A Multiscale Spatio-Temporal Big Data Fusion Algorithm from Point to Satellite Footprint Scales

Dhruva Kathuria[1], Binayak P Mohanty[1], and Matthias Katzfuss[1]

[1]Texas A&M University

November 22, 2022

Abstract

The past six decades has seen an explosive growth in remote sensing data across air, land, and water dramatically improving predictive capabilities of physical models and machine-learning (ML) algorithms. Physical models, however, suffer from rigid parameterization and can lead to incorrect inferences when little is known about the underlying physical process. ML models, conversely, sacrifice interpretation for enhanced predictions. Geostatistics are an attractive alternative since they do not have strong assumptions like physical models yet enable physical interpretation and uncertainty quantification. In this work, we propose a novel multiscale multi-platform geostatistical algorithm which can combine big environmental datasets observed at different spatio-temporal resolutions and over vast study domains. As a case study, we apply the proposed algorithm to combine satellite soil moisture data from Soil Moisture Active Passive (SMAP) and Soil Moisture and Ocean Salinity (SMOS) with point data from U.S Climate Reference Network (USCRN) and Soil Climate Analysis Network (SCAN) across Contiguous US for a fifteen-day period in July 2017. Using an underlying covariate-driven spatio-temporal process, the effect of dynamic and static physical controls—vegetation, rainfall, soil texture and topography—on soil moisture is quantified. We successfully validate the fused soil moisture across multiple spatial scales (point, 3 km, 25 km and 36 km) and compute five-day soil moisture forecasts across Contiguous US. The proposed algorithm is general and can be applied to fuse many other environmental variables.

A Multiscale Spatio-Temporal Big Data Fusion Algorithm from Point to Satellite Footprint Scales

Dhruva Kathuria[1], Binayak P. Mohanty[1], and Matthias Katzfuss[2]

[1]Biological and Agricultural Engineering, Texas A&M University, College Station, Texas, USA

[2]Department of Statistics, Texas A&M University, College Station, Texas, USA.

Corresponding author: Binayak P. Mohanty (bmohanty@tamu.edu)

Corresponding author address: 301A Scoates Hall, Texas A&M University, College Station, Texas, 77840, USA.

Keywords: data fusion, remote sensing, big data, soil moisture, geostatistics, multi-scale, spatio-temporal data

Abstract

The past six decades has seen an explosive growth in remote sensing data across air, land, and water dramatically improving predictive capabilities of physical models and machine-learning (ML) algorithms. Physical models, however, suffer from rigid parameterization and can lead to incorrect inferences when little is known about the underlying physical process. ML models, conversely, sacrifice interpretation for enhanced predictions. Geostatistics are an attractive alternative since they do not have strong assumptions like physical models yet enable physical interpretation and uncertainty quantification. In this work, we propose a novel multiscale multi-platform geostatistical algorithm which can combine big environmental datasets observed at different spatio-temporal resolutions and over vast study domains. As a case study, we apply the proposed algorithm to combine satellite soil moisture data from Soil Moisture Active Passive (SMAP) and Soil Moisture and Ocean Salinity (SMOS) with point data from U.S Climate Reference Network (USCRN) and Soil Climate Analysis Network (SCAN) across Contiguous US for a fifteen-day period in July 2017. Using an underlying covariate-driven spatio-temporal process, the effect of dynamic and static physical controls—vegetation, rainfall, soil texture and topography—on soil moisture is quantified. We successfully validate the fused soil moisture across multiple spatial scales (point, 3 km, 25 km and 36 km) and compute five-day soil moisture forecasts across Contiguous US. The proposed algorithm is general and can be applied to fuse many other environmental variables.

1 Introduction

On April 1, 1960, (National Aeronautics and Space Administration) NASA launched the Television and Infrared Observation Satellite (TIROS 1) demonstrating that satellites could observe weather patterns, marking the advent of remote sensing (RS) to observe global environmental phenomena. Sixty years and the launch of several satellites later, rapid progress has been made in observing Earth-system processes (across air, land, and water) accompanied by an explosion in the availability of data. This so called "big data" are often spatio-temporal (indexed by a spatial coordinate and a time stamp) resulting in an increased interest in space-time problems in the past two decades (Gelfand et al., 2010; Wikle et al., 2019). Usually, environmental data are 1) spatio-temporally dependent, 2) available at multiple resolutions from various instruments, and 3) observed with gaps and noise. It is unreasonable to expect one source of data to fill all the gaps across space and time. However, combining multi-sensor data, while accounting for individual strengths and weaknesses, can lead to novel insights into Earth-system Science. Paradigms facilitating the fusion of disparate data while handling the sheer size of datasets are thus critical.

