NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

10-13

Statistical Modeling of MODIS Cloud Data Using the Spatial Random Effects Model

Aritra Sengupta, Noel Cressie, Richard Frey and Brian H. Kahn

Copyright © 2013 by the National Institute for Applied Statistics Research Australia, UOW.
Work in progress, no part of this paper may be reproduced without permission from the Institute.

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

Statistical Modeling of MODIS Cloud Data Using the Spatial Random Effects Model

Aritra Sengupta [*] Noel Cressie [*†] Richard Frey [‡] Brian H. Kahn [§]

Abstract

Remote sensing of the earth by satellites yields datasets that can be massive in size. To overcome computational challenges, we make use of the reduced-rank Spatial Random Effects (SRE) model in our statistical analysis of cloud mask data from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) instrument on board NASA's Terra satellite, launched in December 1999. A set of retrieval algorithms has been developed by members of the MODIS atmospheric team for detecting clouds. Clouds play an important role in climate studies, and hence an accurate quantification of the the spatial distribution of clouds is necessary. In this paper, we build a statistical model for the underlying clear-sky-probability (or conversely, the cloud-probability) process, and we quantify the uncertainty in our predictions. We consider a hierarchical statistical model for analyzing the cloud data, where we postulate a hidden process for the probability of clear sky that makes use of the SRE model. Its advantages are considerable: It can represent many types of spatial behavior, it permits fast computations when datasets are very large, and it has attractive change-of-support properties.

Key Words: empirical hierarchical model (EHM); massive dataset; optimal spatial prediction; spatial GLMM; uncertainty quantification

1. Introduction

Clouds are generally characterized by higher reflectances and lower temperatures than Earth's surface (Ackerman et al., 2010). They play an important role in climate research and must be accurately described in order to properly assess climatic processes and climate change. The accuracy of remote sensing retrievals of several atmospheric quantities can be affected by cloud contamination of the atmospheric column. If it is highly cloud-contaminated, no retrievals are reported for atmospheric quantities that require a clear sky (e.g., aerosols). The Moderate Resolution Imaging Spectroradiometer (MODIS) offers the opportunity for multispectral approaches to cloud detection.

Our interest is in the MODIS instrument on board the Terra satellite, which was launched by NASA in December 1999. The Level-2 MODIS cloud mask product (Platnick et al., 2003) is produced for pixel arrays at a spatial resolution of 1 km$\times$1 km. Each MODIS product file covers data collected over a five-minute time interval, which is called a granule, that contains data on approximately 2.75×10^6 pixels of 1 km$\times$1 km resolution. In this proceedings paper, a granule of Terra MODIS data will be used to illustrate our statistical-modeling approach. The granule corresponds to June 29, 2006, 12:45 UTC. A true-color composite image of the granule is shown in Figure 1. The processing of this granule is available at the Goddard Data and Information Services Center (DISC) (see `http://daac.gsfc.nasa.gov/`).

[*]Department of Statistics, The Ohio State University
[†]Centre for Statistical and Survey Methodology, University of Wollongong
[‡]Cooperative Institute for Meteorological Satellite Studies, University of Wisconsin-Madison
[§]Jet Propulsion Laboratory, California Institute of Technology

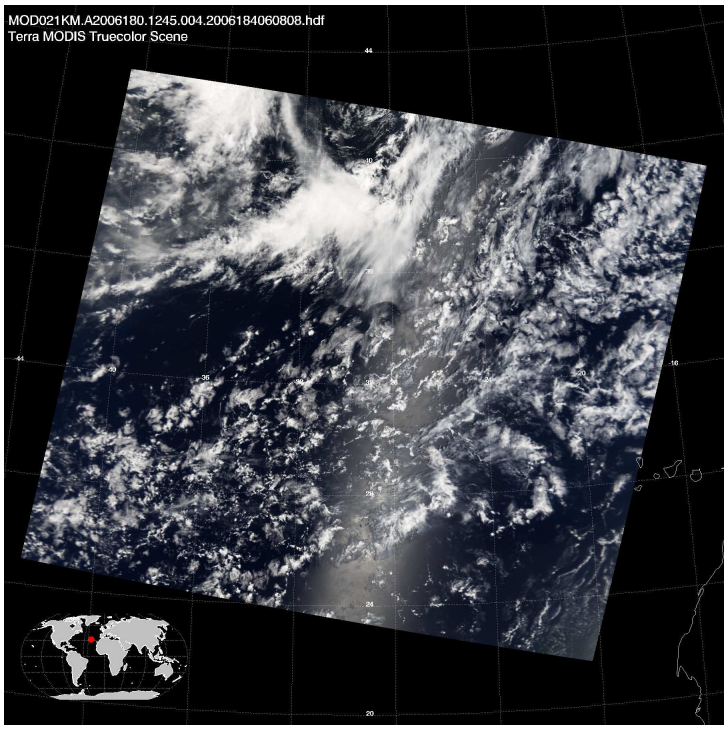

Figure 1: An example of a granule image obtained by the MODIS instrument on board NASA's Terra satellite (June 29, 2006, 12:45 UTC). The inset shows the location of the granule on a world map. (Source: `modis-atmos.gsfc.nasa.gov`.)

The MODIS instrument collects data on spectral radiances that are then processed at NASA using the MODIS cloud detection algorithm (e.g., Platnick et al., 2003; Ackerman et al., 1998, 2010) to produce a Level-2 cloud mask classification (MOD06 product). The MODIS cloud detection algorithm is based on a number of spectral tests; different tests can have different results for a particular pixel. The results from all tests are then combined to determine an overall "confidence," $Q(\mathbf{s})$, for a pixel located at $\mathbf{s}$ to be clear (i.e., cloud free). If $Q(\mathbf{s}) = 1$, it signifies high confidence for the pixel to be clear, and if $Q(\mathbf{s}) = 0$, it signifies high confidence for the pixel to be cloudy. Then, "clear-sky restoral" tests are performed that check for unambiguous clear-sky signals. We denote MODIS's cloud mask product as $Q(\cdot)$, and we review the algorithm that results in $Q(\cdot)$ in Section 2.

