RESPONSE TO COMMENTS FROM REVIEWER 2

1. *I was left wondering how other alternative models would perform in prediction. For example, you might consider a separable or nonseparable covariance function (Gneiting 2002; JASA) although I recognize there are only 5 depths. Or perhaps even something as simple as a separable LMC (e.g., using a kronecker product of the 55 covariance matrix with the 1717 spatial correlation matrix). I know this is a very simple specification but it might allow you to explore the need for a space-varying covariance structure more fully.*

   Valid point. While we did explore simpler, alternate models, we failed to include corresponding discussion in the first draft of the article. We remedied this by comparing the final model to a non-spatially varying model (described on page 12), and discussing the shortcomings of a univariate model in three dimensions (section 1.2) with the following text:

   > "An intuitive approach would be to model WHC as a Gaussian process over a three-dimensional domain (lat-lon-depth). This model presents two major issues, however, which make it impracticable for our application. First, when considering all 5 depths, the number 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."

   Our subsequent comparison of simpler models showed that the regional model was preferred (in terms of DIC) to simpler approaches (see Table 2 in the main article).

2. *I was surprised that all inference was based on only 4800 iterations of a single MCMC chain. Was the computation so onerous that you could not run multiple chains out longer and consider more formal convergence diagnostics?*

   In fact, the opposite was true. Using our parameterization, this is a rather simple model at each depth (multiple linear regression). Hence, we didn't need a large number of iteration to achieve sufficient MC standard error and desirable mixing properties. Figure 1 of this document illustrates this by displaying several example trace plots. Additionally, the design piece of this analysis is computationally costly so, given the adequate mixing at 4800 draws, we elect for fewer draws.

   In the article is Section 4, we added the following point in this regard:

   > "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."

3. *There was not enough information provided about the various candidate model specifications presented in Table 2. The proposed model (2) is fine but I had trouble following the construction of the alternative models described starting with the third paragraph in Section 4.1. What model is being used with the regionalization approach? What are the implications to other strata specifications for these regional models. Without better understanding these alternative models it is difficult to assess the results.*

   We added the following discussion to the article to better clarify the candidate models:

   > "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 the both models, the basis function matrix in Equation (5) is of the form
   >
   > $$\boldsymbol{B}_{jk} = (\mathbf{1}_n, \boldsymbol{W}), \tag{1}$$
   >
   > where $\mathbf{1}_n$ denotes a length n vector of ones, and $\mathbf{W} = \{w_{i\ell}\}$ 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
   >
   > $$w_{i1} = \mathbb{1}_{\mathcal{R}_1}(\boldsymbol{s}), \tag{2}$$
   >
   > where $\mathcal{R}_1$ 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 $\boldsymbol{B}_{jk} = \mathbf{1}_n$."

4. *It would be helpful to consider holdout validation scoring rules beyond RMSE, e.g., CRPS (Gneiting and Raftery 2007; JASA).*

   As suggested, we added CRPS as an additional metric in Table 2.

5. *I recognize Table 3 provides the space-varying loading estimates, but it would be useful to generate cross-correlation (or cross-covariance) summaries comparable to Table 1. How well is the fully space-varying model capturing the nonstationarity observed in the exploratory data analysis?*

   This is a good point. It does seem more natural to include the estimated correlations instead of the less-interpretable loading estimates. We changed Table 3 to now include estimated correlations between depths.

6. *Following from the previous point, it would be helpful to see a simulated data analysis. One of the conclusions was there was just not enough data to allow for substantial differentiation among the candidate models. A synthetic data analysis would allow for further exploration of the proposed modeling framework strengths.*

We agree that a simulation study would be a valuable addition to the paper as a proof of concept. We therefore added Section 5.1 to the paper, which outlines the design and results of a simulation study with the purpose of ensuring that model parameters are learned from the data rather than overly influenced by the prior specification, and validating the use of the model selection criteria summarized in Table 2. The results of the simulation show comparable performance in terms of predictive power, but DIC is indeed best in the Regionally Constant non-Markov model (agreeing with results in Table 2).

7. *Although acknowledged in Section 5, the MAR assumption is clearly violated and the solution for missing data imputation is not very satisfying and, given the small sample size with substantial missingness 14/17 locations effected, could have substantial inferential implications.*

   We agree that the effects of the MAR violation could have far-reaching implications. For this application, according to those who collected the data, the data is missing if the WHC is very low which immediately implies the data are not missing at random. Thus, we altered the imputation approach by constraining the imputed values to fall below the minimum observed WHC (in this case, 0.01). This constraint on the imputed values is a more realistic treatment of the missing values.

   We added the following description to the article:

   > "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; **?**). 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."

   We also mention this again in Section 4 where we say:

   > "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."