RS data have traditionally been used to update the states and improve parameterization of physically based models. Indeed, the assimilation of satellite data into numerical weather prediction models led to the "quiet revolution" (Bauer et al., 2015) in global weather prediction. Data assimilation has also found success in oceanography (Evensen, 1994; Ghil & Malanotte-Rizzoli, 1991) and land-surface hydrology (Reichle et al., 2002). Physical models are vital for predicting variables poorly observed by RS platforms such as ocean mixed layer (Wang et al., 2000) and root-zone soil moisture

(SM) (Lievens et al., 2017). However, the rigid parameterization of physical models can be a hindrance when knowledge of the underlying spatio-temporal process is incomplete (Girotto et al., 2017). The resulting predictions can suffer from signatures of strong (and sometimes incorrect) assumptions (Akbar et al., 2019). Moreover, RS observations usually need to be pre-processed for correcting bias and scale-mismatch before assimilation in the numerical model (Koster et al., 2009).

The recent decade has seen an incredible rise of Machine Learning (ML) in Earth-System Sciences, which has been instrumental in improving predictive accuracy of disparate physical processes (Camps-valls et al., 2013; Hengl et al., 2017; Jung et al., 2010; Mao et al., 2019; Shi et al., 2017). Though classical ML models are inept at accounting for spatio-temporal dependence, recent research in Deep Learning seems promising (Fang et al., 2017; Shen, 2018; Shi et al., 2017). Accuracy without interpretability, however, is insufficient (Reichstein et al., 2019); the lack of transparency and physical interpretability of many ML models is viewed as a major deficiency. Moreover, current state-of-the-art ML models are ill-equipped to handle some of the major challenges associated with fusing RS data such as accounting for multi-sensor multiscale data, uncertainty in observations and predictions, and missing data (Reichstein et al., 2019).

On an interpretation-prediction spectrum, physical models derived from the first laws of physics lie on one end while ML algorithms using black-box models fall on the other. Geostatistics lie somewhere in the middle and are an attractive alternative for spatio-temporal inference in a data-driven setting. They do not have strong assumptions like physical models yet enable physical interpretation and uncertainty quantification. From

its humble origins in South African mines (Cressie, 1990; Krige, 1952), geostatistics has been widely used in modeling the spatio-temporal distribution of environmental variables including precipitation (Cecinati et al., 2017), temperature (Lanfredi et al., 2015), soil properties (Lark, 2012; Mohanty et al., 1991, 1994; Mohanty & Kanwar, 1994), carbon dioxide (Zhong & Carr, 2019), ground-water quality (Goovaerts et al., 2005) and SM (Joshi & Mohanty, 2010; Kathuria et al., 2019a; Mohanty et al., 2000). Recent work on covariate-driven non-stationary models have also enabled the seamless integration of covariates into geostatistical models (Reich et al., 2011; Risser & Calder, 2015) enabling them to model complex spatio-temporal phenomena.

Geostatistical approaches typically assume an underlying Gaussian process (GP) requiring quadratic memory and cubic time complexity in the number of observations, which make them prohibitive as the data size increases. Various approximations have therefore been proposed for applying geostatistics to massive datasets. Such approaches generally aim at approximating the covariance (e.g., Kaufman et al., 2008) and inverse-covariance matrices (e.g., Nychka et al., 2015). Among these, the *Vecchia* approximation (Vecchia, 1988) is one of the oldest with several advantages such as it is 1) suitable for high-performance parallel computing, 2) accounts for uncertainty in predictions, and 3) outperforms several state-of-the-art approaches in accuracy (Guinness, 2018). Moreover, recent work (Katzfuss et al., 2020; Katzfuss & Guinness, 2017) has shown that *Vecchia* approximation can be generalized to include many existing GP approximation approaches as special cases. However, the use of the *Vecchia* approximation, to the best of the authors' knowledge, has been restricted to single-scale data only.

Thus, the objective of this paper is to investigate whether geostatistics, with its rich parametric inference and uncertainty quantification, can potentially be used with *Vecchia* approximation to fuse spatio-temporal multiscale big data.. We achieve this by applying the *Vecchia* approximation to a geostatistical hierarchical model (Gelfand et al., 2001; Kathuria et al., 2019b). In this paper, we define the term "multiscale big data" as data which are observed from multiple platforms at varying footprints, are massive in size, and are observed over vast extents rendering standard geostatistical (and many other statistical) approaches infeasible.