In this proceedings paper, we propose a hierarchical spatial statistical model for analyzing MODIS cloud data. Our goal is to produce optimal spatial-prediction maps for the underlying clear/cloudy process, along with measures of prediction uncertainties. We concentrate on the particular granule discussed above (see Figure 1). Our data are the MODIS cloud mask product, $Q(\cdot)$, which is available on 1 km$\times$1 km pixels. Henceforth, each of these pixels will be called a "basic areal unit" (BAU). The number of BAUs in the granule shown in Figure 1 is $N = 2,748,620$.

In general, we assume that we have data for n BAUs, where $n \leq N$. For the particular granule that we consider in this paper, we have $n = N$ (i.e., there are no BAUs without data). A full-rank spatial-statistical modeling approach for the granule would require specifying an $N \times N$ covariance matrix for the underlying spatial (transformed) clear-sky-probability

process. To produce optimal spatial statistical predictions, we would need to invert the $N \times N$ covariance matrix, something that is not computationally feasible for N larger than several thousand.

The computational bottleneck that arises due to the computational cost of inverting the $N \times N$ covariance matrix referred to above, is often referred to as a "big N" problem. When the data appear to be Gaussian, reduced-rank-modeling approaches have been developed to deal with this computational challenge (e.g., Wikle and Cressie, 1999; Wikle et al., 2001; Cressie and Johannesson, 2006, 2008; Banerjee et al., 2008; Stein, 2008; Lopes et al., 2008). For data appearing to come from the exponential family of distributions, Lopes et al. (2011) took the hierarchical generalized linear mixed modeling framework proposed by Diggle et al. (1998), and they introduced a new class of spatio-temporal models using a latent factor-analysis structure; their fully Bayesian model allows for dimension reduction and hence fast computations. A number of spatial and spatio-temporal applications for very-large-to-massive datasets center around these reduced-rank representations of a hidden continuous Gaussian process (e.g., see the review in Wikle, 2010).

To solve the "big N" problem that arises in our application, we shall use the reduced-rank modeling approach developed by Cressie and Johannesson (2006, 2008), although our data are bimodal and constrained to $[0,1]$. Our modeling approach is a combination of the GLMM framework of Diggle et al. (1998) and use of the Spatial Random Effects (SRE) model of Cressie and Johannesson (2006, 2008), although they developed it for Gaussian data with a continuous spatial index. We take an empirical hierarchical modeling (EHM) approach and, unlike a Bayesian hierarchical modeling (BHM) approach, we treat the model's parameters as fixed but unknown. We estimate these parameters using an EM algorithm (e.g., Dempster et al., 1977). Computation of optimal spatial predictions are feasible, and no prior specification of parameters is needed. For a more complete discussion of the EHM and BHM approaches, see Cressie and Wikle (2011, Chapter 2).

Cressie and Johannesson (2006, 2008) developed the Spatial Random Effects (SRE) model for optimal spatial predictions from continuous, symmetric data with a continuous spatial index, a methodology that is known as Fixed Rank Kriging (FRK). Cressie and Johannesson (2008) took an EHM approach and gave a method-of-moments estimator for the parameters of the SRE model, and Katzfuss and Cressie (2009) gave an EM algorithm to obtain maximum-likelihood (ML) estimates. A Bayesian version of the SRE model is given in Kang and Cressie (2011). In Sengupta and Cressie (2012a) and Sengupta and Cressie (2012b), a hierarchical spatial statistical model that includes the SRE model as a component of the process model is developed for big, spatial, discrete, and continuous data.

With regard to applications, the SRE model and the methodologies associated with it have been successful in analyzing massive remote sensing datasets (e.g., Cressie and Johannesson, 2006, 2008; Shi and Cressie, 2007; Kang et al., 2010; Katzfuss and Cressie, 2011). The models were Gaussian and additive. In Sengupta and Cressie (2012b), the SRE model was used in a hierarchical framework to analyze highly skewed, non-negative, remotely sensed Aerosol Optical Depth data, where the models were non-Gaussian and non-additive.

The plan of the rest of this paper is as follows: In Section 2, we describe the MODIS cloud mask product. Details of the hierarchical spatial statistical model are presented in Section 3. In Section 4, we analyze the granule of MODIS cloud data shown in Figure 1, using the modeling framework proposed in Section 3; we produce optimal-spatial-prediction maps for the underlying clear-sky/cloudy process, along with maps showing the prediction uncertainties. Discussion and conclusions follow in Section 5.

2. MODIS Cloud Mask Product

The MODIS cloud mask algorithm (e.g., Ackerman et al., 1998, 2010) identifies different conceptual domains according to surface type and solar illumination. Once a pixel is assigned to a domain, a battery of spectral tests is applied, where each test attempts to detect the presence of cloud in the pixel, by returning a confidence level for the pixel to be clear, ranging from 1 (high-confidence clear), to 0 (low-confidence clear, that is, high-confidence cloudy). Individual spectral tests are based on an upper and lower bound (see below).

