Abstract

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 irrigation according to within field variation in soil and topographic conditions. As a precursor, methods to characterize spatial variation of soil hydraulic properties is needed. One such property is soil water holding capacity (WHC). This analysis develops a Bayesian 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 a novel conditional specification of a multivariate Gaussian process with 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.

Agriculture companies have already developed precision irrigation technology, which is the ability to apply different amounts of water over different areas of a field. However, the current state of agricultural science lacks practical means to characterize spatial variation of irrigation needs across a field. One key metric for assessing irrigation needs is the
water holding capacity (WHC) of the soil. Measuring WHC, however, is an expensive and time-consuming process. WHC 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 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 .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 leveraging 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 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
ber of prediction locations approaches 12,000, making prediction computationally infeasible. Furthermore, even if computation were reasonable, by the nature of soil layering, WHC violates the assumption that correlation in depth is determined by distance (e.g., correlation between depths 1 and 2 is different than correlation between depths 4 and 5). Thus, instead of using a univariate spatial model in three dimensions, we employ a multivariate spatial model over a two-dimensional domain, considering each of 5 depths at every spatial location. 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 [Gneiting et al., 2010, Royle and Berliner, 1999, Apanasovich et al., 2012, Apanasovich and Genton, 2010, Gelfand et al., 2004, Genton and Kleiber, 2015, Gelfand and Banerjee, 2010], methods that account for spatially-varying correlations are less so due to the difficulty of ensuring a positive definite 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 et al. [2010] use kernel convolutions to build non-stationary cross-covariances. Sang et al. [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 [Royle and Berliner, 1999, Cressie and Zammit-Mangion, 2015]. 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 small 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, the most realistic and cost effective scenario for farmers is to add only a single observation location at a time rather than several observations at a time. Hence, 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). While space-filling and Latin hypercube designs [Johnson et al., 1990, Sacks et al., 1989] are useful for selecting initial sampling locations, they do not incorporate knowledge of the surface already learned from observations. Likewise, follow-up designs 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 process 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 \( y_1(s), \ldots, y_5(s) \) represent WHC at spatial location \( s \in S \) measured at depths 1 through 5, respectively. Using a conditional specification, we represent the likelihood as

\[
[y_{1:5}(s)] = [y_1(s) | y_{2:5}(s)] [y_2(s) | y_{3:5}(s)]
* [y_3(s) | y_{4:5}(s)] [y_4(s) | y_{5:5}(s)] [y_5(s)] \tag{1}
\]

where \([\cdot]\) denotes an arbitrary distribution and \( y_{i:j}(s) = (y_i(s), \ldots, y_j(s))' \). We assume a Gaussian process model for each depth and include the 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 \( y_1(s), \ldots, y_5(s) \) represent WHC at spatial location \( s \in S \) measured at depths 1 through 5, respectively. Using a conditional specification, we represent the likelihood as

\[
[y_{1:5}(s)] = [y_1(s) | y_{2:5}(s)] [y_2(s) | y_{3:5}(s)]
* [y_3(s) | y_{4:5}(s)] [y_4(s) | y_{5:5}(s)] [y_5(s)] \tag{1}
\]

where \([\cdot]\) denotes an arbitrary distribution and \( y_{i:j}(s) = (y_i(s), \ldots, y_j(s))' \). We assume a Gaussian process model for each depth and include the 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.
the measurement error variance in this application (if included, the measurement error would be confounded with $\epsilon_j(s)$), we utilize $\epsilon_j(s)$ 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 our choice in this regard hinges on prior knowledge from agricultural scientists who, traditionally, consider soil from the deepest layer to the most shallow. This decision is validated via DIC comparison, reported in Section 4. It should also be noted that this model assumes only point-wise 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 point-wise 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

$$y_j(s) \mid \{y_{j+1}(s)\} = x'_j(s)\beta_j + \gamma_{j(j+1)}(s)(y_{j+1}(s) - x'_{j+1}(s)\beta_{j+1}) + w_j(s) + \epsilon_j(s).$$

