The Annals of Applied Statistics
2017, Vol. 11, No. 1, 69–92
DOI: 10.1214/16-AOAS991

MULTIVARIATE SPATIAL MAPPING OF SOIL WATER HOLDING CAPACITY WITH SPATIALLY VARYING CROSS-CORRELATIONS[1]

BY RACHEL M. MESSICK, MATTHEW J. HEATON AND NEIL HANSEN

Brigham Young University

Irrigation in agriculture mitigates the adverse effects of drought and improves crop production and yield. Still, water scarcity remains a persistent issue and water resources need to be used responsibly. To improve water use efficiency, precision irrigation is emerging as an approach where farmers can vary the application of water according to within field variation in soil and topographic conditions. As a precursor, methods to characterize spatial variation of soil hydraulic properties are needed. One such property is soil water holding capacity (WHC). This analysis develops a multivariate spatial model for predicting WHC across a field at various soil depths using sparse WHC observations and covariates such as soil electrical conductivity. To capture spatially varying cross-correlations in an efficient manner, we propose to extend the conditional specification of a multivariate Gaussian process by using spatially varying coefficients. Because data is already sparse, our analysis fully utilizes incomplete observations by imputing missing values that we treat as not missing at random. Additionally, due to the high cost of measuring WHC, we use a multivariate integrated mean square error criterion to choose a new observation location that, after sampling, will result in the least predictive uncertainty across the entire field.

1. Introduction.

1.1. *Research motivation and data.* From 2011 to the present, California has experienced severe drought conditions. A recent assessment report from the National Oceanic and Atmospheric Administration (NOAA) classified California's water resources as "severely depleted" [Seager et al. (2014)]. In addition to drought, increased competition for water resources, aquifer depletion and climate change increase water scarcity for irrigated agriculture. Society's ability to deal with water scarcity while still maintaining sufficient agriculture to support life is dependent upon the efficient use of water; that is, farmers need to efficiently manage their limited water resources by using only the necessary amount of water to grow their crops and allocating additional water for urban use.

Received April 2016; revised September 2016.

[1]This material is based upon work supported by the National Science Foundation Grant DMS-14-17856.

Key words and phrases. Multivariate spatial processes, conditional specification, spatial design, Gaussian process, not missing at random.

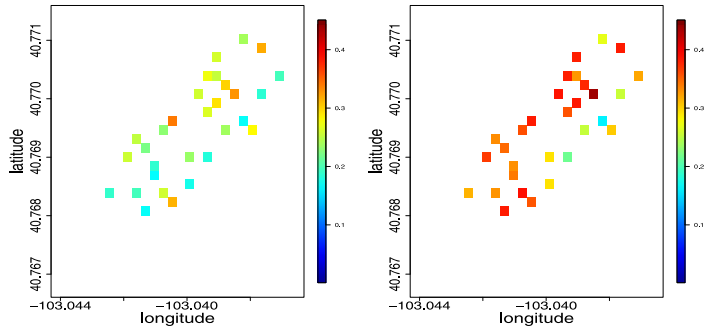
A design goal of a traditional sprinkler irrigation system is to maximize uniformity of water application across the entire irrigated surface. While uniform water application is a technology achievement in itself, irrigated fields are known to vary in topography, soil type and crop productivity [Mzuku et al. (2005)]. Inherent spatial variability can result in a uniform irrigation application being too much or too little water in some areas within the field. Longchamps et al. (2015) showed that soil water content in irrigated fields varies both spatially and temporally, even when fields have been leveled. New technology is available that allows for variable rate irrigation within a field, achieved by varying travel speed and by regulating sections of nozzles along the sprinkler [Sadler et al. (2005)]. Variable rate irrigation has potential to increase efficiency of irrigation water use and to reduce leaching and runoff. To benefit from variable rate irrigation technology, a practical approach is needed to develop prescription maps for variable rate irrigation.

Water holding capacity (WHC) describes how much plant available water a specific soil can store [Allen et al. (1998)] and is determined as the difference between the amount of water in the soil at upper and lower limits. The upper limit, often referred to as the soil field capacity, is the amount of water a well-drained soil retains after being fully wetted and allowed to freely drain. The lower limit, often referred to as the permanent wilting point, is the amount of water remaining in the soil when plants can no longer access it. Soil WHC varies as a function of soil texture. For example, a coarse, sandy textured soil has a lower WHC than silty or clay textured soils because the relatively large soil pores in the sandy soil result in a low upper limit. The WHC of a soil also varies with soil layering by depth. Knowing the variation of WHC for the soils within an irrigated field provides information to the manager about how much water can be supplied to plants from the soil in a specific area of the field and how much irrigation water is required to replenish a depleted soil without leaching. Hence, the goal of this paper is to develop statistical methodology for characterizing the spatial variability of soil available WHC which information can be used in a variable rate irrigation system.

Obtaining measurements of WHC is an expensive and time-consuming process. Water holding capacity can be estimated in the laboratory from soil cores collected from multiple depths and field positions [Klute (1986)] or in the field by measuring water content over a time period long enough to observe a typical range of soil water conditions [Bruce and Luxmoore (1986)]. In the latter case, permanent tubes (that reach a depth of 1.5 meters) must be installed at each location in the field and at regular (e.g., weekly) time intervals; farmers manually insert a neutron probe into each tube to measure, via reflectometry, soil water content at various depths. Thus, the cost and time requirements limit the utility of both of these methods for precision irrigation applications.

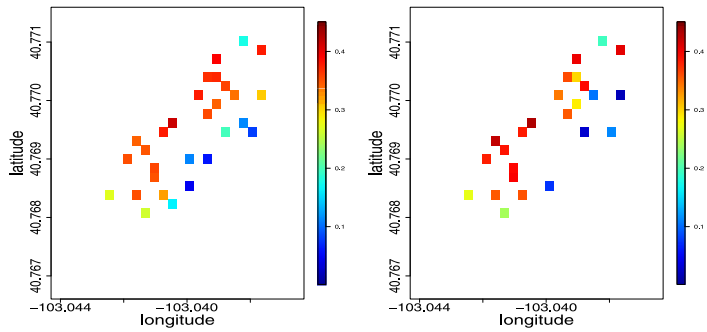
For this research, we consider WHC data collected at 31 different spatial locations across a farm field in Iliff, CO (40°46′ N, 103°2′ W). At each of the 31 latitude/longitude locations, soil water content was measured weekly during the 2012 growing season at up to 5 depths in 0.3 meter increments (the greatest depth being

1.5 m). For each combination of spatial location and soil depth, the maximum observed water content over time was identified as the WHC, without consideration of a lower limit. Plots of the observed WHC data at each depth and the total WHC across depths are provided in Figure 1. All plots and results for WHC are reported



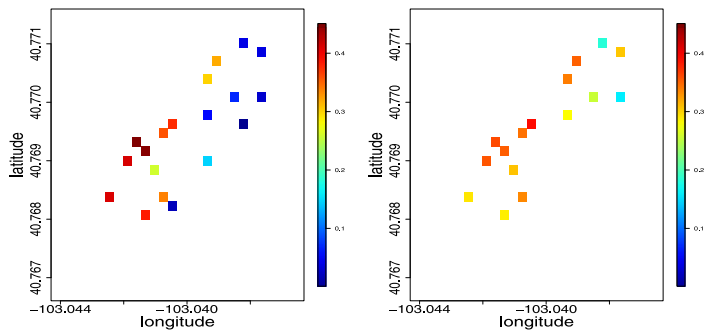
(a) 0-0.3 m

(b) 0.3-0.6 m



(c) 0.6-0.9 m

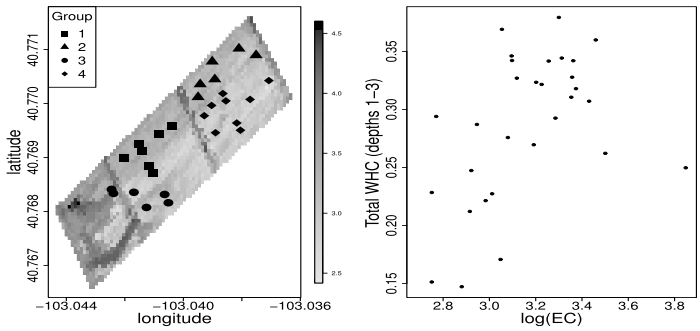
(d) 0.9-1.2 m



(e) 1.2-1.5 m

(f) 0-1.5 m

FIG. 1. *Spatial variation of measured water holding capacity (WHC) in* m^3/m^3 *by depth in* 0.3 *m increments and the total WHC to* 1.5 *m.*



(a) (b)

FIG. 2. (a) *Measured* log(EC) *surface with overlaid WHC sampling locations. Different colored locations correspond to the groupings used to explore spatially-varying correlations.* (b) *Scatterplot of* log(EC) *vs. total WHC* (*first* 3 *depths*) *in* m^3/m^3.

in m^3/m^3, while all analysis was done at the in/ft scale. Of the 31 locations at which WHC is observed, only 17 locations have WHC recorded at each of the 5 depths. Specifically, the data include 31 measurements at depths 1 and 2, 30 measurements at depth 3, 26 measurements at depth 4, and 20 measurements at depth 5.