Tests capable of detecting similar conditions are grouped together. Denote the total number of groups by N_G, and assume that there are m_i spectral tests within the i-th group; $i = 1, \ldots, N_G$. For the j-th test within the i-th group, if the observed light radiance falls below (above) the lower (respectively, upper) bound, then the clear-sky confidence level, F_{ij}, is 0 (respectively, 1). A pictorial illustration of such a spectral test in the MODIS cloud mask algorithm is given in Figure 2: If the observed radiance of the reflected light falls in the "high-confidence cloudy" region (i.e., below the lower bound), then F_{ij} is assigned a value 0 (i.e., cloudy), and if the observed light radiance falls in the "high-confidence clear" region (i.e., above the upper bound), then F_{ij} is assigned a value 1 (i.e., clear). When the observed value falls in the "intermediate" region (i.e., between the lower and upper bounds), F_{ij} is assigned a value between 0 and 1 using linear interpolation; see Figure 2.

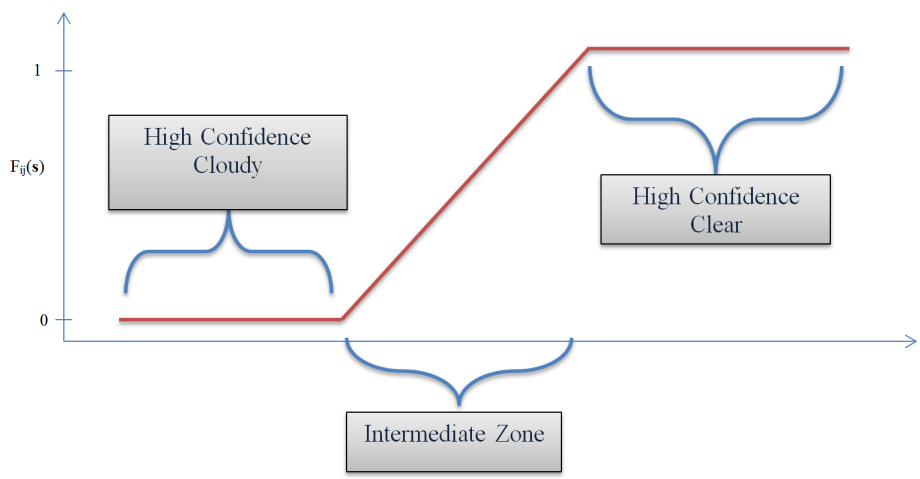

Figure 2: A pictorial illustration of a MODIS cloud mask spectral test, which is based on an upper and lower bound.

For a given pixel, a minimum confidence level is determined for the i-th group as:

$$G_i = min\left\{F_{ij} : j = 1, \ldots, m_i\right\}, \text{ for } i = 1, \ldots, N_G. \tag{1}$$

The overall clear-sky confidence value, Q, for that pixel, is then defined as:

$$Q \equiv \left\{\prod_{i=1}^{N_G} G_i\right\}^{1/N_G}. \tag{2}$$

This approach is clear-sky conservative in the sense that if one of the tests concludes that the pixel is cloudy (i.e., if one $F_{ij} = 0$), then the overall clear-sky confidence value is 0.

The Q-values obtained above (called the "initial" Q-values) are then subject to "clear-sky restoral tests" (e.g., Ackerman et al., 2010; Heidinger, 2010). These tests check for unambiguous clear-sky signals. For example, spectral tests might indicate that a pixel located at $\mathbf{s}$ is cloudy (i.e., $Q(\mathbf{s}) = 0$); but, if all its neighboring pixels are clear, then the pixel is restored as "probably clear" by setting $Q(\mathbf{s}) = 0.96$. Here "cloudy," "probably cloudy," "clear," and "probably clear" are the possible classifications for a pixel, and they are based on thresholding the Q-values (e.g., Platnick et al., 2003). There are other clear-sky restoral tests for different land surfaces, coastal waters, and sun glint. Final Q-values are obtained after applying the clear-sky restoral tests; see Figure 3 for the difference between initial and final Q-values obtained for the granule shown in Figure 1. Noticeable in Figure 1 is a strip of sun glint reflecting off the ocean, which appears in the top panel of Figure 3 (initial Q-values) but not in the bottom panel (final Q-values). Thus, restoral tests are important, since there are geophysical conditions and viewing geometries where the cloud-mask algorithm tends to over-predict clouds (e.g., regions with sun-glint).

In this proceedings paper, we analyze the spatial dataset of *final Q-values* (denoted by $Q(\cdot)$), which we refer to as the *MODIS cloud data*. In the next section, we develop a hierarchical spatial statistical modeling framework that is used in Section 4 for predicting the underlying clear-sky-probability process, given the data. Our approach also allows us to quantify the uncertainty associated with our predictions. These models allow for spatial change-of-support, where our goal is to predict cloud-fraction at any desired resolution coarser than 1 km$\times$1 km; see the discussion in Section 5.

3. Hierarchical Model for the MODIS Cloud Data

In this section, we propose an empirical hierarchical model for final Q-values obtained from the MODIS cloud mask product. We index the the set of BAUs with data as $D_O \equiv \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, and the complimentary set of BAUs without data as $D_U \equiv \{\mathbf{s}_{n+1}, \ldots, \mathbf{s}_N\}$. Hence, our data are $\{Q(\mathbf{s}_i) : i = 1, \ldots n\}$ (see Section 2), where recall that for the granule shown in Figure 1, we have $n = N = 2,748,620$. We introduce a hidden variable $W(\mathbf{s}_i)$, that denotes the state of a pixel, namely 0 or 1 (cloudy or clear), located at $\mathbf{s}_i$; $i = 1, \ldots, N$. Then we assume a hidden spatial process $Y(\cdot)$ that controls the probability of $W(\cdot)$ being 1, where both $W(\cdot)$ and $Y(\cdot)$ are defined over the entire spatial domain, $D \equiv D_O \cup D_U$.