(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 $\gamma_{jk}(s)$. To illustrate, consider $\text{Cov}(y_4(s_1), y_5(s_2))$ (similar derivations exist for any two locations and any two depths). From (2),

$$y_4(s) = x'_4(s)\beta_4 + \gamma_{45}(s)(y_5(s) - x'_5(s)\beta_5) + w_4(s) + \epsilon_4(s),$$

and,

$$y_5(s) = x'_5(s)\beta_5 + w_5(s) + \epsilon_5(s).$$

Standard algebraic manipulations yield,

$$\text{Cov}(y_4(s_1), y_5(s_2)) = \mathbb{E}(y_4(s_1)y_5(s_2)) - \mathbb{E}(y_4(s_1))\mathbb{E}(y_5(s_2))$$

$$= \gamma_{45}(s_1)[\mathbb{E}(y_5(s_1)y_5(s_2))$$

$$- \mathbb{E}(y_5(s_1))\mathbb{E}(y_5(s_2))]$$

$$= \gamma_{45}(s_1)\text{Cov}(y_5(s_1), y_5(s_2))$$

$$= \gamma_{45}(s_1)\sigma^2_{y_5}M_5(\|s_1 - s_2\| \mid \phi_5)$$

$$+ \tau^2_5\mathbb{I}_{\{s_1 = s_2\}}$$

where $\mathbb{I}_A$ is an indicator for the set $A$. Hence, under the conditional specification in (2), the correlation between any two locations $(s_1, s_2)$ at any two corresponding depths $(j, k)$ (for $j < k$) is completely determined by $\gamma_{jk}(s_1)$ which, subsequently, leads to spatially varying cross-correlations.

Let $y_j = (y_j(s_1), \ldots, y_j(s_n))'$ be the vector of observations measured at depth $j$. The process model specification in (2) implies a joint distribution of,

$$y_j \mid \{y_k : k > j\} \sim \mathcal{N}\left(X_j\beta_j + \sum_{k>j} D_k\gamma_{jk}, \sigma^2_{y_j}M_j + \tau^2_j\mathbb{I}_n\right)$$

(4)

where $X_j$ is the $n \times P$ design matrix with $i^{\text{th}}$ row $x'_{j}(s_i)$, $D_k = \text{diag}(y_k - X_k\beta_k)$ is the $n \times n$ diagonal matrix whose diagonal elements are formed from the error vector $(y_k - X_k\beta_k)$, and $\gamma_{jk} = (\gamma_{jk}(s_1), \ldots, \gamma_{jk}(s_n))'$ is the length $n$ vector of $\gamma$. 5
coefficients, $M_j$ is the $n \times n$ matrix of correlations at depth $j$ with $i\ell$th element $M_{i\ell,j}(\|s_i - s_\ell\| \mid \phi_j)$, and $I_n$ 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 $\gamma$ coefficients. That is, let $\gamma_{jk}(s) = \gamma_{jk}^\ast + \sum_{\ell=1}^{L} b_{jk,\ell}(s) \gamma_{jk,\ell}^\ast = b'_{jk}(s) \gamma_{jk}^\ast$ where $\gamma_{jk}^\ast$ is an overall mean and $b_{jk,\ell}(\cdot)$ is a basis function with associated coefficient $\gamma_{jk,\ell}^\ast$. 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 Sanso, 2009, Nychka et al., 2015] or kernel convolutions [Higdon, 2002]. Using basis function expansions, $\gamma_{jk}$ is represented as

$$\gamma_{jk} = B_{jk} \gamma_{jk}^\ast$$

where $B_{jk}$ is a $n \times (L + 1)$ matrix of known basis functions with $i$th row $b'_{jk}(s_i)$. Substituting (5) into (4) results in,

$$y_j | \{y_k : k > j\} \sim N(\mathbf{X}^\ast_j \mathbf{\theta}_j, \sigma_j^2 M_j + \tau_j^2 I_n)$$

where $\mathbf{X}^\ast_j = [\mathbf{X}_j; D_{j+1} B_{j(j+1)}; \cdots; D_n B_{j\delta}]$ and $\mathbf{\theta}_j = (\beta'_j, \gamma_{j(j+1)}^\ast, \ldots, \gamma_{j\delta}^\ast)^T$. 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,

$$y_j | \{y_k : k > j\} \sim N(\mathbf{X}^\ast_j \mathbf{\theta}_j, \kappa_j^2 (\omega_j M_j + (1 - \omega_j) I_n))$$