As displayed in Figure 1, WHC on a single field can vary widely from one point to another. However, due to high monetary and opportunity costs, measuring WHC at many locations across a field is not a reasonable option. Alternatively, obtaining a measurement of the soil electrical conductivity (EC) (a correlated covariate with WHC) is affordable and is already common practice among farmers [Kitchen et al. (2003)]. For example, one widely used approach is to directly measure the electrical conductivity of soil with sensors installed on coulter disks that are pulled through the soil with a tractor and couple measurements with GPS coordinates. The left panel of Figure 2 displays log(EC) measurements at 2291 locations across the research field (measurements capture EC between 0 m and 0.75 m in depth) with the colored points indicating the 31 locations at which WHC is also measured (log scale used to lessen the effect of outliers). The right panel of Figure 2 displays a scatterplot of log(EC) versus the total WHC at depths 1–3 (where complete data is available). Note from the left panel of Figure 2 that EC varies over the field, while the right panel displays a positive relationship between EC and WHC. By successfully exploiting the relationship between EC and WHC to predict WHC, farmers would better be able to assess the irrigation needs of their agriculture fields (and hence better manage water resources).

1.2. *Research challenges and contributions.* In this paper, we have two goals: (i) estimate how WHC varies across a field using EC and accurately quantify uncertainty in the predictions and (ii) identify new locations that, if sampled, will reduce uncertainty in the prediction. Goal (i) can be accomplished by statistically

modeling the WHC data but, in doing so, a number of intricacies must be accounted for. First and foremost, recall that, at each spatial location, WHC is observed at 5 different depths. An intuitive approach would be to model WHC as a Gaussian process over a three-dimensional domain (lat-lon-depth). However, soil science postulates that correlation in WHC is more determined by soil type than by depth [see Natural Resources Conservation Service (1997), Plaster (2013)]. The modeling repercussions of this fact are that the correlation would be discontinuous as a function of depth with discontinuities at the depths where the soil type changes. Given the requirement that a correlation function must be positive definite, statistical modeling of a discontinuous correlation function would be difficult. Thus, instead of using a univariate spatial model in three dimensions, we employ a spatial model over a two-dimensional domain, considering each of 5 depths at every spatial location as a multivariate outcome. To complicate modeling issues, however, cross-correlations between WHC at each depth potentially vary over the spatial domain (the so-called spatially varying cross-correlation problem). To illustrate, we clustered the set of locations into four spatially contiguous groups that correspond with the different colored points in Figure 2. Empirical correlations between depths for each group are displayed in Table 1. From Table 1, note that the empirical correlations between depths vary by group (spatial location), showing possible space-varying inter-depth correlations.

While multivariate spatial data models are well developed [Apanasovich and Genton (2010), Apanasovich, Genton and Sun (2012), Gelfand and Banerjee (2010), Gelfand et al. (2004), Genton and Kleiber (2015), Gneiting, Kleiber and Schlather (2010), Royle and Berliner (1999)], methods that account for spatially varying correlations are less so due to the difficulty of ensuring a positive definite

TABLE 1

Empirical cross-correlations between WHC at various depths and spatial locations. Group assignments correspond to those displayed in Figure 2

	Correlation Between Depths			
Depths	**Group 1**	**Group 2**	**Group 3**	**Group 4**
(1, 2)	0.77	0.67	0.43	0.99
(1, 3)	0.73	0.66	0.66	0.73
(1, 4)	0.71	0.72	0.89	0.30
(1, 5)	0.33	−0.21	−0.79	0.99
(2, 3)	0.75	0.99	0.40	0.73
(2, 4)	0.12	0.98	0.35	0.30
(2, 5)	0.13	0.56	−0.86	0.99
(3, 4)	0.49	0.99	0.92	0.87
(3, 5)	−0.28	0.58	−0.74	0.79
(4, 5)	0.16	0.47	−0.77	0.38

covariance matrix at each location. However, the following are a few notable exceptions. Gelfand et al. (2004) capture spatially varying correlations by allowing the coefficients in a linear model of coregionalization (LMC) to vary over space. Fuentes and Reich (2013) use a spatial stick-breaking prior to construct a spatially varying distribution for the multivariate process and then smooth the processes with a spatially varying kernel. Guhaniyogi et al. (2013) develop low-rank spatially varying cross-covariance processes that allow for interpolated cross-covariances at arbitrary locations. Majumdar, Paul and Bautista (2010) use kernel convolutions to build nonstationary cross-covariances. Sang, Jun and Huang (2011) use parametric regression to predict cross-covariances dependent upon informative predictors or covariates. Kleiber and Genton (2013) take a more theoretical approach and derive sufficient conditions for positive definiteness of a spatially varying cross-covariance matrix. From a practical standpoint, many of these existing methods require large data sets to estimate the associated parameters. For this research, we take a fundamentally different and novel approach to the problem of spatially varying correlations by using spatially varying coefficients [Gelfand et al. (2003)] in a conditional specification for multivariate spatial fields [Cressie and Zammit-Mangion (2015), Royle and Berliner (1999)]. This approach is not only computationally simple and interpretable but also allows for low-rank representations of the cross-correlations through basis function expansions. Furthermore, these methods can be effectively applied in small data settings.

A second notable challenge in modeling WHC is the presence of incomplete observations. Recall that, of the 31 locations where WHC is observed, only 17 locations have WHC recorded at each of the 5 depths and the remaining 14 locations have varying degrees of missing data. Exclusion of all incomplete data points would eliminate 14 locations from the analysis (45% of locations). According to knowledge from those who collected the data, a measurement is missing if the resulting WHC is very low, which suggests a not missing-at-random mechanism [NMAR; Little and Rubin (2002)]. For this analysis, we assume that a WHC measurement is missing if it falls below a threshold of 0.01 (the smallest observed WHC). We subsequently adopt a Bayesian approach to impute the missing values, working under the constraint that such values must lie between 0 and 0.01.

Given the relatively sparse spatial data available for accomplishing goal (i), various regions in the spatial domain may have undesirably high uncertainty in WHC. For this reason, additional observations of WHC may be desired to rein in uncertainty. Due to the high cost of data collection, farmers are often hesitant to allow further substantial data collection. Hence, given the scenario under which such data is collected, we desire to find the single sampling location on the spatial field that, when WHC is measured, will result in the largest decrease in predictive uncertainty [Goal (ii) above]. This "sequential design" problem is discussed by Santner, Williams and Notz (2003) who outline a collection of sequential spatial design techniques (e.g., space-filling and criterion-based designs). Additionally, the com-

puter science literature considers the issue of sensor placement [see, e.g., Krause, Singh and Guestrin (2008) and the references therein]. Notably, relevant sampling designs, such as space-filling and Latin hypercube designs [Johnson, Moore and Ylvisaker (1990), Sacks, Schiller and Welch (1989)], are primarily useful for selecting initial sampling locations in that they do not incorporate knowledge of the surface already learned from observations. Likewise, the most common follow-up designs in spatial and geospatial statistics based on prediction error, expected improvement [Kleiber et al. (2013)], entropy [Currin et al. (1991)] or integrated mean square error [IMSE; Ranjan et al. (2011)] are primarily for univariate random variables. Here, we propose a simple multivariate extension of the IMSE criterion of Ranjan et al. (2011) to select follow-up locations that reduce prediction uncertainty.

To reiterate, the primary statistical contributions of this article are to (i) propose a conditional model for multivariate spatial processes that incorporates spatially varying cross-correlation through the use of spatially varying coefficients and (ii) extend the spatial IMSE design criterion of Ranjan et al. (2011) to the multivariate setting. Additionally, in terms of agricultural science, this article seeks to help farmers understand the variation in WHC across a field using limited data. This understanding will give farmers more information in efficiently allocating scarce water resources. The remainder of this paper is outlined as follows. Section 2 outlines our statistical model, and Section 3 describes the multivariate IMSE criterion for selecting additional observation locations. Section 4 applies our model and IMSE criterion to the WHC data, and Section 6 provides discussion and additional areas of research.

2. A spatial model for water holding capacity. Let $y_1(\boldsymbol{s}), \ldots, y_5(\boldsymbol{s})$ represent WHC at spatial location $\boldsymbol{s} \in \mathcal{S}$ measured at depths 1 through 5, respectively. Using a conditional specification, we represent the likelihood as

$$\begin{aligned}[\boldsymbol{y}_{1:5}(\boldsymbol{s})] &= [y_1(\boldsymbol{s})|\boldsymbol{y}_{2:5}(\boldsymbol{s})][y_2(\boldsymbol{s})|\boldsymbol{y}_{3:5}(\boldsymbol{s})] \\ &\quad \times [y_3(\boldsymbol{s})|\boldsymbol{y}_{4:5}(\boldsymbol{s})][y_4(\boldsymbol{s})|y_5(\boldsymbol{s})][y_5(\boldsymbol{s})],\end{aligned} \tag{1}$$

where $[\cdot]$ denotes an arbitrary distribution and $\boldsymbol{y}_{i:j}(\boldsymbol{s}) = (y_i(\boldsymbol{s}), \ldots, y_j(\boldsymbol{s}))'$. We assume a Gaussian process model for each depth and include the WHC at subsequent depths in the process mean. Specifically, we model, for $j = 1, \ldots, 5$,

$$\begin{aligned}y_j(\boldsymbol{s})|\{y_k(\boldsymbol{s}) : k > j\} &= \boldsymbol{x}_j'(\boldsymbol{s})\boldsymbol{\beta}_j + \sum_{k>j} \gamma_{jk}(\boldsymbol{s})\big(y_k(\boldsymbol{s}) - \boldsymbol{x}_k'(\boldsymbol{s})\boldsymbol{\beta}_k\big) + \cdots \\ &\quad + w_j(\boldsymbol{s}) + \varepsilon_j(\boldsymbol{s}),\end{aligned} \tag{2}$$

where $\boldsymbol{x}_j(\boldsymbol{s})$ is a P-vector of covariates (with intercept) specific to depth j measured at location $\boldsymbol{s}$ (e.g., the EC measurements) with coefficients $\boldsymbol{\beta}_j$, $\gamma_{jk}(\boldsymbol{s})$ is the location-specific "loading" of depth k onto depth j, $w_j(\boldsymbol{s})$ is a mean zero, Gaussian process, independent of all other quantities, where $\mathbb{C}\mathrm{ov}(w_j(\boldsymbol{s}_1), w_j(\boldsymbol{s}_2)) =$