Our hierarchical spatial statistical model consists of a data model and a two-stage process model. We model the pixel-level conditional probabilities, $\{[Q(\mathbf{s}_i)|W(\mathbf{s}_i), \textit{parameters}] : i = 1, \ldots, n\}$, using a "zero-one inflated" Beta distribution. Conditional on $W(\mathbf{s}_i) = 0$, $Q(\mathbf{s}_i)$ will be modeled using a zero-inflated Beta distribution; and conditional on $W(\mathbf{s}_i) = 1$, $Q(\mathbf{s}_i)$ will be modeled using a one-inflated Beta distribution. The zero-one inflation deals with those $\{Q(\mathbf{s}_i)\}$ that are exactly zero or one. Then our *data model* is: For $i = 1, \ldots, n$, independently,

$$[Q(\mathbf{s}_i)|W(\mathbf{s}_i) = 0, P_0, \alpha_0] = \left\{ P_0 I(Q(\mathbf{s}_i) = 0) + (1 - P_0) f_{1,\alpha_0}(Q(\mathbf{s}_i)) \right\}; \tag{3}$$

and, for $i = 1, \ldots, n$, independently,

$$[Q(\mathbf{s}_i)|W(\mathbf{s}_i) = 1, P_1, \alpha_1] = \left\{ P_1 I(Q(\mathbf{s}_i) = 1) + (1 - P_1) f_{1,\alpha_1}(Q(\mathbf{s}_i)) \right\}. \tag{4}$$

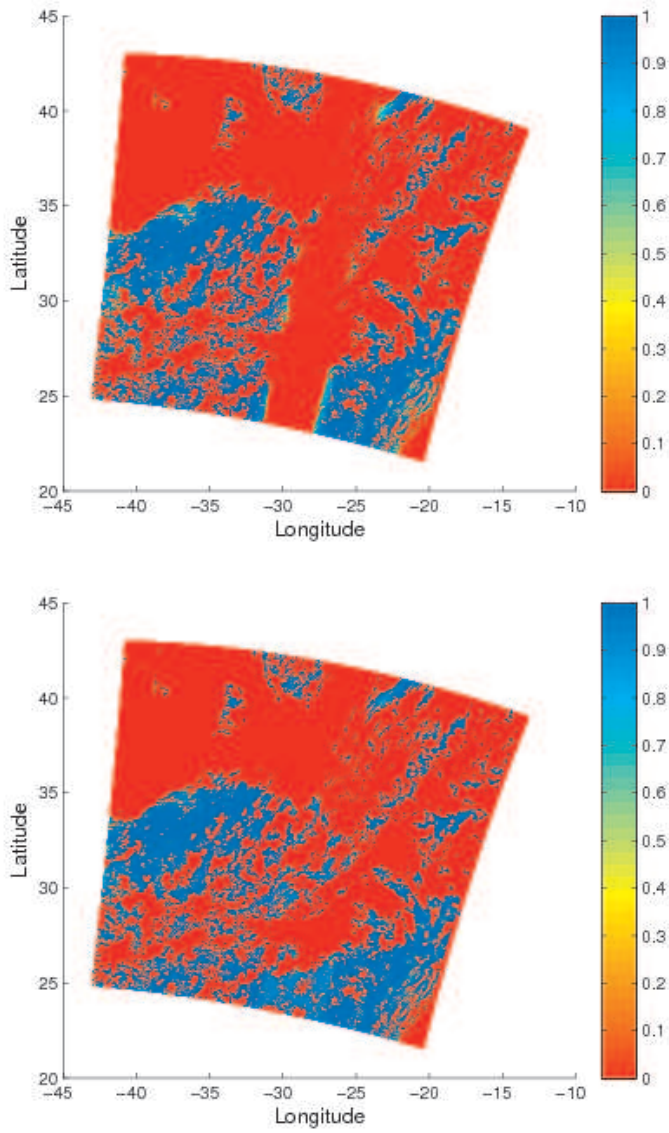

Figure 3: Initial Q-values (top panel) and final Q-values (bottom panel) corresponding to the granule shown in Figure 1.

In (3) and (4),

$$f_{a,b}(Q(\mathbf{s}_i)) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} Q(\mathbf{s}_i)^{a-1}(1-Q(\mathbf{s}_i))^{b-1} I(0 < Q(\mathbf{s}_i) < 1), \tag{5}$$

which is the density of a Beta(a,b) random variable, where $a > 0$ and $b > 0$. The parameters P_0, α_0, P_1, and α_1 in the data model are unknown and need to estimated.

Next, we specify the *two-stage process model*. "Process model 1" represents the distribution of $\{W(\mathbf{s}_i) : i = 1,\ldots,N\}$, conditional on the hidden spatial process $Y(\cdot)$. We assume a set of independent Bernoulli random variables for *process model 1*: For $i = 1,\ldots N$, inde-

pendently,

$$W(\mathbf{s}_i)|Y(\cdot) \sim \text{Bernoulli}\left(\frac{\exp(Y(\mathbf{s}_i))}{1+\exp(Y(\mathbf{s}_i))}\right), \tag{6}$$

where recall that $W(\mathbf{s}_i) = 1$ (respectively, 0) means that the pixel located at $\mathbf{s}_i$ is clear (respectively, cloudy). Then $Y(\cdot)$ is the logit transform of the clear-sky-probability process, $p(\cdot)$, and conversely,

$$p(\cdot) = \left(\frac{\exp(Y(\cdot))}{1+\exp(Y(\cdot))}\right). \tag{7}$$

At the second stage of the process model ("process model 2"), we use the reduced-rank Spatial Random Effects (SRE) model (e.g., Cressie and Johannesson, 2006, 2008) to define the smooth spatial dependence in $Y(\cdot)$. *Process model 2* is:

$$Y(\mathbf{s}_i) = \mathbf{X}(\mathbf{s}_i)^\top \boldsymbol{\beta} + \mathbf{S}(\mathbf{s}_i)^\top \boldsymbol{\eta} + \xi(\mathbf{s}_i);\ i = 1, \ldots N, \tag{8}$$

where $\mathbf{X}(\mathbf{s}_i)$ is a vector of known covariates; $\boldsymbol{\beta}$ denotes the set of unknown regression coefficients; $\mathbf{S}(\cdot) \equiv (S_1(\cdot), \ldots, S_r(\cdot))^\top$ is a vector of r (not necessarily orthogonal) spatial basis functions, where $r << N$ is fixed; $\boldsymbol{\eta}$ is an r-dimensional vector of spatial random effects assumed to have a $\text{Gau}(\mathbf{0}, \mathbf{K})$ distribution, where the covariance matrix $\mathbf{K}$ is unknown and needs to be estimated; $\xi(\cdot)$ is a fine-scale-variation process modeled as independent $\text{Gau}(0, \sigma^2_\xi)$ random variables, where σ^2_ξ is unknown and needs to be estimated.

4. Spatial Statistical Analysis of MODIS Cloud Data

In this section, we carry out a spatial statistical analysis of the granule of MODIS data shown in Figure 1, using the hierarchical model specified in Section 3. For the purpose of this analysis, we selected as basis functions, $\mathbf{S}(\cdot)$, the bisquare functions (e.g., Cressie and Johannesson, 2006, 2008). The generic form of a bisquare function is,

$$b(\mathbf{s}) = \left\{1 - \left(\frac{||\mathbf{s}-\mathbf{c}||}{w}\right)^2\right\}^2 I(||\mathbf{s}-\mathbf{c}|| < w), \tag{9}$$

where $\mathbf{c}$ is the center of the basis function, $I(A)$ is an indicator function that is 1 if A is true, and 0 otherwise. Centers $\{\mathbf{c}_i\}$ in D are usually chosen according to a multi-resolution scheme (e.g., a quad-tree). Then the "aperture" w is given by,

$$w = 1.5 \times \text{shortest great arc distance between like-resolution center points}$$

A pictorial illustration of the bisquare basis function is given in Figure 4. Other choices for basis functions are also possible (e.g., EOFs in Wikle and Cressie, 1999; W-wavelets in Shi and Cressie, 2007).

As in Cressie and Johannesson (2008), we employ several resolutions of the basis functions to capture the different scales of spatial variability; here we use three scales of resolutions to obtain $\{b_i(\mathbf{s}) : i = 1, \ldots, (r_1 + r_2 + r_3)\}$, where $r_1 = 12$, $r_2 = 34$, and $r_3 = 102$, are the number of basis functions at the three resolutions. The centers of the bisquare basis functions were selected using a quad-tree structure (e.g., Cressie and Kang, 2010), ensuring that the centers for the different resolutions do not match. The number of basis functions were determined to ensure full coverage of the spatial domain. We also included centers of the bisquare function outside the study region to account for the boundary effects (e.g.,

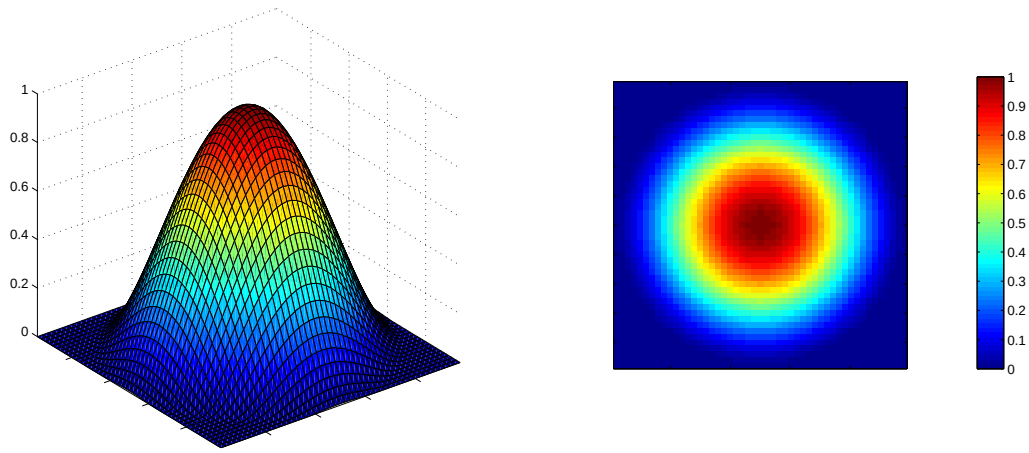

Figure 4: A two-dimensional bisquare function as a 3-D plot (left) and as an image plot (right).

Cressie and Kang, 2010). We further standardized the bisquare function $b_i(\cdot)$ to obtain the i-th basis function,

$$S_i(\mathbf{s}) \equiv \frac{b_i(\mathbf{s}) - \text{ave}_{\mathbf{s}\in D}(b_i(\mathbf{s}))}{\{\text{var}_{\mathbf{s}\in D}(b_i(\mathbf{s}))\}^{1/2}};\ i = 1,\ldots,(r_1+r_2+r_3), \tag{10}$$

where $\text{ave}_{\mathbf{s}\in D}(\cdot)$ and $\text{var}_{\mathbf{s}\in D}(\cdot)$ are spatial moments taken over the domain of interest D. The locations of the basis-function centers for all three resolutions are shown in Figure 5.

Consider now the covariates $\mathbf{X}(\cdot)$ in (8). We include the vector $\mathbf{1}$ and latitude as a covariate. Further, instead of using the coarsest-resolution $S_1(\cdot),\ldots,S_{r_1}(\cdot)$ as spatial basis functions in the SRE model, we use them as covariates in $\mathbf{X}(\cdot)$ (e.g., Shi and Cressie, 2007).

