Shi (2010), we lift the eigenvalues of the left-hand side in Equation (B.1) to yield positive-definiteness while preserving total variability. First, we define binned

versions of  $\Sigma_F$  and  $\mathbf{U}_F$  as follows:

$$\hat{\Sigma}_F \equiv \begin{bmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{bmatrix} \text{ and } \hat{\mathbf{U}}_F \equiv \begin{bmatrix} \bar{\sigma}_{\xi}^2 \bar{\mathbf{E}}_{11} + \bar{\sigma}_{1\epsilon}^2 \bar{\mathbf{V}}_1 & \bar{\sigma}_{\xi}^2 \bar{\mathbf{E}}_{12} \\ \bar{\sigma}_{\xi}^2 \bar{\mathbf{E}}_{21} & \bar{\sigma}_{\xi}^2 \bar{\mathbf{E}}_{22} + \bar{\sigma}_{2\epsilon}^2 \bar{\mathbf{V}}_2 \end{bmatrix}. \quad (\text{B.2})$$

Then, we transform  $\hat{\Sigma}_F - \hat{\mathbf{U}}_F$  into  $\mathbf{A}_F$  by normalization:

$$\mathbf{A}_F \equiv \bar{\mathbf{U}}_F^{-\frac{1}{2}} (\hat{\Sigma}_F - \hat{\mathbf{U}}_F) \bar{\mathbf{U}}_F^{-\frac{1}{2}}.$$

Then, we ‘‘lift’’ the eigenvalues of  $\mathbf{A}_F$  to make it positive-definite (Kang, Cressie, and Shi 2010):

$$\lambda^* = \begin{cases} \lambda & \lambda > \lambda_0 \\ \lambda_0 \exp(a(\lambda - \lambda_0)) & \lambda \leq \lambda_0, \end{cases}$$

where  $a, \lambda_0 > 0$ ;  $a$  is a given constant defined below,  $\lambda$  represents an eigenvalue of  $\mathbf{A}_F$ , and  $\lambda^*$  is the lifted version. The lifting procedure above does not modify eigenvalues larger than the chosen threshold,  $\lambda_0$ . Below this threshold, the lifting procedure shrinks the positive eigenvalues and makes the negative eigenvalues positive. After lifting, all eigenvalues of  $\mathbf{A}_F$  are positive, the original order of the eigenvalues is preserved, and the total variation (defined by the trace) of  $\mathbf{A}_F$  is preserved.

Following Kang, Cressie, and Shi (2010), we choose  $\lambda_0$  to be the  $(M_1 + M_2 - r)/(M_2 + M_2)$  quantile of all the original eigenvalues, where  $M_k$  and  $r$  are defined above. The constant  $a$  is chosen so that total variation defined by the sum of eigenvalues remains unchanged after lifting. That is, choose  $a$  so that

$$\text{trace}(\hat{\Sigma}_F) = \text{trace}(\Sigma_F^*),$$

where

$$\Sigma_F^* \equiv \bar{\mathbf{U}}_F^{\frac{1}{2}} \mathbf{A}_F^* \bar{\mathbf{U}}_F^{\frac{1}{2}} + \bar{\mathbf{U}}_F,$$

and  $\mathbf{A}_F^*$  represents the lifted version of  $\mathbf{A}_F$ . Finally, the lifted version,  $\Sigma_F^*$ , is positive-definite and preserves the total variability of the original matrix  $\hat{\Sigma}_F$ .

[Received October 2010. Revised October 2011.]

### REFERENCES

- Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), ‘‘Gaussian Prediction Process Models for Large Spatial Data Sets,’’ *Journal of the Royal Statistical Society, Series B*, 70, 825–848. [1006]
- Berrocal, V., Gelfand, A. E., and Holland, D. M. (2010), ‘‘A Spatio-Temporal Downscaler for Output From Numerical Models,’’ *Journal of Agricultural, Biological, and Environmental Statistics*, 15(2), 176–197. [1006]
- Bradley, J., Cressie, N., and Shi, T. (2011), ‘‘Selection of Rank and Basis Functions in the Spatial Random Effects model,’’ in *2011 Proceedings of the Joint Statistical Meetings*, Alexandria, VA: American Statistical Association, pp. 3393–3406. [1016]
- Carr, D., Kahn, R., Sahr, K., and Olsen, T. (1998), ‘‘ISEA Discrete Global Grids,’’ *Statistical Computing and Statistical Graphics Newsletter*, 8(8), 31–39. [1008]
- Cressie, N. (1993), *Statistics for Spatial Data* (rev. ed.), New York: Wiley-Interscience. [1010,1012,1016]
- (1996), ‘‘Change of Support and the Modifiable Areal Unit Problem,’’ *Geographical Systems*, 3, 159–180. [1006]
- Cressie, N., and Hawkins, D. M. (1980), ‘‘Robust Estimation of the Variogram,’’ *Journal of the International Association of Mathematical Geology*, 1(12), 115–125. [1016]
- Cressie, N., and Johannesson, G. (2006), ‘‘Spatial Prediction for Massive Datasets,’’ in *Mastering the Data Explosion in the Earth and Environmental Sciences: Proceedings of the Australian Academy of Science, Elizabeth and Frederick White Conference*, pp. 1–11. [1007]
- (2008), ‘‘Fixed Rank Kriging for Very Large Spatial Data Sets,’’ *Journal of the Royal Statistical Society, Series B*, 70(1), 209–226. [1006,1007,1008,1011,1016,1017]