$\sigma_j^2 M_{\nu_j}(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|\phi_j)$ is the Matérn covariance function with smoothness ν_j and decay parameter ϕ_j, and $\varepsilon_j(\boldsymbol{s})$ is a mean zero, white noise Gaussian process, also independent of all other quantities, with variance τ_j^2. An additional error term accounting for measurement error might be appropriate in the case that measurement error variance is known, but, as we have no knowledge of the measurement error variance in this application [if included, the measurement error would be confounded with $\varepsilon_j(\boldsymbol{s})$], we utilize $\varepsilon_j(\boldsymbol{s})$ to account for both small-scale spatial variation and measurement error. Note that in (2) we choose to order the conditioning from the deepest point to the shallowest (e.g., depth 4 is conditional on depth 5, depth 3 is conditional on depth 4 and 5, etc.). Cressie and Zammit-Mangion (2015) note that this ordering is arbitrary, but we validate this decision via DIC comparison reported in Section 4. It should also be noted that this model assumes only pointwise interaction between depths—a justifiable assumption due to the natural horizontal layering of the soil; that is, soil types are more similar horizontally than vertically. Hence, knowing the soil type (or, in this case, the WHC) of the soil layer directly below is sufficient. Additionally, from a statistical perspective, including more than pointwise interactions can greatly increase the parameter space. Given the small amount of available data, it is likely that we would be unable to estimate such a large number of parameters.

Given the large number of parameters in (2), it is reasonable to consider employing the Markov assumption, such that (2) becomes

$$\begin{aligned} y_j(\boldsymbol{s})|\{y_{j+1}(\boldsymbol{s})\} &= \boldsymbol{x}_j'(\boldsymbol{s})\boldsymbol{\beta}_j + \gamma_{j(j+1)}(\boldsymbol{s})\big(y_{j+1}(\boldsymbol{s}) - \boldsymbol{x}_{j+1}'(\boldsymbol{s})\boldsymbol{\beta}_{j+1}\big) + \cdots \\ &\quad + w_j(\boldsymbol{s}) + \varepsilon_j(\boldsymbol{s}). \end{aligned} \tag{3}$$

The Markov and non-Markov approaches will be formally compared in Section 4, but until that point, without loss of generality, we continue to consider the non-Markov model.

The spatially varying cross-correlations between depths in (2) are captured by the location-specific loadings $\gamma_{jk}(\boldsymbol{s})$. To illustrate, consider $\mathbb{C}\mathrm{ov}(y_4(\boldsymbol{s}_1), y_5(\boldsymbol{s}_2))$ (similar derivations exist for any two locations and any two depths). From (2),

$$y_4(\boldsymbol{s}) = \boldsymbol{x}_4'(\boldsymbol{s})\boldsymbol{\beta}_4 + \gamma_{45}(\boldsymbol{s})\big(y_5(\boldsymbol{s}) - \boldsymbol{x}_5'(\boldsymbol{s})\boldsymbol{\beta}_5\big) + w_4(\boldsymbol{s}) + \varepsilon_4(\boldsymbol{s}),$$

and

$$y_5(\boldsymbol{s}) = \boldsymbol{x}_5'(s)\boldsymbol{\beta}_5 + w_5(\boldsymbol{s}) + \varepsilon_5(s).$$

Standard algebraic manipulations yield

$$\begin{aligned} \mathbb{C}\mathrm{ov}\big(y_4(\boldsymbol{s}_1), y_5(\boldsymbol{s}_2)\big) &= \mathbb{E}\big(y_4(\boldsymbol{s}_1)y_5(\boldsymbol{s}_2)\big) - \mathbb{E}\big(y_4(\boldsymbol{s}_1)\big)\mathbb{E}\big(y_5(\boldsymbol{s}_2)\big) \\ &= \gamma_{45}(\boldsymbol{s}_1)\big[\mathbb{E}\big(y_5(\boldsymbol{s}_1)y_5(\boldsymbol{s}_2)\big) - \mathbb{E}\big(y_5(\boldsymbol{s}_1)\big)\mathbb{E}\big(y_5(\boldsymbol{s}_2)\big)\big] \\ &= \gamma_{45}(\boldsymbol{s}_1)\,\mathbb{C}\mathrm{ov}\big(y_5(\boldsymbol{s}_1), y_5(\boldsymbol{s}_2)\big) \\ &= \gamma_{45}(\boldsymbol{s}_1)\big(\sigma_5^2 M_{\nu_5}\big(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|\phi_5\big) + \tau_5^2 \mathbb{1}_{\{\boldsymbol{s}_1 = \boldsymbol{s}_2\}}\big), \end{aligned}$$

where $\mathbb{1}_{\mathcal{A}}$ is an indicator for the set $\mathcal{A}$. Hence, under the conditional specification in (2), the correlation between any two locations $(\boldsymbol{s}_1, \boldsymbol{s}_2)$ at any two corresponding depths (j, k) (for $j < k$) is completely determined by $\gamma_{jk}(\boldsymbol{s}_1)$, which subsequently leads to spatially varying cross-correlations.

Let $\boldsymbol{y}_j = (y_j(\boldsymbol{s}_1), \ldots, y_j(\boldsymbol{s}_n))'$ be the vector of observations measured at depth j. The process model specification in (2) implies a joint distribution of

$$\boldsymbol{y}_j|\{\boldsymbol{y}_k : k > j\} \sim \mathcal{N}\Big(\boldsymbol{X}_j\boldsymbol{\beta}_j + \sum_{k>j} \boldsymbol{D}_k\boldsymbol{\gamma}_{jk}, \sigma_j^2\boldsymbol{M}_j + \tau_j^2\boldsymbol{I}_n\Big), \tag{4}$$

where $\boldsymbol{X}_j$ is the $n \times P$ design matrix with ith row $\boldsymbol{x}_j'(\boldsymbol{s}_i)$, $\boldsymbol{D}_k = \operatorname{diag}(\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta}_k)$ is the $n \times n$ diagonal matrix whose diagonal elements are formed from the error vector $(\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta}_k)$, and $\boldsymbol{\gamma}_{jk} = (\gamma_{jk}(\boldsymbol{s}_1), \ldots, \gamma_{jk}(\boldsymbol{s}_n))'$ is the length n vector of γ coefficients, $\boldsymbol{M}_j$ is the $n \times n$ matrix of correlations at depth j with $i\ell$th element $M_{\nu_j}(\|\boldsymbol{s}_i - \boldsymbol{s}_\ell\|/\phi_j)$, and $\boldsymbol{I}_n$ is the rank n identity matrix. The joint model in (4) is obviously overparameterized because, for depth j, there are $P + (5-j) \times n + 4$ unknown parameters. However, much of this overparameterization can be remedied by using low-rank, basis function representations of the γ coefficients; that is, let $\gamma_{jk}(\boldsymbol{s}) = \gamma_{jk}^\star + \sum_{\ell=1}^{L} b_{jk,\ell}(\boldsymbol{s})\gamma_{jk,\ell}^\star = \boldsymbol{b}_{jk}'(\boldsymbol{s})\boldsymbol{\gamma}_{jk}^\star$, where $\gamma_{jk}^\star$ is an overall mean and $b_{jk,\ell}(\cdot)$ is a basis function with associated coefficient $\gamma_{jk,\ell}^\star$. While many choices of basis functions are available, we recommend those commonly employed in a spatial setting such as bisquare basis functions [Cressie and Johannesson (2008), Kang and Cressie (2011)], predictive processes [Banerjee et al. (2008), Finley et al. (2009)], compactly supported basis functions [Lemos and Sansó (2009), Nychka et al. (2015)] or kernel convolutions [Higdon (2002)]. Using basis function expansions, $\boldsymbol{\gamma}_{jk}$ is represented as

$$\boldsymbol{\gamma}_{jk} = \boldsymbol{B}_{jk}\boldsymbol{\gamma}_{jk}^\star, \tag{5}$$

where $\boldsymbol{B}_{jk}$ is a $n \times (L+1)$ matrix of known basis functions with ith row $\boldsymbol{b}_{jk}'(\boldsymbol{s}_i)$. Substituting (5) into (4) results in

$$\boldsymbol{y}_j|\{\boldsymbol{y}_k : k > j\} \sim \mathcal{N}\big(\boldsymbol{X}_j^\star\boldsymbol{\theta}_j, \sigma_j^2\boldsymbol{M}_j + \tau_j^2\boldsymbol{I}_n\big), \tag{6}$$

where $\boldsymbol{X}_j^\star = [\boldsymbol{X}_j; \boldsymbol{D}_{j+1}\boldsymbol{B}_{j(j+1)}; \ldots; \boldsymbol{D}_5\boldsymbol{B}_{j5}]$ and $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j', \boldsymbol{\gamma}_{j(j+1)}^{\star\prime}, \ldots, \boldsymbol{\gamma}_{j5}^{\star\prime})'$. Using this basis function approach, there are now $P + (5-j) \times L + 4$ unknown parameters. Careful choice of L can ensure that $n > (P + (5-j) \times L + 4)$ and the parameters are well identified.