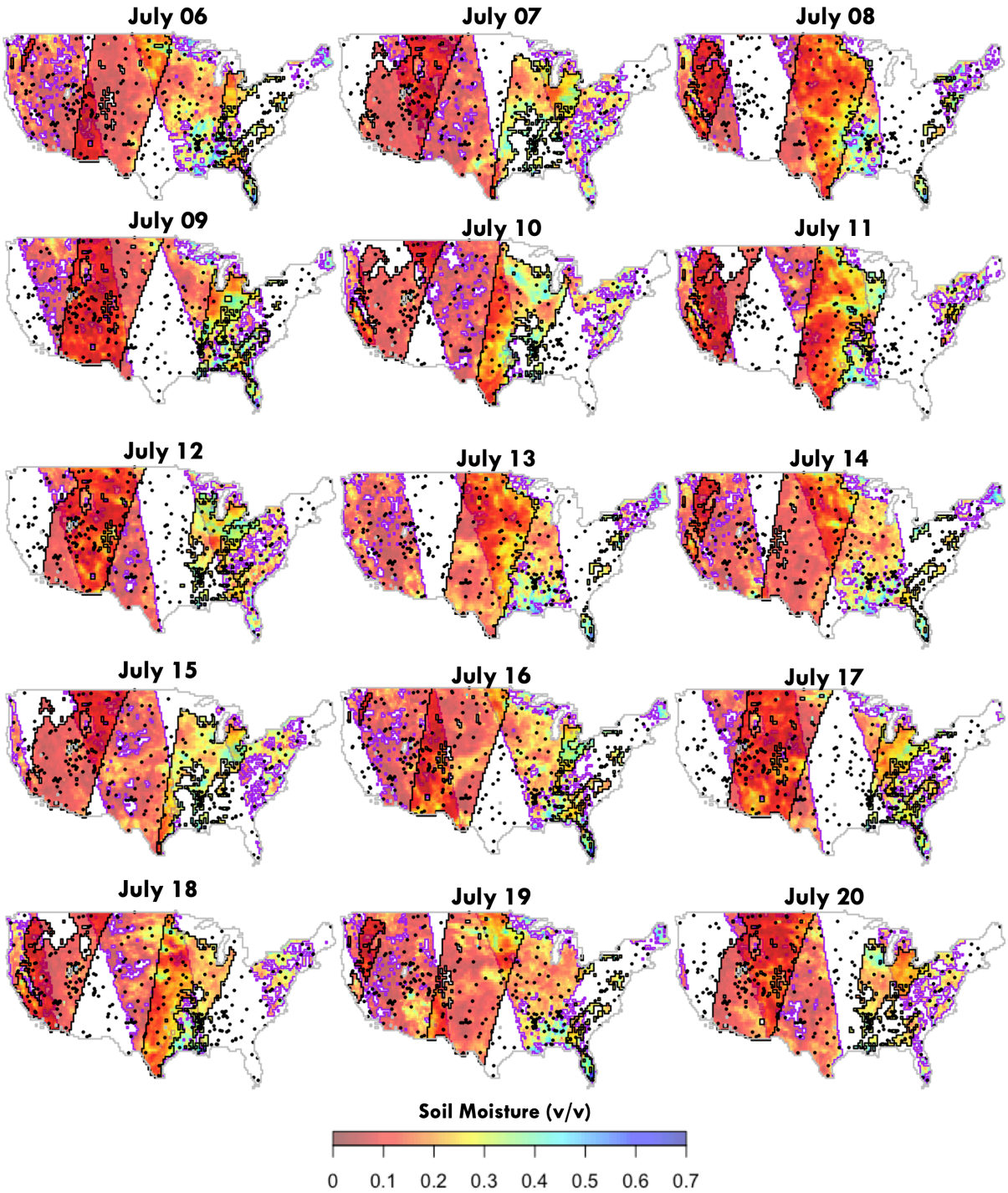
We explore the utility of the approximation using simulations, and by fusing real SM datasets as a case study. SM is a critical variable governing land-atmosphere interactions and contains significant information about physical processes such as rainfall (Koster et al., 2016), streamflow (Koster et al., 2018) and evapotranspiration (ET) (Akbar et al., 2019). SM is highly correlated in space and time resulting from dynamic interactions between surface and atmospheric controls making it a prime candidate for geostatistics driven multiscale data fusion. Kathuria et al. (2019b) previously proposed a geostatistical data fusion scheme for combining multiscale SM data but its application was restricted to regions with small extent and small data size limiting its utility. We also choose SM as a case study application for our proposed algorithm to provide a big data closure for Kathuria et al. (2019b). The rest of the paper is organized as follows. We describe the SM datasets used in the case study in Section 2. The data fusion algorithm along with its big data extension is detailed in Section 3. This is followed by the discussion of results in Section 4 before we conclude in Section 5. Note that in the following sections, all vectors are assumed to be column vectors.

2 Study Area and Data

2.1 Case Study: Soil moisture

We apply the proposed algorithm to combine daily point surface (top 0-5 cm) SM data from U.S. Climate Reference Network (USCRN) (Diamond et al., 2013) and Soil Climate Analysis Network (SCAN) (Schaefer et al., 2007) with satellite data from Soil Moisture Ocean Salinity (SMOS) (Barré et al., 2008) and Soil Moisture Active Passive (SMAP) (Entekhabi et al., 2010) for Contiguous US (CONUS) for July 06-20, 2017. This fifteen-day time interval was randomly chosen for the warm summer period so that the effect of snow on SM estimation is minimal. For any given day, there are approximately 143 sites for USCRN and SCAN while individual satellites partially observe SM across CONUS with some overlap between the two data sets (Figure 1).