- Cressie, N., Shi, T., and Kang, E. L. (2010), "Fixed Rank Filtering for Spatio-Temporal Data," *Journal of Computational and Graphical Statistics*, 19(3), 724–745. [1016]
- Diner, D. J., Beckert, J. C., Reilly, T. H., Bruegge, C. J., Conel, J. E., Kahn, R., Martonchik, J. V., Ackerman, T. P., Davies, R., Gerstl, S. A. W., Gordon, H. R., Muller, J.-P., Myneni, R., Sellers, R. J., Pinty, B., and Verstraete, M. M. (1998), "Multi-Angle Imaging SpectroRadiometer (MISR) Description and Experiment Overview," *IEEE Transactions on Geoscience and Remote Sensing*, 36, 1072–1087. [1004]
- Fuentes, M., and Raftery, A. E. (2005), "Model Evaluation and Spatial Interpolation by Bayesian Combination of Observations With Outputs From Numerical Models," *Biometrics*, 61(1), 36–45. [1006,1011,1014,1015]
- Furrer, R., Genton, M. G., and Nychka, D. (2006), "Covariance Tapering for Interpolation of Large Spatial Datasets," *Journal of Computational and Graphical Statistics*, 15, 502–523. [1006]
- Gotway, C. A., and Young, L. J. (2002), "Combining Incompatible Spatial Data," *Journal of the American Statistical Association*, 97, 632–648. [1006]
- Hastie, T. (1996), "Pseudosplines," *Journal of the Royal Statistical Society, Series B*, 58, 379–396. [1016]
- Henderson, H., and Searle, S. (1981), "On Deriving the Inverse of a Sum of Matrices," *SIAM Review*, 23, 53–60. [1007]
- IPCC (2001), *Climate Change 2001 : The Scientific Basis*, Geneva: Intergovernmental Panel on Climate Change. [1004]
- Kahn, R. A., Nelson, D. L., Garay, M. J., Levy, R. C., Bull, M. A., Diner, D. J., Martonchik, J. V., Paradise, S. R., Hansen, E. G., and Remer, L. A. (2009), "MISR Aerosol Product Attributes and Statistical Comparisons With MODIS," *IEEE actions on Geoscience and Remote Sensing*, 47(12), 4095–4114. [1006]
- Kang, E. L., and Cressie, N. (2011), "Bayesian Inference for the Spatial Random Effects Model," *Journal of the American Statistical Association*, 106(495), 972–983. [1016]
- Kang, E. L., Cressie, N., and Shi, T. (2010), "Using Temporal Variability to Improve Spatial Mapping With Application to Satellite Data," *Canadian Journal of Statistics*, 38(2), 271–289. [1017]
- Kang, E. L., Liu, D., and Cressie, N. (2009), "Statistical Analysis of Small-Area Data Based on Independence, Spatial, Non-Hierarchical, and Hierarchical Models," *Computational Statistics and Data Analysis*, 53, 3016–3022. [1016]
- Katzfuss, M., and Cressie, N. (2012), "Bayesian Hierarchical Spatio-Temporal Smoothing for Massive Datasets," *Environmetrics*, 23, 94–107. [1016]
- King, M. D., Kaufman, Y. J., Menzel, W. P., and Tanre, D. (1992), "Remote Sensing of Cloud, Aerosol, and Water Vapor Properties From the Moderate Resolution Imaging Spectrometer (MODIS)," *IEEE Transactions on Geoscience and Remote Sensing*, 30, 2–27. [1004]
- 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(4), 423–498. [1006]
- Lohmann, U., and Feichter, J. (2005), "Global Indirect Aerosol Effects: A Review," *Atmospheric Chemistry and Physics*, 5, 715–737. [1004]
- Nguyen, H., Katzfuss, M., Cressie, N., and Braverman, A. (2011), *Spatio-Temporal Data Fusion for Very Large Remote Sensing Datasets*, Technical Report 859, Columbus, OH: Department of Statistics, The Ohio State University. [1016]
- Nychka, D., Wikle, C. K., and Royle, J. A. (2002), "Multiresolution Models for Nonstationary Spatial Covariance Functions," *Statistical Modelling*, 2, 315–331. [1006]
- Paradise, S., Braverman, A., Kahn, R., and Wilson, B. (2007), "Characterizing Error Distributions for MISR and MODIS Optical Depth Data," in *American Geophysical Union, Fall Meeting 2007*, abstract no. A23B-0292. [1005,1011]
- Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1986), *Numerical Recipes: The Art of Scientific Computing*, New York: Cambridge University Press. [1016]
- 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. [1007,1017]
- Stein, M. L., and Jun, M. (2008), "Nonstationary Covariance Models for Global Data," *The Annals of Applied Statistics*, 2, 1271–1289. [1006]
- Ver Hoef, J., and Cressie, N. (1993), "Multivariable Spatial Prediction," *Mathematical Geology*, 25, 219–240. [1016]
- Wikle, C. K., and Berliner, L. M. (2005), "Combining Information Across Spatial Scales," *Technometrics*, 47, 80–91. [1006]
- Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M. (2001), "Spatiotemporal Hierarchical Bayesian Modeling: Tropical Ocean Surface Winds," *Journal of the American Statistical Association*, 96, 382–397. [1016]