Using basis function expansions, the spatially varying coefficients model in (4) reduces to the simple spatial model in (6), which facilitates ease in estimation despite a complex dependency structure among the variables. To simplify estimation further, we reparameterize (6) to

$$\boldsymbol{y}_j|\{\boldsymbol{y}_k : k > j\} \sim \mathcal{N}\big(\boldsymbol{X}_j^\star\boldsymbol{\theta}_j, \kappa_j^2\big(\omega_j\boldsymbol{M}_j + (1-\omega_j)\boldsymbol{I}_n\big)\big), \tag{7}$$

where $\kappa_j^2 = \sigma_j^2 + \tau_j^2$ is the total variance, and $\omega_j = \sigma_j^2/(\sigma_j^2 + \tau_j^2)$ is the percent of the total variance attributable to spatial variation. Using this parameterization, the parameters $\boldsymbol{\theta}_j$ and κ_j^2, under certain prior assumptions, will have closed-form complete conditional distributions facilitating sampling in a Markov chain Monte Carlo (MCMC) framework. Additionally, because $\omega_j \in [0, 1]$, discretizing the support of the prior distribution for ω_j to a fine grid over $[0, 1]$ enables direct sampling of ω_j from its complete conditional distribution with minimal loss of information. Of course, a beta prior for ω_j would also be appropriate, but the complete conditional distribution would not be closed form and a Metropolis–Hastings algorithm would be required. Finally, recall that the Matérn covariance $\boldsymbol{M}_j$ is governed by unknown parameters ϕ_j (controlling the spatial decay) and ν_j (controlling the spatial smoothness). Following results from Zhang (2004), we recommend, without loss of predictive power, to simply fix each ν_j and estimate ϕ_j. Traditionally, gamma, inverse-gamma or log-normal priors are used for ϕ_j, but we opt for the computationally simpler discrete uniform prior advocated by Diggle and Ribeiro (2002). Specifically, we construct a discrete uniform prior for ϕ_j by choosing a correlation target, say 0.5, and considering a sequence of distances $\{d_k : 0 < d_1 < d_2 < \cdots < d_K\}$ such that two points d_k units apart have correlation 0.5. We then back transform the distances into corresponding ϕ_j values, resulting in reasonable values for ϕ_j that are given equal prior weight.

3. Sequential design for sampling WHC. The data presented in Section 1 include $n = 17$ locations, where WHC is measured at each depth and $n = 14$ locations with varying degrees of incomplete or missing data. Given the few number of data points for parameter estimation, the model presented in Section 2 may yield predictions at certain spatial locations that have a high degree of uncertainty. To reduce uncertainty in WHC to reasonable values, additional sampling may be required. However, the cost of collecting more WHC data is high, and we wish to ensure that additional sampling locations reduce predictive uncertainty across the entire field; that is, given n locations at which data is gathered, we seek the $(n + 1)$th location which, when appended to the data, minimizes prediction uncertainty. As discussed in the Introduction and validated in Section 5.2, we consider a multivariate extension of the work of Ranjan et al. (2011) who propose the integrated mean square error (IMSE) criterion for selecting additional sampling locations.

Let $\boldsymbol{y}(\boldsymbol{s}_i)$ be the vector of WHC observations at location $\boldsymbol{s}_i$, $\mathcal{Y} = \{\boldsymbol{y}(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ represent the set of WHC observations and $\boldsymbol{\psi}$ be the vector of all parameters in the model described in Section 2 with parameter space Ψ. Furthermore, let $\boldsymbol{A}$ represent a $Q \times 5$ user-specified matrix such that $\boldsymbol{A}\boldsymbol{y}(\boldsymbol{s})$ represents predictive quantities of interest. For example, if $Q = 1$ and $\boldsymbol{A} = (1, 0, 0, 0, 0)$, then $\boldsymbol{A}\boldsymbol{y}(\boldsymbol{s}) = y_1(\boldsymbol{s})$, suggesting that WHC at depth 1 is of predictive interest. Naturally, if $\boldsymbol{A} = \boldsymbol{I}$, then prediction at all 5 depths is of interest. If a new observation

$\boldsymbol{y}(\boldsymbol{s}_0)$ is gathered at $\boldsymbol{s}_0$, then the mean square error (MSE) at location $\boldsymbol{s}$ is given by

$$\text{MSE}(\boldsymbol{s}|\boldsymbol{y}(\boldsymbol{s}_0), \boldsymbol{\psi}) = \mathbb{E}\big[\|\boldsymbol{A}\boldsymbol{y}(\boldsymbol{s}) - \boldsymbol{A}\hat{\boldsymbol{y}}(\boldsymbol{s})\|^2 | \boldsymbol{y}(\boldsymbol{s}_0), \mathcal{Y}, \boldsymbol{\psi}\big], \tag{8}$$

where $\|\cdot\|$ denotes a vector norm (here, we set $\|\cdot\|$ to be the L_2-norm, but other norms may be more suitable depending on the application) and $\hat{\boldsymbol{y}}(\boldsymbol{s}) = \mathbb{E}(\boldsymbol{y}(\boldsymbol{s})|y(\boldsymbol{s}_0), \mathcal{Y}, \boldsymbol{\psi})$ is the predicted value of $\boldsymbol{y}(\boldsymbol{s})$ given the data $\mathcal{Y}$, the "new" observation $\boldsymbol{y}(\boldsymbol{s}_0)$ and the parameters $\boldsymbol{\psi}$. Note that (8) is defined in terms of a vector where $\boldsymbol{A}$ controls the main quantities of predictive interest. A new location $\boldsymbol{s}_{\text{new}}$ is chosen such that

$$\begin{aligned} \boldsymbol{s}_{\text{new}} &= \arg\min_{\boldsymbol{s}_0} \text{IMSE}(\boldsymbol{s}_0) \\ &= \arg\min_{\boldsymbol{s}_0} \int_{\Psi}\int_{\mathcal{S}} \text{MSE}(\boldsymbol{s}|\boldsymbol{y}(\boldsymbol{s}_0), \mathcal{Y}, \boldsymbol{\psi})[\boldsymbol{\psi}|\mathcal{Y}]\, d\boldsymbol{s}\, d\boldsymbol{\psi}, \end{aligned} \tag{9}$$

where $[\boldsymbol{\psi}|\mathcal{Y}]$ is the posterior distribution $\boldsymbol{\psi}$.

The integrals in (9) are not available in closed form. Hence, we approximate these integrals using Riemann and Monte Carlo integration. Specifically, we use the fact that the double integrals in (9) can be expressed as

$$\begin{aligned} &\mathbb{E}_{[\boldsymbol{\psi}|\mathcal{Y}]}\left[\int_{\mathcal{S}} \text{MSE}(\boldsymbol{s}|\boldsymbol{y}(\boldsymbol{s}_0), \mathcal{Y}, \boldsymbol{\psi})\, d\boldsymbol{s}\right] \\ &\quad\approx \mathbb{E}_{[\boldsymbol{\psi}|\mathcal{Y}]}\left[\Delta \sum_{g=1}^{G} \text{MSE}(\boldsymbol{s}_g^\star|\boldsymbol{y}(\boldsymbol{s}_0), \mathcal{Y}, \boldsymbol{\psi})\right], \end{aligned} \tag{10}$$

where $\{\boldsymbol{s}_1^\star, \ldots, \boldsymbol{s}_G^\star\}$ is a regularly spaced grid of locations on $\mathcal{S}$ with equal grid area Δ. Equation (10) suggests an intuitive Monte Carlo approach to calculating IMSE. Namely, given a potential sampling location $\boldsymbol{s}_0$, at each iteration in a Markov chain Monte Carlo algorithm, to sample $\boldsymbol{\psi}$ from its posterior distribution, (i) draw $\boldsymbol{y}(\boldsymbol{s}_0)$ from its predictive distribution, (ii) draw $\{\boldsymbol{y}(\boldsymbol{s}_g^\star) : g = 1, \ldots, G\}$ from the predictive distribution conditional on $\boldsymbol{y}(\boldsymbol{s}_0)$, and (iii) retain $\Delta \sum_{g=1}^{G} \|\boldsymbol{A}\boldsymbol{y}(\boldsymbol{s}_g^\star) - \boldsymbol{A}\hat{\boldsymbol{y}}(\boldsymbol{s}_g^\star)\|^2$. Calculation of (10) is then carried out via Monte Carlo integration using the retained values of $\Delta \sum_{g=1}^{G} \|\boldsymbol{A}\boldsymbol{y}(\boldsymbol{s}_g^\star) - \boldsymbol{A}\hat{\boldsymbol{y}}(\boldsymbol{s}_g^\star)\|^2$. We note that, while this is a straightforward algorithm, it is computationally demanding when considering a large number of possible "new locations."