Both SMOS and SMAP use L-band radiometers to measure surface brightness temperature (T_b) at an average revisit time of three days (Colliander et al., 2017; Pablos et al., 2019). Both the satellites apply (different) retrieval algorithms to T_b and generate composite daily L3 SM products resampled, at 36 km for SMAP (L3) and 25 km for SMOS (Barcelona Expert Center L3), to an Equal Area Scalable Earth (EASE)-2 grid. For the SMAP data we remove the pixels where 1) the retrieval was unsuccessful (using flag data), and 2) where the vegetation water content is greater than 5 kg/m^2 (O'Neill et al., 2018). For consistency we use the morning overpass for both satellites— 6 AM local time. For the covariate data, daily rainfall data were extracted from Parameter-elevation Regressions on Independent Slopes Model (PRISM) at 4 km resolution. PRISM provides gridded rainfall data across CONUS at a daily scale using a combination of climatological and statistical methods (Daly et al., 1994). Soil and elevation data were



157
158

Figure 1. Fifteen day soil moisture data from USCRN and SCAN (black cross), SMOS (swath - black outline) and SMAP (swath - purple outline) for July 06-20, 2017. For individual days, both SMOS and SMAP observe different regions of Contiguous US (CONUS) and there is a significant overlap between the data. The size of the SM data and the extent of study domain (CONUS) are both massive making data fusion computationally demanding.

159 extracted from Soil Survey Geographic Database (1 km) (Soil Survey Staff, 2020) and
160 Leaf Area Index (LAI) (as a proxy for vegetation) were extracted from Moderate
161 Resolution Imaging Spectroradiometer (MCD15A3H, 500m) (Myneni et al., 2015).
162
163 **3 Methodology**
164 **3.1 Multiscale data fusion**
165 Let the environmental variable varying across space and time (such as SM, ET,
166 temperature, etc.) be denoted by y. We assume that $y(.)$ is a Gaussian Process (GP) (a
167 standard geostatistical assumption) at the point scale in a domain or extent $\mathfrak{D}$ in d
168 dimensions ($d = 1, 2, 3 \ldots$). For instance, if y represents daily land-surface temperature
169 (LST) varying spatially (latitude and longitude) and temporally (days), then d equals 3.
170 The variable y is defined at the point scale using a mean function μ and a covariance
171 function C:
172
173 $$y(.) \sim GP(\mu, C). \quad (1)$$
174
175 For any environmental variable y, in addition to point data, we might observe data at
176 aggregate resolutions from RS platforms or large-scale numerical models. For instance,
177 surface SM is observed at aggregate resolutions from SMAP ($\sim 36\ km \times 36\ km$, daily)
178 and SMOS ($\sim 25\ km \times 25\ km$, daily) while ET is observed using ECOSTRESS ($\sim$
179 $70m \times 70m$, daily) and MODIS ($\sim 500m \times 500m$, 8-day). Since y is defined at point
180 scale, for any aggregate pixels A_i and A_j, $y(A_i) = \frac{1}{|A_i|}\int_{A_i} y(s)ds$, with the corresponding
181 mean and covariance as:

182

183 $\mu(A_i) = \frac{1}{|A_i|}\int_{A_i}\mu(s)ds$ and

184 $$C(A_i, A_j) = \frac{1}{|A_i|}\frac{1}{|A_j|}\int_{A_i}\int_{A_j} C(s_1, s_2)ds_1 ds_2, \quad (2)$$

185

186 where $|A_i|$ is the d-dimensional resolution of pixel A_i and s represents a point in d
187 dimensions. If A_i and A_j represent coordinates of point data, the mean of data at A_i is
188 simply $\mu(A_i)$ and the covariance between A_i and A_j is given as $C(A_i, A_j)$. If A_i is an areal
189 pixel and A_j represents a point, then the covariance $C(A_i, A_j)$ is given as $\frac{1}{|A_i|}\int_{A_i} C(s, A_j)ds$.

190

191 Let the total number of observed pixels be n and be denoted by $\mathcal{A} = \{A_1, \dots, A_n\}$ with
192 $A_i \subset \mathfrak{D}$. The joint distribution of $y(\mathcal{A}) = (y(A_1), \dots, y(A_n))$ can be shown to be
193 multivariate normal (Gelfand et al., 2001):

194

195 $$y(\mathcal{A}) = \mathcal{N}_n(\mu(\mathcal{A}), C(\mathcal{A}, \mathcal{A})), \quad (3)$$

196

197 where $\mu(\mathcal{A})$ is a vector of length n and $C(\mathcal{A}, \mathcal{A})$ is a matrix of size $n \times n$. The
198 individual elements of $(\mu(\mathcal{A}))_i$ and $(C(\mathcal{A}, \mathcal{A}))_{ij}$ are given by equation 2. Since we
199 cannot always analytically solve the above integrals, we use a numerical approximation
200 (Gelfand et al., 2001) by assuming an equidistant numerical grid $\mathcal{G}$ over the extent $\mathfrak{D}$
201 with $n_{\mathcal{G}}$ number of grid points such that $\mathcal{G} = \{g_1, \dots, g_{n_{\mathcal{G}}}\}$ or equivalently $\mathcal{G} = \{g_k : k = 1, \dots, n_{\mathcal{G}}\}$.
202 Here g_k denotes the location of the k^{th} grid point in $\mathcal{G}$. We can then
203 approximate $y(A_i)$ as:

$$y(A_i) \approx \frac{1}{n_{A_i}} \sum_{g_k \in \mathcal{G}_{A_i}} y(g_k), \tag{4}$$