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 $\mathbf{\theta}_j$ and $\kappa_j^2$, 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 $\omega_j$ to a fine grid over $[0, 1]$ enables direct sampling of $\omega_j$ from its complete conditional distribution with minimal loss of information. Of course, a beta prior for $\omega_j$ 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 $M_j$ is governed by unknown parameters $\phi_j$ (controlling the spatial decay) and $\nu_j$ (controlling the spatial smoothness). Following results from Zhang [2004], we recommend, without loss of predictive power, to simply fix each $\nu_j$ and estimate $\phi_j$. Traditionally, gamma, inverse-gamma, or log-normal priors are used for $\phi_j$ but we opt for the computationally simpler discrete uniform prior. Specifically, we construct a discrete uniform prior for $\phi_j$ 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 $\phi_j$ values resulting in reasonable values for $\phi_j$ that are given equal prior weight [Diggle and Ribeiro, 2002].
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. Ranjan et al. [2011] propose the integrated mean square error (IMSE) criterion for selecting additional sampling locations. However, the IMSE criterion, as originally proposed, is for a univariate representation of all parameters in the model described in Section 2. The data presented in Section 1 include \( \{s_i\} : i = 1, \ldots, n \) represent the set of WHC observations and \( \psi \) be the vector of all parameters in the model described in Section 2 with parameter space \( \Psi \). Furthermore, let \( A \) represent a \( Q \times 5 \) user-specified matrix such that \( Ay(s) \) represents predictive quantities of interest. For example, if \( Q = 1 \) and \( A = (1, 0, 0, 0, 0) \) then \( Ay(s) = y_1(s) \) suggesting that WHC at depth 1 is of predictive interest. Naturally, if \( A = I \) then prediction at all 5 depths is of interest. If a new observation \( y(s_0) \) is gathered at \( s_0 \), the mean square error (MSE) at location \( s \) is given by

\[
\text{MSE}(s \mid y(s_0), \psi) = \mathbb{E} \left[ \| Ay(s) - Ay(s) \|^2 \mid y(s_0), \mathcal{Y}, \psi \right]
\]  

(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 \( \hat{y}(s) = \mathbb{E}(y(s) \mid y(s_0), \mathcal{Y}, \psi) \) is the predicted value of \( y(s) \) given the data \( \mathcal{Y} \), the “new” observation \( y(s_0) \) and the parameters \( \psi \). Note that (8) is defined in terms of a vector where \( A \) controls the main quantities of predictive interest. A new location \( s_{\text{new}} \) is chosen such that,

\[
s_{\text{new}} = \arg \min_{s_0} \text{IMSE}(s_0) = \arg \min_{s_0} \int_{\psi} \int_{S} \text{MSE}(s \mid y(s_0), \psi) d\psi d\mathcal{Y} \]  

(9)

where \( \psi \mid \mathcal{Y} \) is the posterior distribution \( \psi \).

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

\[
\mathbb{E}_{[\psi \mid \mathcal{Y}]} \left[ \int_{S} \text{MSE}(s \mid y(s_0), \psi) d\psi \right] 
\]

\[
\approx \mathbb{E}_{[\psi \mid \mathcal{Y}]} \left[ \sum_{g=1}^{G} \text{MSE}(s^*_g \mid y(s_0), \psi) \right]
\]

(10)

where \( \{s^*_1, \ldots, s^*_G\} \) is regularly spaced grid of locations on \( S \) with equal grid area \( \Delta \). Equation (10) suggests an intuitive Monte Carlo approach to calculating IMSE. Namely, given a potential sampling location \( s_0 \), at each iteration in a Markov chain Monte Carlo algorithm to sample \( \psi \) from its posterior distribution (i) draw \( y(s_0) \) from its predictive distribution, (ii) draw \( \{y(s^*_g) : g = 1, \ldots, G\} \) from the
predictive distribution conditional on \(y(s_0)\) and (iii) retain \(\Delta \sum_{g=1}^{G} \|Ay(s^*_g) - \hat{Ay}(s^*_g)\|^2\). Calculation of (10) is then carried out via Monte Carlo integration using the retained values of \(\Delta \sum_{g=1}^{G} \|Ay(s^*_g) - \hat{Ay}(s^*_g)\|^2\). 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 \(\{\theta_j, \kappa^2_j, \omega_j, \phi_j\}\) for \(j = 1, \ldots, 5\). Non-informative, Jeffery’s priors were chosen for \(\theta_j\) and \(\kappa^2_j\), leading to closed form complete conditional distributions. As mentioned above, the priors for \(\omega_j\) and \(\phi_j\) were assumed to be discrete uniform with 20 values for each \(\omega_j\) and 10 values for each \(\phi_j\). We note that these discrete uniform prior distributions also lead to closed form complete conditional distributions for each \(\omega_j\) and \(\phi_j\). 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 \(U(0, 0.01)\) proposal distributions to update all missing values simultaneously.

To obtain draws from the joint posterior distribution, we ran a Gibbs sampler for 5,000 iterations after a burn-in of 200 iterations. Despite the few number of iterations, we found convergence to be acceptable among the estimated parameters (trace plots showed adequate mixing) and 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.