The second term of (8) involves an r-dimensional vector, $\mathbf{S}(\cdot)$, of spatial basis functions, which in our case is made up of the bisquare functions at the second and the third resolutions (see Figure 5). Now, there are regions in the study region that are affected by sun-glint (see Figure 1), which the MODIS cloud algorithm attempts to account for by doing clear-sky restoral tests. Nevertheless, the presence or absence of sun glint is a source of variability that exists for the granule we consider. Hence, we include the sun-glint indicator flag (which takes a value 1 if a pixel is affected by sun glint, and is 0 otherwise) as a column in $\mathbf{S}(\cdot)$. That is, $r = 1+r_2+r_3 = 1+34+102 = 137$.

Recall that our goal is to produce optimal spatial-prediction maps for the underlying clear-sky-probability process, along with measures of prediction uncertainty. This can be achieved by generating samples from the predictive distribution, $[\mathbf{W},\mathbf{Y}|\mathbf{Q}_O,\boldsymbol{\theta}]$, where $\mathbf{Q}_O \equiv (Q(\mathbf{s}_1),\ldots,Q(\mathbf{s}_n))^\top$, $\mathbf{W} \equiv (W(\mathbf{s}_1),\ldots,W(\mathbf{s}_N))^\top$, $\mathbf{Y} \equiv (Y(\mathbf{s}_1),\ldots,Y(\mathbf{s}_N))^\top$, and $\boldsymbol{\theta} \equiv \left\{P_0,\alpha_0,P_1,\alpha_1,\boldsymbol{\beta},\mathbf{K},\sigma^2_\xi\right\}$. (Recall that for the data shown in Figure 3, $n = N$, and hence

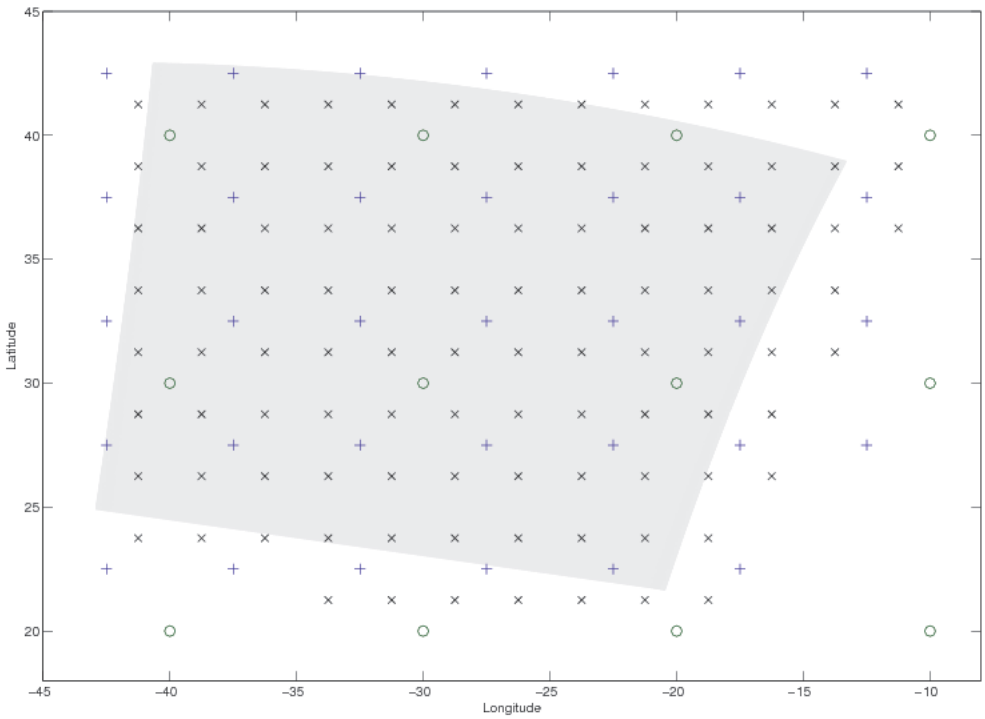

Figure 5: Centers of the basis function; 'o', '+', and 'x' are use to distinguish the three scales of resolution.

the set of pixels where there are no observations, D_U, is empty.) This can be achieved by equivalently generating samples from the predictive distribution, $[\mathbf{W}, \boldsymbol{\eta}, \boldsymbol{\xi} | \mathbf{Q}_O, \boldsymbol{\theta}]$, where $\boldsymbol{\xi} \equiv (\xi(\mathbf{s}_1), \ldots, \xi(\mathbf{s}_N))^\top$. Using Bayes' Theorem, this predictive distribution is

$$[\mathbf{W}, \boldsymbol{\eta}, \boldsymbol{\xi} | \mathbf{Q}_O, \boldsymbol{\theta}] \propto [\mathbf{Q}_O | \mathbf{W}, \boldsymbol{\theta}][\mathbf{W} | \boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{\theta}][\boldsymbol{\eta}, \boldsymbol{\xi} | \boldsymbol{\theta}]. \tag{11}$$

However, due to the unknown proportionality constant (which is a function of the data $\mathbf{Q}_O$), the predictive distribution is not available in closed form, nor are the parameters, $\boldsymbol{\theta}$, known. Here we use a combination of EM estimation of $\boldsymbol{\theta}$ to yield $\hat{\boldsymbol{\theta}}_{EM}$, and an MCMC algorithm (e.g., Robert and Casella, 2004) to yield samples from the (empirical) predictive distribution $[\mathbf{W}, \boldsymbol{\eta}, \boldsymbol{\xi} | \mathbf{Q}_O, \boldsymbol{\theta}]$, where $\hat{\boldsymbol{\theta}}_{EM}$ is substituted in for $\boldsymbol{\theta}$.