4. Application to WHC.

4.1. *Spatial mapping of WHC.* Based on the model proposed in Section 2, the unknown parameters are $\{\boldsymbol{\theta}_j, \kappa_j^2, \omega_j, \phi_j\}$ for $j = 1, \ldots, 5$. Noninformative, Jeffery's priors were chosen for $\boldsymbol{\theta}_j$ and κ_j^2, leading to closed-form complete conditional distributions. As mentioned above, the priors for ω_j and ϕ_j were assumed to be discrete uniform with 20 values for each ω_j and 10 values for each ϕ_j. We

note that these discrete uniform prior distributions also lead to closed-form complete conditional distributions for each ω_j and ϕ_j. Though it can be shown that the complete conditional distribution of the constrained missing data follows a truncated multivariate Gaussian distribution, sampling directly from this distribution is complex due to the correlation structure and the NMAR constraint that each value lies in $(0, 0.01)$. Hence, we elect to use a Metropolis algorithm with independent $\mathcal{U}(0, 0.01)$ proposal distributions to update all missing values simultaneously.

To obtain draws from the joint posterior distribution, we ran 5 chains of a Gibbs sampler for 5000 iterations after a burn-in of 200 iterations. Despite the few number of iterations, we found convergence to be acceptable among the estimated parameters; that is, trace plots showed adequate mixing and Gelman–Rubin diagnostics [Gelman and Rubin (1992)] supported convergence. Additionally, most Monte Carlo standard errors [Jones et al. (2006)] fell below 0.01, with a few lying between 0.01 and 0.04. While computation time for parameter estimation and WHC prediction is reasonable, and would easily allow for more iterations, the design portion of the process is computationally expensive. Thus, with parameter convergence being satisfactory even with only 4800 iterations, we find the relatively small number of iterations to be well justified.

In regards to the effect of EC on WHC, we considered both linear and nonlinear models and found that, across all metrics described below, a linear relationship was preferred. Additionally, while continuous basis function expansions of $\boldsymbol{\gamma}_{jk}$ in (5), such as kernel convolutions, are attractive in many settings, they may overfit the sparse observed data in this application. Hence, we compare a regionally constant model with two regions for $\boldsymbol{\gamma}_{jk}$ to a model using Gaussian kernel basis functions. For both models, the basis function matrix in equation (5) is of the form

$$\boldsymbol{B}_{jk} = (\mathbf{1}_n, \boldsymbol{W}), \tag{11}$$

where $\mathbf{1}_n$ denotes a length n vector of ones, and $\boldsymbol{W} = \{w_{i\ell}\}$ is an $n \times L$ matrix of "weights." In the case of the Gaussian kernel we set

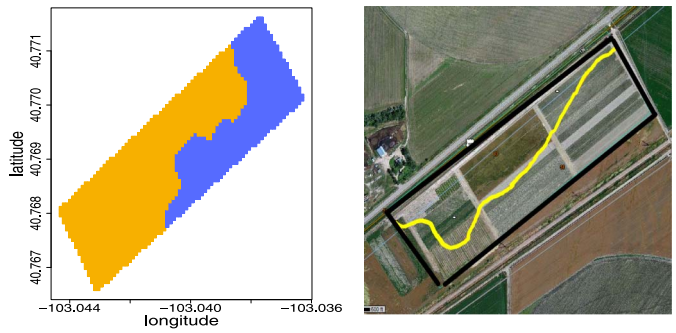
$$w_{i\ell} = \frac{1}{\sqrt{2\pi\lambda^2}} \exp\left\{-\frac{\|\boldsymbol{s}_i - \boldsymbol{s}_\ell^\star\|^2}{2\lambda^2}\right\},$$

where we have L "knots" $\boldsymbol{s}_1^\star, \ldots, \boldsymbol{s}_L^\star$ and λ^2 denotes the (unknown) variance of the kernel (for this application we consider $L = 5$). Alternatively, in the regionally constant model with two regions (which are defined below), $L = 1$ and

$$w_{i1} = \mathbb{1}_{\mathcal{R}_1}(\boldsymbol{s}), \tag{12}$$

where $\mathcal{R}_1$ denotes the set of locations corresponding to region 1.

To define the regions in the regionally constant model, we cluster the observations using gradient clustering [Heaton, Christensen and Terres (2017)]. The left panel of Figure 3 includes the regionalization resulting from gradient clustering, and the right panel delineates the field into regions according to the data in the



(a) Gradient Clustering (b) USDA Web Soil Map

FIG. 3. *Field regions according to* (a) *Gradient Clustering and* (b) *the Web Soil Survey. Due to the similarity, we used the gradient clustering regionalization in* (a) *to define the regionally constant model.*

USDA Soil Survey (i.e., the two regions have fundamentally different soils according to the USDA). The USDA soil survey primarily uses remote sensing data [Natural Resources Conservation Service (2016)], while our partitioning is based on the observed data; hence we opt to use our partitioning but draw confidence in our spatial partition from that fact that the two partitions are similar. For each of the three proposed models (regionally constant with 2 regions, Gaussian kernel convolutions and spatially constant), we consider both Markov [as in (3)] and non-Markov conditioning.

To further compare our modeling choice to simpler alternatives, we also considered four additional model specifications. First, to assess the need for spatially varying cross-correlations, we consider a spatially constant model with $\boldsymbol{B}_{jk} = \mathbf{1}_n$ in both a Markov and non-Markov conditioning framework [see equation (3)]. Second, we considered a "3-D" model where correlation is determined by distance in space and depth using the correlation function $\text{Corr}(d_1, d_2) = \exp\{-\alpha_1 d_1 - \alpha_2 d_2\}$, where d_1 is the distance in space and d_2 is the distance in depth such that α_i governs the decay across dimension i. Finally, we considered a nonstationary version of this 3-D model, where α_1 and α_2 were allowed to be specific to the same regions as the regionally constant model.

Table 2 compares the six different models in terms of leave-one-out cross-validated bias, root mean square error (RMSE), coverage, coverage of the total WHC, predictive interval width of total WHC, continuous rank probability score [CRPS; Gneiting and Raftery (2007)] and DIC. The cross-validated bias is defined as

$$\text{Bias} = \frac{1}{5 \times 17} \sum_{i=1}^{17} \sum_{d=1}^{5} \big(y_d(\boldsymbol{s}_i) - \hat{y}_{d,-i}(\boldsymbol{s}_i)\big),$$

TABLE 2

Leave-one-out model comparison results in terms of bias, root mean square error (RMSE), predictive interval coverage (CVG), total coverage (TCVG), predictive interval width (PIW), continuous rank probability score (CRPS) and deviance information criterion (DIC). "R," "SC" and "G" denote the regional, spatially constant and Gaussian kernel models, while "M" denotes the use of the Markov assumption in equation (3). The convolution models are slightly better at prediction, but have wider predictive intervals. Additionally, the regionally constant models fit the data better (in terms of DIC)

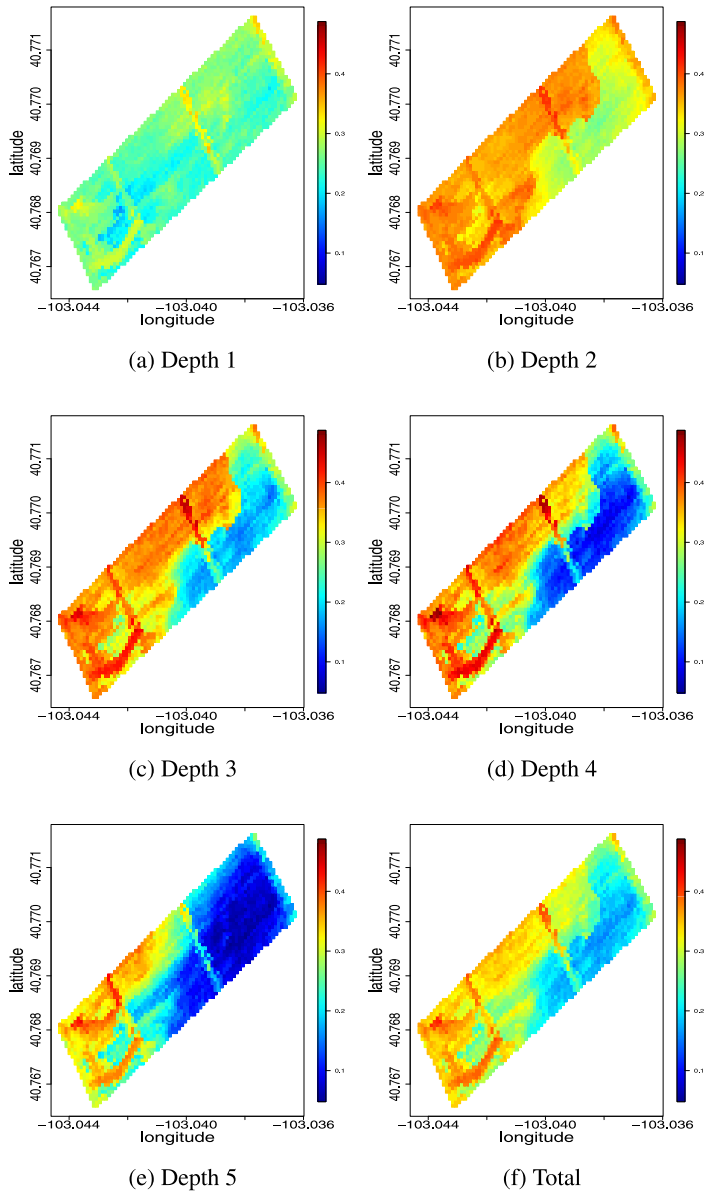
	Bias	**RMSE**	**CVG**	**TCVG**	**PIW**	**CRPS**	**DIC**
R	0.038	0.110	0.941	0.882	0.248	4.931	422.130
R-M	0.038	0.113	0.906	0.941	0.238	4.959	444.405
SC	0.042	0.109	0.953	0.882	0.235	4.799	493.499
SC-M	0.043	0.111	0.894	0.824	0.238	4.900	509.154
G	0.031	0.090	0.988	0.941	0.260	4.186	512.303
G-M	0.031	0.088	0.976	0.941	0.260	4.046	496.531
3D	0.064	0.227	0.884	0.832	0.473	5.750	555.305
3D-R	0.051	0.154	0.903	0.882	0.315	5.284	523.454

where $\hat{y}_{d,-i}(\boldsymbol{s}_i)$ is the posterior predictive mean of $y_d(\boldsymbol{s}_i)$ with the ith observation omitted. Likewise, RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{5 \times 17} \sum_{i=1}^{17} \sum_{d=1}^{5} \big(y_d(\boldsymbol{s}_i) - \hat{y}_{d,-i}(\boldsymbol{s}_i)\big)^2}.$$