While continuous basis function expansions of \(\gamma_{jk}\) 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 \(\gamma_{jk}\) to a model using Gaussian kernel basis functions. For the both models, the basis function matrix in Equation (5) is of the form

\[
B_{jk} = (1_n, W),
\]

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

\[
w_{i\ell} = \frac{1}{\sqrt{2\pi\lambda^2}} \exp\left\{-\frac{\|s_i - s^*_\ell\|^2}{2\lambda^2}\right\}
\]

where we have \(L\) “knots” \(s^*_1, \ldots, s^*_L\) and \(\lambda^2\) 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

\[
w_{i1} = \mathbb{1}_{R_1}(s),
\]

where \(R_1\) denotes the set of locations corresponding to region 1. Additionally, to assess the need for spatially varying cross-correlations, we consider a spatially constant model with \(B_{jk} = 1_n\).

To define the regions in the regionally constant model, we cluster the observations using gradient clustering [Heaton et al., 2015] and knowledge of the
soil construct of the field. 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 USDA Soil Survey (i.e. the two regions have fundamentally different soils according to the USDA). Though there is slight discrepancy between the two, we are comfortable using the gradient clustering method as both delineations have some level of noise. 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.

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, CRPS, and DIC. The cross-validated bias is defined as

$$\text{Bias} = \frac{1}{5 \times 17} \sum_{i=1}^{17} \sum_{d=1}^{5} (\hat{y}_{d}(s_i) - \hat{y}_{d,-i}(s_i))$$

where $\hat{y}_{d,-i}(s_i)$ is the posterior predictive mean of $y_{d}(s_i)$ with the $i^{th}$ observation omitted. Likewise, RMSE is defined as

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

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. 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 compare the DICs from the regionally-constant, non-Markov models for each approach. The respective DICs are 422.130 and 452.96, supporting the deepest to shallowest conditioning.

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 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 $\beta_{1j}$, 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 Future 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 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 A matrices. One simple choice in A is the matrix $A_1 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \end{pmatrix}$ (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 A is the matrix

$$A_2 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$ (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 $A_1$ and the right using $A_2$. 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 Models

To demonstrate that the methods introduced in this paper indeed produce valid results, 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 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 of the six models described in Section 4. That is, we considered three different specifications of basis functions $B_{jk}$ (regionally constant bases, Gaussian kernel bases, and non-spatially 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.

Having successfully validated our model selection process, we examine the estimated coverages for $\beta_j$ and $\gamma_{jk}$ for all depths $j = 1\ldots5$ 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 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 specify 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 $32^{nd}$ observation where this additional observation was chosen using one of three criteria: (i) the proposed IMSE criterion with $A = I$, (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, not only are we interested in point prediction but the uncertainty associated with these predictions. As such, define the cumulative generalized predictive variance as

$$\text{CGV} = \sum_{d=1}^{5} |\Sigma_{d,U|O}|$$

as a measure of predictive uncertainty where $\Sigma_{d,U|O}$ is the conditional variance of the unobserved locations given the observed locations at the $d^{th}$ depth and $|$ 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

As stated in the Introduction, 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.

While WHC is, generally, temporally static, the soil water content is not. For example, water can be added to the soil through irrigation or rain and removed from the soil by the crops or evaporation (a process called evapotranspiration). An interesting extension of this work would be to temporally model water levels in the soil resulting in a spatio-temporal multivariate model with non-stationary cross-correlations. This would be a natural extension to the WHC problem, in that instead of relying only on the soil’s water holding capacity, farmers could understand how the actual soil water content varies over time and space.

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.

Acknowledgements

The authors would like to thank Dr. Shane Reese for providing helpful discussion and insight into the multivariate IMSE criterion. This material is based upon work supported by the National Science Foundation under Grant Number DMS-1417856. 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.
Tables and Figures

Figure 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
Figure 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 $m^3/m^3$.

Table 1: Empirical cross-correlations between WHC at various depths and spatial locations. Group assignments correspond to those displayed in Figure 2.