The EM algorithm is employed for estimation of the parameters $\boldsymbol{\theta}$; for more details on the methodology, see the review in McLachlan and Krishnan (2008). For the hierarchical model described in Section 3, the process vector $\mathbf{W}$, the random effect $\boldsymbol{\eta}$, and the fine-scale-variation component $\boldsymbol{\xi}$ are not observed, but can be considered as missing data. The EM algorithm involves iterating between an E (expectation) step and an M (maximization) step. Here, the E-step is the most problematic, which we resolve by using Laplace approximations to evaluate the expectations. In the M-step, maximization with respect to (wrt) P_0, P_1, $\mathbf{K}$, and σ^2_ξ is easy and is available in closed form. However, since maximization wrt α_0, α_1, and $\boldsymbol{\beta}$ are not available in closed form, we use a one-step Newton-Raphson update in each of the iterations of the EM algorithm. Technical details of the EM algorithm

used in this and related problems can be found in Sengupta (2012, Ch. 4) and Sengupta and Cressie (2012a,b). Estimates, $\hat{\boldsymbol{\theta}}_{EM}$, obtained for the MODIS cloud data are given in Sengupta (2012, Ch. 4).

Once we obtain the parameter estimates, $\hat{\boldsymbol{\theta}}_{EM}$, we substitute them into the MCMC algorithm to obtain samples from the (empirical) predictive distribution, $[\mathbf{W}, \boldsymbol{\eta}, \boldsymbol{\xi} | \mathbf{Q}_O, \hat{\boldsymbol{\theta}}_{EM}]$. We generated 10,000 MCMC samples, after discarding 1,000 samples as burn-in. Because of storage issues involved with storing the N-dimensional vector $\boldsymbol{\xi}$, we saved every fifth MCMC sample generated. The EM algorithm converged after 14 iterations, and the computational time for the EM algorithm was 27.76 minutes. The computational time for the MCMC was 12.73 hours. All the computations were performed on a dual quad core 2.8 GHz 2x Xeon X5560 processor, with 96 Gbytes of memory.

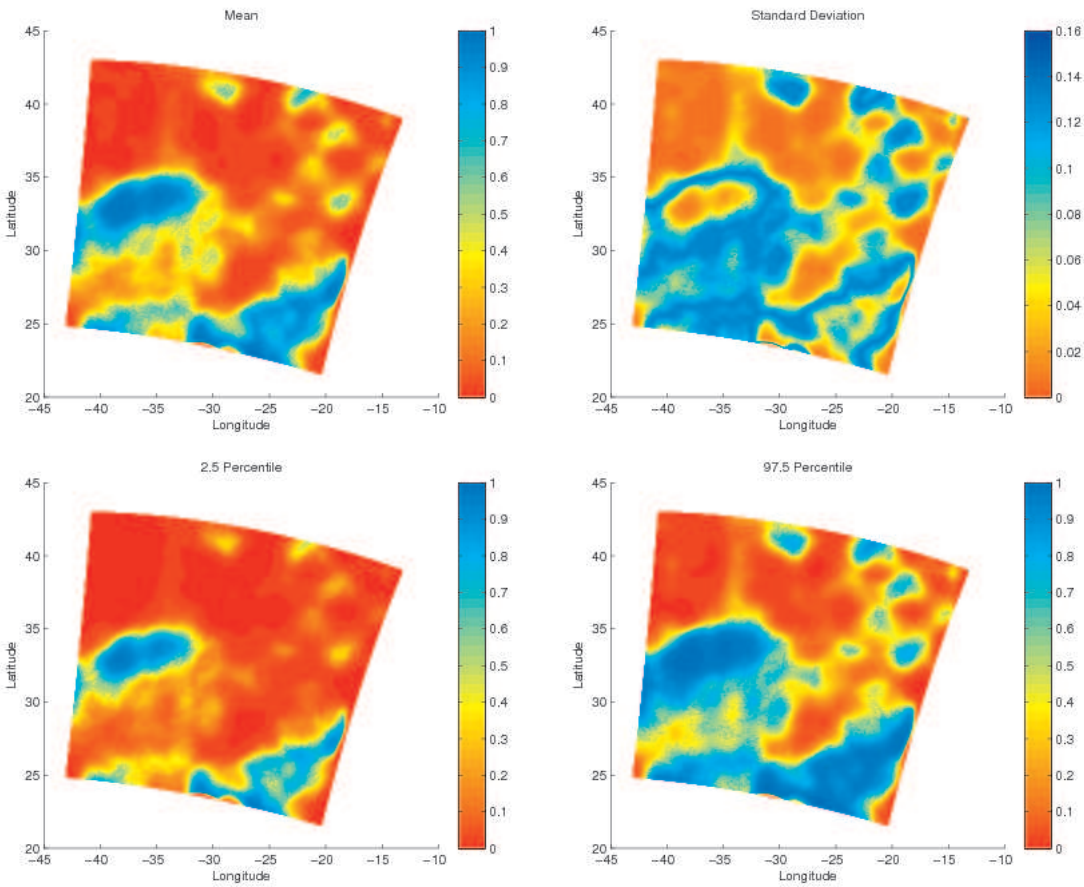

Figure 6: Maps showing the predictive mean (top-left panel), the pixelwise predictive standard-deviation (top-right panel), the pixelwise 2.5 percentile (bottom-left panel) and the pixelwise 97.5 percentile (bottom-right panel) for the predictive distribution of the clear-sky-probability process.