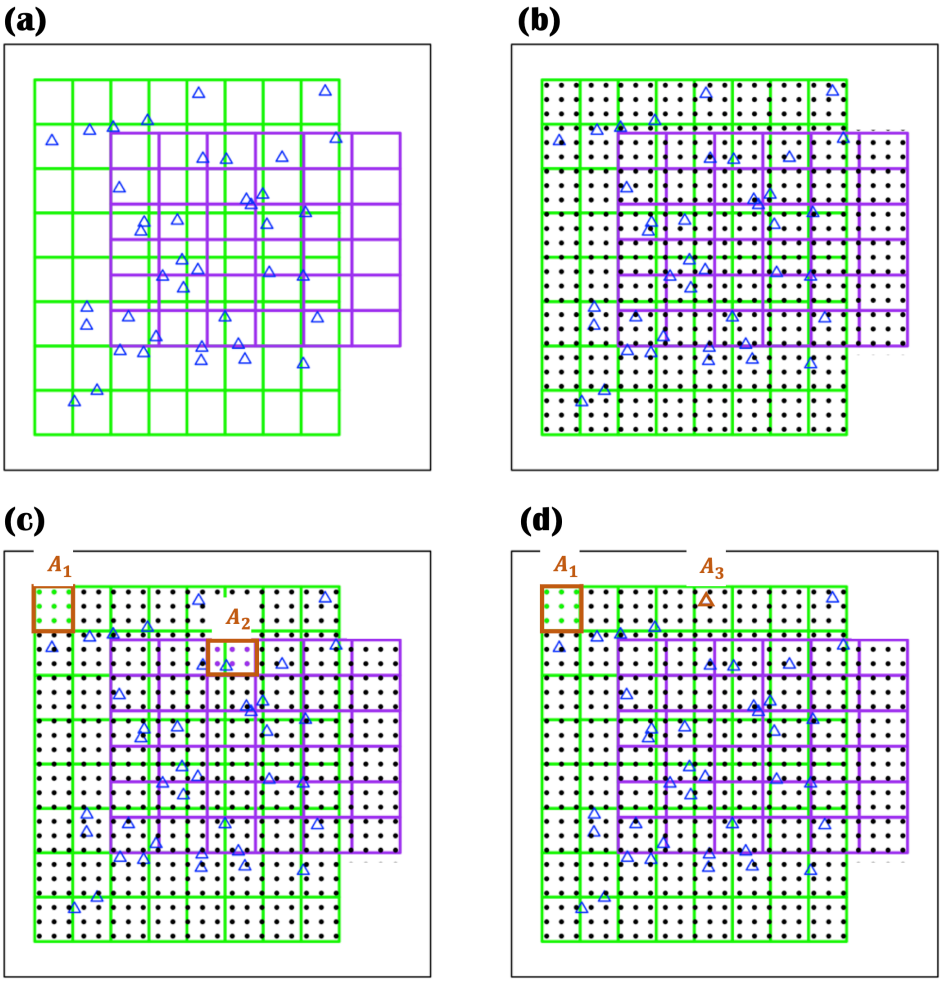
where $\mathcal{G}_{A_i}$ denotes the subset of the total grid points $\mathcal{G}$ lying inside the pixel A_i ,and n_{A_i} denotes the number of grid points in $\mathcal{G}_{A_i}$. The corresponding approximations for the mean and covariance can be written as:

$$\mu(A_i) \approx \frac{1}{n_{A_i}} \sum_{g_k \in \mathcal{G}_{A_i}} \mu(g_k),$$

$$C(A_i, A_j) \approx \frac{1}{n_{A_i}} \frac{1}{n_{A_j}} \sum_{g_k \in \mathcal{G}_{A_i}} \sum_{g_l \in \mathcal{G}_{A_j}} C(g_k, g_l). \tag{5}$$

We illustrate the numerical approximation using a hypothetical example in Figure 2. Figure 2 (a) represents three partially overlapping datasets which cover different extents and have different resolutions: two areal datasets R_1 (64 green pixels) and R_2 (36 purple pixels), and point dataset P_1 (40 blue triangles). Figure 2 (b) represents the equidistant grid $\mathcal{G}$ (black dots) over the study domain. Assuming the mean and covariance functions are known at the point scale, the mean of pixel A_1 (A_2) and the covariance between pixels A_1 and A_2 in Figure 2 (c) are given by equation 5. Here $\mathcal{G}_{A_1}(\mathcal{G}_{A_2})$ are subset of the total grid points $\mathcal{G}$, color-coded as green (purple), lying inside A_1 (A_2) with $n_{A_1} = 9$ ($n_{A_2} = 6$). Similarly, the mean function at point A_3 in Figure 2 (d) is simply given as $\mu(A_3)$ while $C(A_1, A_3)$ is given by $\frac{1}{n_{A_1}} \sum_{g_k \in \mathcal{G}_{A_1}} C(g_k, A_3)$.