<table>
<thead>
<tr>
<th>Correlation Between Depths</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,2)</td>
<td>0.7684212</td>
<td>0.6864782</td>
<td>0.4295057</td>
<td>0.9999946</td>
</tr>
<tr>
<td>(1,3)</td>
<td>0.7343041</td>
<td>0.6554364</td>
<td>0.6609153</td>
<td>0.7321721</td>
</tr>
<tr>
<td>(1,4)</td>
<td>0.7054258</td>
<td>0.7218852</td>
<td>0.8907092</td>
<td>0.300151</td>
</tr>
<tr>
<td>(1,5)</td>
<td>0.3267703</td>
<td>-0.2109739</td>
<td>-0.7901776</td>
<td>0.9960362</td>
</tr>
<tr>
<td>(2,3)</td>
<td>0.7474385</td>
<td>0.9961967</td>
<td>0.3976585</td>
<td>0.7299231</td>
</tr>
<tr>
<td>(2,4)</td>
<td>0.1199806</td>
<td>0.9814045</td>
<td>0.3542135</td>
<td>0.2970052</td>
</tr>
<tr>
<td>(2,5)</td>
<td>0.1338251</td>
<td>0.5607749</td>
<td>-0.8645057</td>
<td>0.9957376</td>
</tr>
<tr>
<td>(3,4)</td>
<td>0.4863032</td>
<td>0.9891101</td>
<td>0.9202209</td>
<td>0.8694766</td>
</tr>
<tr>
<td>(3,5)</td>
<td>-0.2781055</td>
<td>0.5815133</td>
<td>-0.7387598</td>
<td>0.7898544</td>
</tr>
<tr>
<td>(4,5)</td>
<td>0.1634437</td>
<td>0.4748326</td>
<td>-0.7733717</td>
<td>0.3838083</td>
</tr>
</tbody>
</table>
Figure 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.

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).

<table>
<thead>
<tr>
<th></th>
<th>Bias</th>
<th>RMSE</th>
<th>CVG</th>
<th>TCVG</th>
<th>PIW</th>
<th>CRPS</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>0.038</td>
<td>0.110</td>
<td>0.941</td>
<td>0.882</td>
<td>0.248</td>
<td>4.931</td>
<td>422.130</td>
</tr>
<tr>
<td>R-M</td>
<td>0.038</td>
<td>0.113</td>
<td>0.906</td>
<td>0.941</td>
<td>0.238</td>
<td>4.959</td>
<td>444.405</td>
</tr>
<tr>
<td>SC</td>
<td>0.042</td>
<td>0.109</td>
<td>0.953</td>
<td>0.882</td>
<td>0.235</td>
<td>4.799</td>
<td>493.499</td>
</tr>
<tr>
<td>SC-M</td>
<td>0.043</td>
<td>0.111</td>
<td>0.894</td>
<td>0.824</td>
<td>0.238</td>
<td>4.900</td>
<td>509.154</td>
</tr>
<tr>
<td>G</td>
<td>0.031</td>
<td>0.090</td>
<td>0.988</td>
<td>0.941</td>
<td>0.260</td>
<td>4.186</td>
<td>512.303</td>
</tr>
<tr>
<td>G-M</td>
<td>0.031</td>
<td>0.088</td>
<td>0.976</td>
<td>0.941</td>
<td>0.260</td>
<td>4.046</td>
<td>496.531</td>
</tr>
</tbody>
</table>
Figure 4: The location-specific posterior means of the posterior predictive distribution for WHC (in $m^3/m^3$) across the spatial field.
Figure 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.

Table 3: Estimated correlations between depths for each of the two regions

<table>
<thead>
<tr>
<th>Depths</th>
<th>Southwest Region</th>
<th>Northeast Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,2)</td>
<td>0.800</td>
<td>0.285</td>
</tr>
<tr>
<td>(1,3)</td>
<td>0.731</td>
<td>-0.181</td>
</tr>
<tr>
<td>(1,4)</td>
<td>0.192</td>
<td>-0.158</td>
</tr>
<tr>
<td>(1,5)</td>
<td>-0.084</td>
<td>-0.387</td>
</tr>
<tr>
<td>(2,3)</td>
<td>0.577</td>
<td>0.597</td>
</tr>
<tr>
<td>(2,4)</td>
<td>-0.020</td>
<td>0.695</td>
</tr>
<tr>
<td>(2,5)</td>
<td>-0.073</td>
<td>0.455</td>
</tr>
<tr>
<td>(3,4)</td>
<td>0.582</td>
<td>0.903</td>
</tr>
<tr>
<td>(3,5)</td>
<td>-0.216</td>
<td>0.791</td>
</tr>
<tr>
<td>(4,5)</td>
<td>-0.008</td>
<td>0.793</td>
</tr>
</tbody>
</table>
Figure 6: WHC ($m^3/m^3$) 95% Predictive Interval Widths
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.

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

Figure 7: IMSE for all 100 candidate locations, the lowest being indicated by a star.
Figure 8: Comparison of true to estimated IMSE from simulation. Best locations indicated with star.
References