Using the MCMC samples refered to above, we computed the predictive mean and the predictive standard deviation of the clear-sky-probability process, $p(\mathbf{s}_i)$, given by (7). We also obtained the pixelwise 2.5 and 97.5 percentiles of each of the N elements of $\mathbf{p} \equiv (p(\mathbf{s}_1), \ldots, p(\mathbf{s}_N))^\top$. These summaries were obtained from $[\mathbf{p} | \mathbf{Q}_O, \hat{\boldsymbol{\theta}}_{EM}]$. Figure 6 shows

maps of the pixelwise predictive mean, the pixelwise predictive standard deviation, and the pixelwise 2.5 and 97.5 percentiles, respectively; the latter two quantities are the end-points of a pixelwise 95% prediction interval.

5. Discussion

In this proceedings paper, we have developed a hierarchical spatial statistical model for analyzing a remote sensing dataset on clouds from NASA's MODIS instrument. The data are at a very fine scale of resolution (1 km$\times$1 km), and they are massive in size ($n = 2,748,620$). However, use of the reduced-rank SRE model to capture the spatial covariance of the latent process $Y(\cdot)$ allows for very fast computations. For such a massive dataset, we were able to perform EM estimation in 27.76 minutes and then implement the MCMC algorithm in 12.73 hours.

We took an empirical hierarchical modeling (EHM) approach, where the unknown model parameters were estimated using an EM algorithm. Alternatively, one could take a Bayesian hierarchical modeling (BHM) approach, where a prior distribution is put on the parameters. In the context of the SRE model, Kang and Cressie (2011) developed a "Givens angle prior" for $\mathbf{K}$, which could be adapted to the cloud data in much the same way as was done for count data in Sengupta and Cressie (2012b). They found that while the prediction intervals obtained using an EHM approach tended to be too liberal when compared to those using a BHM approach, EHM was an order of magnitude faster.

Within the hierarchical-modeling framework that we developed in this article, we used the SRE model to define an underlying Gaussian field for the hidden process $Y(\cdot)$. These models do not rely on specifying a spatial-weights matrix, and no assumptions of homogeneity, stationarity, or isotropy were made. The SRE model used for $Y(\cdot)$ is particularly adept at handling change-of-support, which involves inferring cloud fraction at any desired scale coarser than 1 km$\times$1 km.

To our knowledge, this is the first attempt to develop a hierarchical spatial statistical model for a cloud dataset at such a fine resolution. The spatial model developed here could be extended to a spatio-temporal setting that might be useful for the evaluation of climate model processes, as well as for improvements in their subgrid-scale physical parameterization. In the long term, we would like to develop data-fusion methodology to incorporate cloud data (e.g., fuse water vapor from AIRS with cloud data from MODIS).

Acknowledgments

This research was supported by NASA's Earth Science Technology Office through its Advanced Information Systems Technology Program. We are grateful to Amy Braverman, Mathias Schreier, and Robert Pincus for their generous contributions to this research.

References

Ackerman, S. A., Strabala, K. I., Menzel, W. P., Frey, R. A., Moeller, C. C., and Gumley, L. E. (1998). "Discriminating clear sky from clouds with MODIS." *Journal of Geophysical Research - Atmospheres*, 103, 32141–32157.

— (2010). "Discrimininiting clear sky from clouds with MODIS algorithm theoretical basis

document (MOD35)." Version 6.1. Website: http://modis.gsfc.nasa.gov/data/atbd/atbd_mod06.pdf.

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.

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

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

Cressie, N. and Kang, E. (2010). "High-resolution digital soil mapping: Kriging for very large datasets." In *Proximal Soil Sensing*, eds. R. A. Viscarra-Rossel, A. B. McBratney, and B. Minasny, vol. 1 of *Progress in Soil Science*, 49–63. Dordrecht: Springer.

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

Dempster, A. P., Laird, N., and Rubin, D. (1977). "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). "Model-based geostatistics." *Journal of the Royal Statistical Society, Series C*, 47, 299–350.

Heidinger, A. (2010). "ABI Cloud Mask." *NOAA NESDIS Center for Satellite Applications and Research; Algorithm Theoretical Basis Document*, Version 2. Website: http://www.goes-r.gov/products/ATBDs/baseline/Cloud_CldMask_v2.0_no_color.pdf.

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

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

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

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

Lopes, H. F., Gamerman, D., and Salazar, E. (2011). "Generalized spatial dynamic factor models." *Computational Statistics and Data Analysis*, 55, 1319 – 1330.

Lopes, H. F., Salazar, E., and Gamerman, D. (2008). "Spatial dynamic factor analysis." *Bayesian Analysis*, 3, 759–792.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. 2nd ed. New York, NY: Wiley-Interscience.

Platnick, S., King, M., Ackerman, S., Menzel, W., Baum, B., Riedi, J., and Frey, R. (2003). "The MODIS cloud products: algorithms and examples from Terra." *IEEE Transactions on Geoscience and Remote Sensing*, 41, 459 – 473.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. New York, NY: Springer.

Sengupta, A. (2012). "Empirical Hierarchical Modeling and Predictive Inference for Big, Spatial, Discrete, and Continuous Data." PhD Thesis, The Ohio State University, Columbus, OH. In preparation.

Sengupta, A. and Cressie, N. (2012a). "Empirical hierarchical modeling for count data using the Spatial Random Effects model." *Spatial Economic Analysis*, forthcoming.

— (2012b). "Hierarchical statistical modeling of big spatial datasets using the exponential family of distributions." Tech. Rep. 870, Department of Statistics, The Ohio State University, Columbus, OH.

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

Stein, M. L. (2008). "A modeling approach for large spatial datasets." *Journal of the Korean Statistical Society*, 37, 3 – 10.

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 and Hall/CRC.

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

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.