225



226
227

Figure 2. (a) Example depicting two areal (green and purple) and one point (blue triangles) data platforms (b) Equidistant point grid assumed throughout the study domain (c) The mean and covariance of pixels A_1 and A_2 approximated using the numerical grid (d) The mean and covariance between a pixel A_1 and point observation A_3.

We can write $\frac{1}{n_{A_i}}\sum_{g_k \in \mathcal{G}_{A_i}} y(g_k)$ (equation 4) in matrix form as $h_{A_i}^T y_{A_i}$, where h_{A_i} is a vector of length n_{A_i} with each element equal to $1/n_{A_i}$ or $h_{A_i} = (1/n_{A_i}, \ldots\ldots, 1/n_{A_i})$, and y_{A_i} is a vector of length n_{A_i} with elements $\{y(g_k): g_k \in \mathcal{G}_{A_i}\}$. Similarly in equation 5, $\mu(A_i)$ can be written as $h_{A_i}^T \mu_{A_i}$ (with μ_{A_i} having elements $\{\mu(g_k): g_k \in \mathcal{G}_{A_i}\}$). We also write $C(A_i, A_j)$ in equation 5 in matrix form as $h_{A_i}^T (C(\mathcal{G}_{A_i}, \mathcal{G}_{A_j})) h_{A_j}$, where (as mentioned before) $\mathcal{G}_{A_i}$ denotes the subset of the total grid points $\mathcal{G}$ lying inside the pixel A_i.

Retrievals of an environmental variable from different platforms are typically subject to systematic (bias) and stochastic (random) errors (e.g. refer Fan et al. (2020) and Reichle & Koster (2004) for SM, Li et al. (2014) and Westermann et al. (2012) for LST, Klees et al. (2007) for water storage, Hu et al. (2015) and Velpuri et al. (2013) for ET). Thus, for any observed pixel A_i, it is important to differentiate between the noisy observation from a platform (denoted as $z(A_i)$) and the latent environmental variable $y(A_i)$ that is uncorrupted by the parameterized errors. For a given observation $z(A_i)$ (from a data platform) for pixel A_i, we thus write:

$$z(A_i) = y(A_i) + \delta(A_i) + \kappa(A_i) y(A_i) + \epsilon(A_i), \quad (6)$$

where $\delta(A_i)$, $\kappa(A_i)$ and $\epsilon(A_i)$ are respectively the additive bias, multiplicative bias, and random measurement error associated with $z(A_i)$. We parameterize the random error as $\epsilon(A_i) \sim \mathcal{N}(0, \tau_{A_i}^2)$ with variance $\tau_{A_i}^2$. We then write:

$$z(A_i) \approx h_{A_i}^T y_{A_i} + \delta(A_i) + \kappa(A_i) h_{A_i}^T y_{A_i} + \epsilon(A_i)$$
$$= \big(1 + \kappa(A_i)\big) h_{A_i}^T y_{A_i} + \delta(A_i) + \epsilon(A_i)$$
$$= (h_{A_i}^{\kappa})^T \, y_{A_i} + \delta(A_i) + \epsilon(A_i), \quad (7)$$

where $h_{A_i}^{\kappa} = \big(1 + \kappa(A_i)\big) h_{A_i}^T$. The mean ($\mu$) and covariance function (C) in equation 1 are thus given parametric forms based on the environmental variable y while the additive bias ($\delta(A_i)$), multiplicative bias ($\kappa(A_i)$) and error-variance ($\tau_{A_i}^2$) for a pixel A_i in equation 7 are parameterized depending on the data platforms. Let all the parameters used to parameterize the mean, covariance, bias and random error be denoted by the vector θ. Elements of θ can either assumed to be known or be estimated from the observations. If the total number of observations from all platforms is equal to n, we denote $z(\mathcal{A}) = \{z(A_1), z(A_2), \dots, z(A_n)\}$. The parameter vector θ is estimated by maximizing the likelihood $f(z(\mathcal{A})|\theta)$ where $f(A|B)$ denotes the probability density of A given B. For our model, it can be easily derived that the (log-) likelihood is:

$$-2log(f(z(\mathcal{A})|\theta) = log(det(\Sigma_z)) + (z(\mathcal{A}) - \mu_z)^T \Sigma_z^{-1} (z(\mathcal{A}) - \mu_z) + nlog(2\pi), \quad (8)$$

where the i^{th} element of the vector μ_z (size n) and the $(i,j)^{th}$ element of the matrix Σ_z (size $n \times n$) in equation 8 are given as:

$$\mu_{z,i} \approx (h_{A_i}^{\kappa})^T \mu_{A_i} + \delta(A_i),$$
$$\Sigma_{z,ij} \approx (h_{A_i}^{\kappa})^T (C(\mathcal{G}_{A_i}, \mathcal{G}_{A_j})) h_{A_j}^{\kappa} + \tau_{A_{i,j}}^2, \quad (9)$$

where $\tau^2_{A_{i,j}} = \begin{cases} \tau^2_{A_i}, i = j \\ 0, i \neq j \end{cases}$. However this data fusion algorithm becomes computationally infeasible when the size of the datasets and/or the extent of study domain becomes large. We therefore propose an approximation to the fusion algorithm for such cases in the next Section.