Coverage is defined as the percent of all predictive intervals that include the left-out value, and predictive interval width is the length between the interval endpoints.

The model comparison results in Table 2 are mixed. At first glance, the convolution models appear to be best for predicting, with the lowest RMSE, lowest CRPS and best coverage. However, the predictive interval widths are considerably higher, suggesting that these convolution models are highly variable due to the sparse nature of the data. In contrast, the regionally constant models still attain adequate coverage (the 0.882 coverage value is within binomial sampling variability due to the fact that only 17 values were able to be left out) while fitting the data substantially better (smallest DIC) and obtaining smaller predictive interval widths. The spatially constant models are comparable to the others in terms of most criteria, but have considerably higher DIC values than do the regionally constant models. Likewise, the 3-D alternatives are not preferred, which we attribute to the fact that correlation across depth is governed more by soil type than distance. For these reasons, we use the regionally constant, non-Markov model as our final model in this application. Additionally, to substantiate our decision to condition from deepest to shallowest as opposed to shallowest to deepest, we compared the DICs from the regionally constant, non-Markov models for each approach. The respective DICs were 422.130 and 452.96, supporting the deepest to shallowest conditioning.



(a) Depth 1 (b) Depth 2

(c) Depth 3 (d) Depth 4

(e) Depth 5 (f) Total

FIG. 4. *The location-specific posterior means of the posterior predictive distribution for WHC (in* m^3/m^3*) across the spatial field.*

Predictions resulting from this model are included in Figure 4. Note the consistency between the plots in Figure 4 and those in Figure 1, suggesting adequate use of the data to achieve spatial kriging. Also, the predictions appear consistent with certain agricultural principles, such as the fact that there is less spatial variation

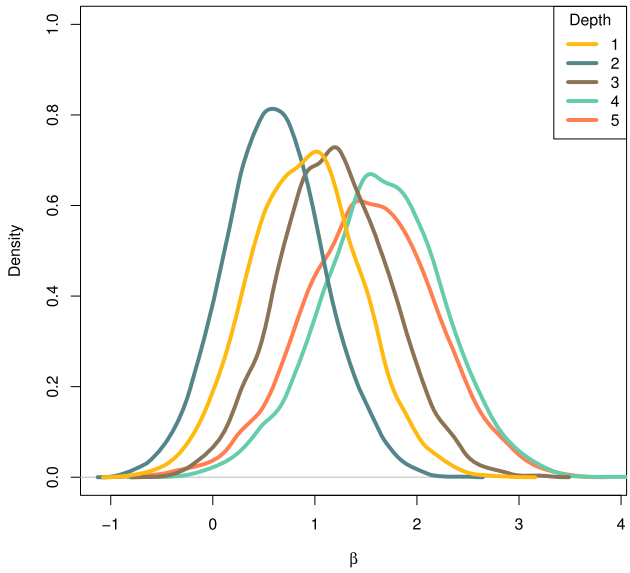


FIG. 5. *Posterior densities of* $\beta_{1,j}$ *for depths* $j = 1, \ldots, 5$. *The posterior probabilities of* $\beta_{1,j} > 0$ *for* $j = 1, \ldots, 5$ *are, respectively,* 0.956, 0.895, 0.989, 0.996 *and* 0.989.

in WHC within the topsoil layer, but as depth increases and the topsoil layer ends, there is much more spatial variation due to varying depth in soil horizons or layers.

An apparent feature from Figure 4 is that the EC measurements tend to be highly correlated with WHC at lower depths, as various spatial features in EC are also present in the predicted values of WHC. This result is further displayed by Figure 5, which reports the posterior kernel density estimates of $\beta_{1,j}$, the coefficient for the relationship between WHC and log(EC) at depth j. From Figure 5, notice the general strengthening relationship as j increases. Specifically, the posteriors suggest the strongest positive relationship between log(EC) and WHC to be at depth 4, while at depth 2 there appears to be no significant relationship. This may be the result of the EC measurements being a better reflection of EC at greater depths than at shallow depths.

Table 3 presents the estimated correlations between depths in both regions, and illustrates the difference in inter-depth cross-correlations between the two regions as hypothesized. Notably, because the regions defined by the gradient clustering closely match changes in soil composition as seen in Figure 3, the difference in correlations is likely dependent upon the soil composition.

4.2. *Selection of additional sampling locations*. Figure 6 shows the predictive interval widths (a measure of predictive uncertainty) for the WHC predictions in Figure 4. From Figure 6, low depths seem to have greater uncertainty than shallow depths (particularly for the far southwest region of the field). Additionally, this

TABLE 3
Estimated colocated correlations between depths for each of the two regions along with 95% credible intervals

	Correlation Between Depths					
	Southwest Region			Northeast Region		
Depths	**Lower**	**Center**	**Upper**	**Lower**	**Center**	**Upper**
(1, 2)	0.63	0.80	0.90	−0.05	0.28	0.56
(1, 3)	0.53	0.73	0.87	−0.51	−0.18	0.20
(1, 4)	−0.18	0.19	0.54	−0.55	−0.16	0.23
(1, 5)	−0.54	−0.08	0.35	−0.71	−0.39	0.10
(2, 3)	0.28	0.58	0.79	0.31	0.60	0.80
(2, 4)	−0.41	−0.02	0.36	0.44	0.69	0.86
(2, 5)	−0.47	−0.07	0.37	0.06	0.46	0.75
(3, 4)	0.24	0.58	0.79	0.80	0.90	0.96
(3, 5)	−0.60	−0.22	0.26	0.57	0.79	0.92
(4, 5)	−0.46	−0.01	0.46	0.54	0.79	0.92

southwest region of the field has high variability in EC, and further data may be desired to better estimate the relationship between EC and WHC in this region.

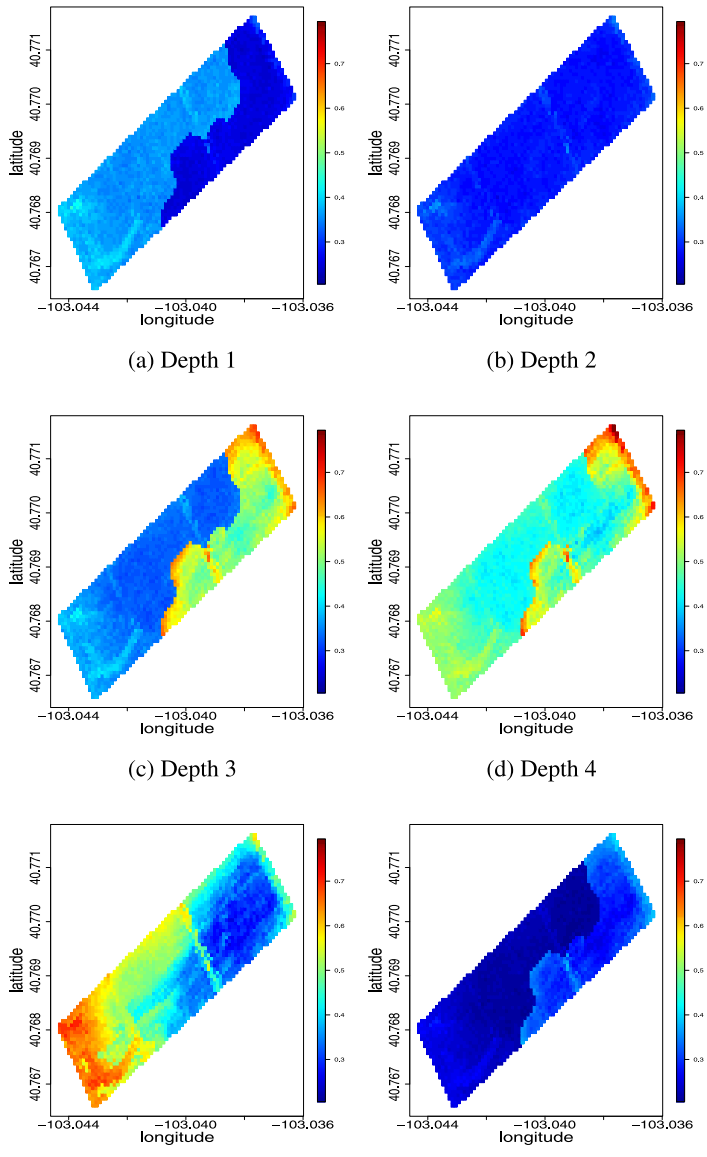
Using the IMSE method outlined in Section 3, we approximate IMSE for each of 100 candidate "new" locations, considering two different $\boldsymbol{A}$ matrices. One simple choice in $\boldsymbol{A}$ is the matrix $\boldsymbol{A}_1 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \end{pmatrix}$ (the sum across depths), accounting for the uncertainty in all depths. However, because most of the uncertainty occurs in depths 3, 4 and 5, a second choice of $\boldsymbol{A}$ is the matrix

$$\boldsymbol{A}_2 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{13}$$

The IMSE results are summarized in Figure 7, which displays the IMSE for all 100 candidate locations (the location with the lowest IMSE indicated by a star) with the left panel using $\boldsymbol{A}_1$ and the right using $\boldsymbol{A}_2$. The plots align with what might be expected based on where data is limited; that is, the locations with especially sparse data are those with the lowest IMSE.

5. Validation of statistical methods. To demonstrate that the methods introduced in this paper are indeed valid, we present the results of two simulation studies. In the first, we validate the model results and model comparison methods, and in the second, we validate that the design criterion outlined in Section 3 selects an appropriate location.

5.1. *Validation of statistical model.* The main purposes of this simulation study are to, first, ensure that model parameters are learned from the data rather



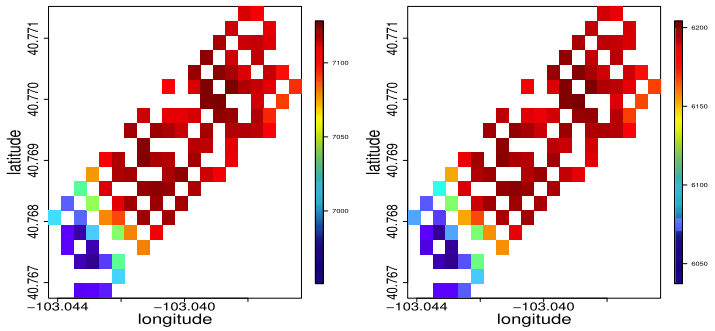
(a) Depth 1 (b) Depth 2

(c) Depth 3 (d) Depth 4

(e) Depth 5 (f) Total

FIG. 6. *WHC* (m^3/m^3) 95% *predictive interval widths.*

than overly influenced by the prior specification, and, second, validate the use of the model selection criteria summarized in Table 2. To do this, we generated 50 full fields of WHC data (at all 2291 prediction locations) using the posterior mean of the regionally constant, non-Markov analysis described in Section 4 (which was our selected model). For each of the 50 simulated fields, we fit and compared each



(a) IMSE plot using $\boldsymbol{A}_1$ (b) IMSE plot using $\boldsymbol{A}_2$

FIG. 7. *IMSE for all* 100 *candidate locations*.

of the models described in Section 4 except the 3-D models (which were excluded because correlation in depth is more related to soil layering). Specifically, we considered three different specifications of basis functions $\boldsymbol{B}_{jk}$ (regionally constant bases, Gaussian kernel bases and nonspatially varying bases) and two different dependency structures (Markov and non-Markov) to the simulated WHC at the same 31 locations as observed in the true data.

Table 4 compares the six models according to the same criteria as in Table 2. From Table 4, the models are comparable in terms of predictive power with a slight preference to the R-M model in terms of CRPS. Interestingly, the greatest discrepancy between the models comes in terms of DIC, where the regionally constant non-Markov model is clearly preferred. This result, therefore, reassures our decision to utilize DIC as an accurate model selection criterion.

TABLE 4

Simulation model comparison results in terms of bias, root mean square error (*RMSE*), *predictive interval coverage* (*CVG*), *total coverage* (*TCVG*), *predictive interval width* (*PIW*), *continuous rank probability score* (*CRPS*) *and deviance information criterion* (*DIC*). *"R," "SC" and "G" denote the regional, spatially constant and Gaussian kernel models, while "M" denotes the use of the Markov assumption in equation* (3). *Models are comparable in terms of all metrics except DIC, in which the regionally constant non-Markov clearly outperforms the others*

	Bias	**RMSE**	**CVG**	**TCVG**	**PIW**	**CRPS**	**DIC**
R	0.002	0.106	0.954	0.944	0.267	927.982	407.913
R-M	0.001	0.106	0.946	0.948	0.268	923.155	457.004
SC	−0.001	0.107	0.941	0.930	0.265	941.044	481.811
SC-M	−0.001	0.106	0.937	0.939	0.269	930.125	491.071
G	−0.003	0.106	0.949	0.935	0.268	931.457	465.190
G-M	−0.003	0.106	0.944	0.942	0.271	927.250	482.811

Having successfully validated our model selection process, we examine the estimated coverages for $\boldsymbol{\beta}_j$ and γ_{jk} for all depths $j = 1, \ldots, 5$ and $k > j$. Again, letting the coverage equal the proportion of all 95% credible intervals containing the true parameter value, all estimated coverages fall between 0.92 and 1.0, offering evidence of sufficient accuracy of parameter estimates. Notably, each of the marginal posterior distributions (not shown for the sake of brevity) were markedly more peaked than the corresponding prior distribution, suggesting adequate statistical learning.

5.2. *Validation of sequential sampling design*. To validate our proposed sequential sampling design using IMSE, we simulated 50 fields of WHC values at all 2291 locations using the posterior mean of the regionally constant, non-Markov analysis from Section 4 and specified the realizations at the 31 locations in the application above as the observed data. For this validation exercise, we compare the predictive performance of our model after adding a 32nd observation where this additional observation was chosen using one of three criteria: (i) the proposed IMSE criterion with $\boldsymbol{A} = \boldsymbol{I}$, (ii) the location with greatest posterior predictive uncertainty and (iii) the location that maximizes a geometric space-filling criteria over the spatial design given the original 31 sampling locations.

The average RMSE when selecting an additional point via IMSE was 7.01 compared to 6.95 and 6.93 when using the greatest predictive uncertainty and space-filling criteria, respectively, indicating nearly identical point prediction performance. However, we are interested in not only point prediction but also the uncertainty associated with these predictions. As such, define the cumulative generalized predictive variance as

$$\text{CGV} = \sum_{d=1}^{5} |\boldsymbol{\Sigma}_{d,\text{U}|\text{O}}| \tag{14}$$

as a measure of predictive uncertainty where $\boldsymbol{\Sigma}_{d,\text{U}|\text{O}}$ is the conditional variance of the unobserved locations given the observed locations at the dth depth and $|\cdot|$ is the determinant. The CGV was 4540 for the IMSE criterion compared to 5471 and 5523 when using greatest predictive uncertainty and space-filling criteria, respectively. This is an approximately 17% decrease in the cumulative predictive variance when adding an observation using the IMSE criterion compared to the other methods [Ranjan et al. (2011) saw similar decreases], suggesting that the IMSE criterion successfully chooses locations that not only aid in point prediction but also decrease predictive uncertainty.

6. Conclusions. This article had two main purposes: (i) use EC to accurately predict WHC at various depths across an agricultural field while properly accounting for predictive uncertainty, and (ii) locate points which, if added to the observed data, would minimize the predictive uncertainty. To accomplish (i), we used a conditional specification of a Gaussian process model. However, in order to account

for potential spatially varying cross-correlation between depths, we proposed the use of spatially varying coefficients in the conditional model. Though we developed a general method for characterizing the space-varying coefficients using basis functions, model comparison revealed that a regionally constant model for the WHC application was preferred. We accomplished (ii) using a multivariate extension of IMSE, choosing as the "next" location the point with the lowest estimated IMSE. Comparison of IMSE for various predictive quantities of interest (total vs. depths 3–5) showed strikingly similar choices of the next location.

Our choice to treat WHC at all depths as a multivariate outcome (rather than modeling correlation based on depth), while scientifically and empirically justified above, precludes us from extrapolating WHC deeper than the deepest measured depth (in this case, 1.5 meters). Note, however, that extrapolation to deeper depths is not of direct interest in this application because root depth for most crops typically does not exceed this depth [see Natural Resources Conservation Service (1997), Chapter 3]. Certainly, there could be applications where such an extrapolation is needed, in which case the modeling strategy would likely need to depend on distance in depth and soil type. We leave the use of such correlation functions for future work.

For the sampling design problem above, we only considered adding one additional sampling location at a time. This decision was made primarily based on personal conversations with farmers who indicated that most farmers are reluctant to allow WHC sampling on their field due to the high cost of data collection. Given this reluctancy, adding one additional sampling location is a constraint under which this research has to operate. However, in many cases, adding more than one sampling location may be possible and desired. In such cases, following Ranjan et al. (2011), the methods described in Section 3 could be extended so as to consider multiple additional sampling locations.

Quantifying the spatial variability of WHC is only one metric necessary to generate prescription maps for a variable rate irrigation system. Other metrics include water inputs, such as precipitation and irrigation, and water losses, including evapotranspiration, runoff and drainage. Each of these variables are spatially and temporally dynamic. While this represents a necessary first step, additional work will be needed to model the spatial and temporal variation of the full water balance equation, leading to prescription maps for precision irrigation [Longchamps et al. (2015)].

In terms of contributions to agricultural science, this research offers farmers and agricultural scientists insight into the WHC of a field using sparse observations. Such information allows them to efficiently utilize scarce water resources. The multivariate predictions also offer them a better understanding of how soil varies at lower depths, as this is not readily observable.

Acknowledgments. The authors would like to thank Dr. C. Shane Reese for providing helpful discussion and insight into the multivariate IMSE criterion. Any

opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

ALLEN, R. G., PEREIRA, L. S., RAES, D., SMITH, M. et al. (1998). Crop evapotranspiration—guidelines for computing crop water requirements—FAO irrigation and drainage paper 56. *FAO, Rome* **300** D05109.

APANASOVICH, T. and GENTON, M. G. (2010). Cross-covariance functions for multivariate random fields based on latent dimensions. *Biometrika* **97** 15–30. MR2594414

APANASOVICH, T. V., GENTON, M. G. and SUN, Y. (2012). A valid Matérn class of cross-covariance functions for multivariate random fields with any number of components. *J. Amer. Statist. Assoc.* **107** 180–193. MR2949350

BANERJEE, S., GELFAND, A. E., FINLEY, A. O. and SANG, H. (2008). Gaussian predictive process models for large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 825–848. MR2523906

BRUCE, R. R. and LUXMOORE, R. J. (1986). Water retention: Field methods. In *Methods of Soil Analysis*: *Part* 1—*Physical and Mineralogical Methods*, 2nd ed. (A. Klute, ed.). 663–686. Soil Science Society of America, American Society of Agronomy, Madison, WI.

CRESSIE, N. and JOHANNESSON, G. (2008). Fixed rank kriging for very large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 209–226. MR2412639

CRESSIE, N. and ZAMMIT-MANGION, A. (2015). Multivariate spatial covariance models: A conditional approach. Available at arXiv:1504.01865v1.

CURRIN, C., MITCHELL, T., MORRIS, M. and YLVISAKER, D. (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *J. Amer. Statist. Assoc.* **86** 953–963. MR1146343

DIGGLE, P. J. and RIBEIRO, P. J. (2002). Bayesian inference in Gaussian model-based geostatistics. *Geogr. Environ. Model.* **6** 129–146.

FINLEY, A. O., SANG, H., BANERJEE, S. and GELFAND, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Comput. Statist. Data Anal.* **53** 2873–2884. MR2667597

FUENTES, M. and REICH, B. (2013). Multivariate spatial nonparametric modelling via kernel processes mixing. *Statist. Sinica* **23** 75–97. MR3076159

GELFAND, A. E. and BANERJEE, S. (2010). Multivariate spatial process models. In *Handbook of Spatial Statistics* 495–515. CRC Press, Boca Raton, FL. MR2730963

GELFAND, A. E., KIM, H.-J., SIRMANS, C. F. and BANERJEE, S. (2003). Spatial modeling with spatially varying coefficient processes. *J. Amer. Statist. Assoc.* **98** 387–396. MR1995715

GELFAND, A. E., SCHMIDT, A. M., BANERJEE, S. and SIRMANS, C. F. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *TEST* **13** 263–312. With discussion by Montserrat Fuentes, Dave Higdon and Bruno Sansó and a rejoinder by the authors. MR2154003

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GENTON, M. G. and KLEIBER, W. (2015). Cross-covariance functions for multivariate geostatistics. *Statist. Sci.* **30** 147–163. MR3353096

GNEITING, T., KLEIBER, W. and SCHLATHER, M. (2010). Matérn cross-covariance functions for multivariate random fields. *J. Amer. Statist. Assoc.* **105** 1167–1177. MR2752612

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

GUHANIYOGI, R., FINLEY, A. O., BANERJEE, S. and KOBE, R. K. (2013). Modeling complex spatial dependencies: Low-rank spatially varying cross-covariances with application to soil nutrient data. *J. Agric. Biol. Environ. Stat.* **18** 274–298. MR3110894

HEATON, M. J., CHRISTENSEN, W. F. and TERRES, M. A. (2017). Nonstationary Gaussian process models using spatial hierarchical clustering from finite differences. *Technometrics* **59** 93–101. MR3604192

HIGDON, D. (2002). Space and space–time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues* (C. Anderson, V. Barnett, P. C. Chatwin and A. H. El-Shaarawi, eds.) 37–56. Springer-Verlag, London.

JOHNSON, M. E., MOORE, L. M. and YLVISAKER, D. (1990). Minimax and maximin distance designs. *J. Statist. Plann. Inference* **26** 131–148. MR1079258

JONES, G. L., HARAN, M., CAFFO, B. S. and NEATH, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **101** 1537–1547. MR2279478

KANG, E. L. and CRESSIE, N. (2011). Bayesian inference for the spatial random effects model. *J. Amer. Statist. Assoc.* **106** 972–983. MR2894757

KITCHEN, N. R., DRUMMOND, S. T., LUND, E. D., SUDDUTH, K. A. and BUCHLEITER, G. W. (2003). Soil electrical conductivity and topography related to yield for three contrasting soil–crop systems. *Agron. J.* **95** 483–495.

KLEIBER, W. and GENTON, M. G. (2013). Spatially varying cross-correlation coefficients in the presence of nugget effects. *Biometrika* **100** 213–220. MR3034334

KLEIBER, W., SAIN, S. R., HEATON, M. J., WILTBERGER, M., REESE, C. S. and BINGHAM, D. (2013). Parameter tuning for a multi-fidelity dynamical model of the magnetosphere. *Ann. Appl. Stat.* **7** 1286–1310. MR3127948

KLUTE, A. (1986). Water retention: Laboratory methods. In *Methods of Soil Analysis*: *Part* 1—*Physical and Mineralogical Methods*, 2nd ed. (A. Klute, ed.) 635–662. Soil Science Society of America, American Society of Agronomy, Madison, WI.

KRAUSE, A., SINGH, A. and GUESTRIN, C. (2008). Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *J. Mach. Learn. Res.* **9** 235–284.

LEMOS, R. T. and SANSÓ, B. (2009). A spatio-temporal model for mean, anomaly, and trend fields of North Atlantic sea surface temperature. *J. Amer. Statist. Assoc.* **104** 5–18. MR2662306

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. MR1925014

LONGCHAMPS, L., KHOSLA, R., REICH, R. and GUI, D. W. (2015). Spatial and temporal variability of soil water content in leveled fields. *Soil Sci. Soc. Amer. J.* **79** 1446–1454.

MAJUMDAR, A., PAUL, D. and BAUTISTA, D. (2010). A generalized convolution model for multivariate nonstationary spatial processes. *Statist. Sinica* **20** 675–695. MR2682636

MZUKU, M., KHOSLA, R., REICH, R., INMAN, D., SMITH, F. and MACDONALD, L. (2005). Spatial variability of measured soil properties across site-specific management zones. *Soil Sci. Soc. Amer. J.* **69** 1572–1579.

NATURAL RESOURCES CONSERVATION SERVICE (1997). National Engineering Handbook: Irrigation guide. U.S. Department of Agriculture.

NATURAL RESOURCES CONSERVATION SERVICE (2016). National Soil Survey Handbook. U.S. Department of Agriculture.

NYCHKA, D., BANDYOPADHYAY, S., HAMMERLING, D., LINDGREN, F. and SAIN, S. (2015). A multiresolution Gaussian process model for the analysis of large spatial datasets. *J. Comput. Graph. Statist.* **24** 579–599. MR3357396

PLASTER, E. (2013). *Soil Science and Management*. Cengage Learning, Independence, KY.

RANJAN, P., LU, W., BINGHAM, D., REESE, S., WILLIAMS, B. J., CHOU, C.-C., DOSS, F., GROSSKOPF, M. and HOLLOWAY, J. P. (2011). Follow-up experimental designs for computer models and physical processes. *J. Stat. Theory Pract.* **5** 119–136. MR2829827

ROYLE, J. A. and BERLINER, L. M. (1999). A hierarchical approach to multivariate spatial modeling and prediction. *J. Agric. Biol. Environ. Stat.* **4** 29–56. MR1812239

SACKS, J., SCHILLER, S. B. and WELCH, W. J. (1989). Designs for computer experiments. *Technometrics* **31** 41–47. MR0997669

SADLER, E. J., EVANS, R., STONE, K. C. and CAMP, C. R. (2005). Opportunities for conservation with precision irrigation. *J. Soil Water Conserv.* **60** 371–378.

SANG, H., JUN, M. and HUANG, J. Z. (2011). Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors. *Ann. Appl. Stat.* **5** 2519–2548. MR2907125

SANTNER, T. J., WILLIAMS, B. J. and NOTZ, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer, New York. MR2160708

SEAGER, R., HOERLING, M., SCHUBERT, S., WANG, H., LYON, B., KUMAR, A., NAKAMURA, J. and HENDERSON, N. (2014). Causes and predictability of the 2011–14 California drought. Assessment report, National Oceanic and Atmospheric Administration, Silver Spring, MD.

ZHANG, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Amer. Statist. Assoc.* **99** 250–261. MR2054303

DEPARTMENT OF STATISTICS
210 TMCB
BRIGHAM YOUNG UNIVERSITY
PROVO, UTAH 84602
USA
E-MAIL: messick.rachel@gmail.com
mheaton@stat.byu.edu
neil_hansen@byu.edu