3.2 *Vecchia-multiscale*: An Approximation for Multiscale Big Data

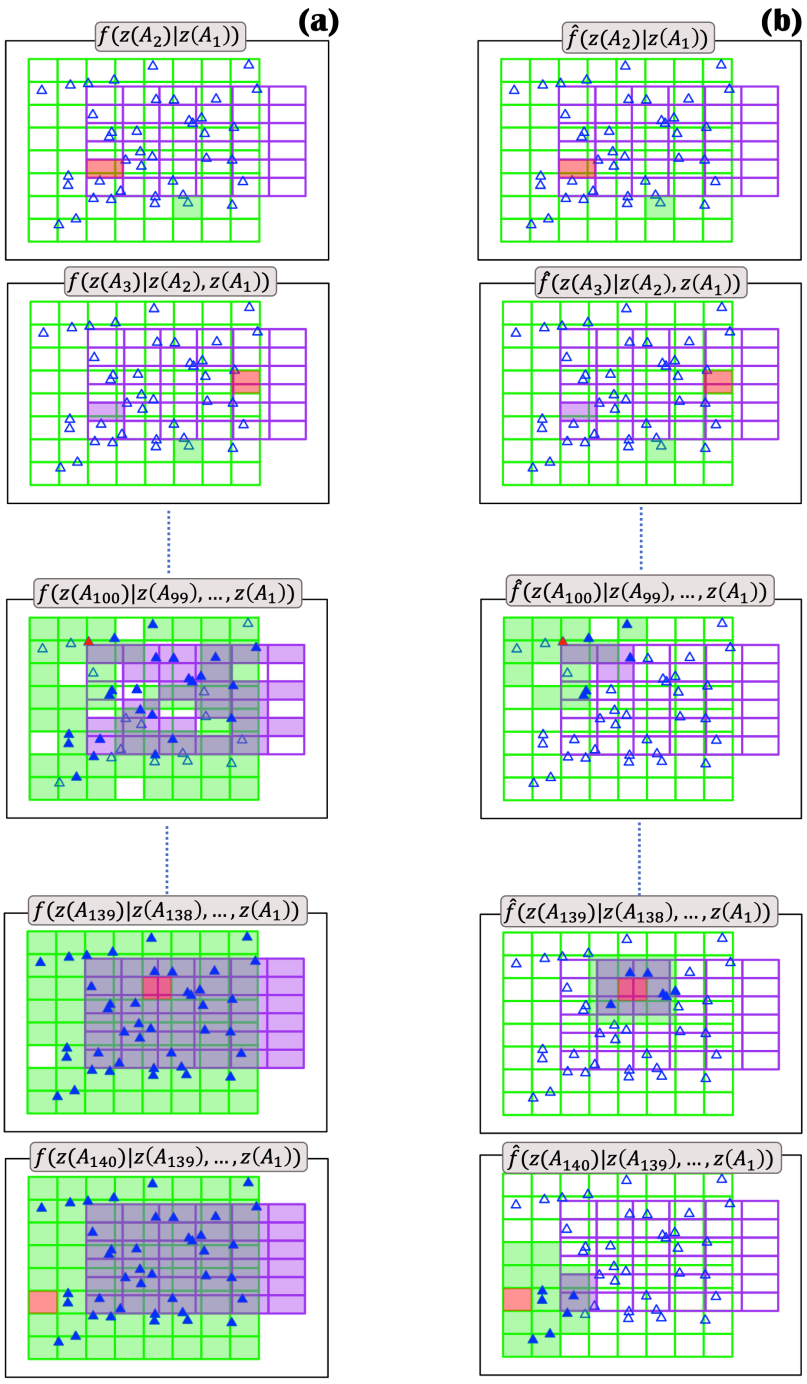
If the total number of observations (governed by the number of data platforms and resolution of pixels for a given study domain) be n, and the number of assumed grid points (governed by the extent of the study domain and distance between individual grid points) be $n_{\mathcal{G}}$, then computing Σ_z and finding its inverse Σ_z^{-1} in equation 8 requires $\mathcal{O}(n_{\mathcal{G}}^2) + \mathcal{O}(n^3)$ floating point operations. This evaluation becomes computationally prohibitive as the number of data and the size of study domain increase (e.g., when combining multiple data platforms for continental scale fusion of an environmental variable), and thus requires an approximation. To approximate the likelihood, we first write the joint distribution in $f(z(\mathcal{A})|\theta)$ as a product of univariate conditional distributions as

$$f(z(\mathcal{A})|\theta) = f(z(A_1)|\theta) \times \prod_{i=2}^{n} f(z(A_i)|\boldsymbol{z}(\boldsymbol{A_{1:i-1}}), \theta), \tag{10}$$

where $\boldsymbol{A_{1:i-1}}$ denotes $\{A_1, \dots, A_{i-1}\}$ and thus $\boldsymbol{z}(\boldsymbol{A_{1:i-1}})$ denotes $\{z(A_1), \dots, z(A_{i-1})\}$. Following Vecchia (1988) we approximate the likelihood $f(z(\mathcal{A})|\theta)$ as:

$$\hat{f}(z(\mathcal{A})|\theta) = f(z(A_1)|\theta) \times \prod_{i=2}^{n} f(z(A_i)|\boldsymbol{z}(\boldsymbol{A_{m_i}}), \theta), \tag{11}$$

where $\boldsymbol{A}_{\boldsymbol{m}_i}$ is a subvector of $\boldsymbol{A}_{\boldsymbol{1:i-1}}$ of length m_i such that $m_i = \begin{cases} i-1, i \leq m \\ m, i > m \end{cases}$. Here m is an integer lying between 1 and $n-1$ with $m = n-1$ representing the exact likelihood in equation 10. The elements of subvector $\boldsymbol{A}_{\boldsymbol{m}_i}$ consist of m_i elements from $\boldsymbol{A}_{\boldsymbol{1:i-1}}$ which are closest to A_i in space. The subvector $\boldsymbol{z}(\boldsymbol{A}_{\boldsymbol{m}_i})$ is the observed data vector corresponding to $\boldsymbol{A}_{\boldsymbol{m}_i}$. To illustrate the approximation, we again use the hypothetical example in Figure 2 (a) comprising three datasets: areal data R_1 (64 green pixels) and R_2 (36 purple pixels), and point data P_1 (40 blue triangles), making the total number of observations $n = 140$. For this data, the univariate conditional distributions are illustrated in Figure 3 using a random permutation of the pixels $\mathcal{A}$ and choosing $m = 20$. Column (a) presents the conditional distributions in equation 10 corresponding to the exact likelihood while column (b) consist of the corresponding conditional distributions resulting from the *Vecchia* approximation. The i^{th} pixel A_i in equations 10 and 11 (where $i = 2, \ldots, 140$ increases from top to bottom in the columns) is color-filled in red while the pixels (or points) of the conditioning vector $\boldsymbol{z}(\boldsymbol{A}_{\boldsymbol{1:i-1}})$ (equation 10) or $\boldsymbol{z}(\boldsymbol{A}_{\boldsymbol{m}_i})$ (equation 11) are color-filled in green (R_1), purple (R_2) and blue (P_1). It can be seen in Figure 3 that for $i > m$, the Vecchia approximation selects a subset of m pixels (or points) for each A_i. It can be shown that this approximation is equivalent to inducing sparsity (large percentage of zeros) in the inverse Cholesky factor matrix Λ ($\Lambda^T\Lambda = \Sigma_z^{-1}$). This leads to fast evaluation of Σ_z^{-1} (and consequently the likelihood) in equation 8 used for estimating the parameter vector θ as well as doing subsequent predictions. The detailed algorithm for parameter estimation and subsequent predictions is given in Appendix A1. We call this approximation *Vecchia-multiscale.*



315

Figure 3. Illustration of the Vecchia-multiscale to the hypothetical data in Figure 2(a) consisting of 64 green pixels (R_1), 36 purple pixels (R_2) and 40 point data P_1 (blue triangles). Column (a) denotes the conditional distributions as implied by the the exact likelihood while column (b) gives the conditional distributions using Vecchia-multiscale approximation with maximum size of the conditioning vector m equal to 20. The i^{th} pixel A_i (where $i = 2, \ldots, 140$ increases from top to bottom in the columns) is color-filled in red while the pixels (or points) of the conditioning vector are color-filled in green (R_1), purple (R_2) and blue (P_1).

3.2.1 Permutation in Vecchia-multiscale

There are two criteria we seek in the approximation: speed and accuracy. For the *Vecchia-multiscale*, significant computational and memory benefits can be achieved by selecting $m \ll n$. Further, equation 11 results in a product of independent univariate distributions which is readily parallelized for faster computations.

Regarding accuracy for a fixed value of m, as the right side of equation 11 consists of an "ordered" sequence of conditional probability distributions, the approximation depends on the order in which the pixels appear in $\mathcal{A}$. This is because in equation 11, for a pixel A_i ($i \geq 2$), we select the subset $\boldsymbol{A}_{\boldsymbol{m}_i}$ (of length m_i) from elements of $\boldsymbol{A}_{\boldsymbol{1:i-1}}$ which are closest in space to A_i. This leads to different values for $\boldsymbol{z}(\boldsymbol{A}_{\boldsymbol{m}_i})$ in equation 11 based on how we permute $\{A_1, \ldots, A_n\}$. Thus, the approximation accuracy will depend upon what permutation of $\{A_1, \ldots, A_n\}$ we choose for the pixels (and points) for computing $\hat{f}(z(\mathcal{A})|\theta)$ in equation 11. When the size of the multiscale data is massive, it is infeasible to explore all such permutations. For point data, Guinness (2018) found that certain permutations of $\mathcal{A}$ give more accurate approximations when compared with the exact likelihood $f(z(\mathcal{A})|\theta)$. In this paper we explore the same for multiscale data. We use four popular permutations (Guinness, 2018): 1) *Joint-Coordinate* (ordering the locations based on increasing coordinate values), 2) *Joint-Middleout* (ordering locations based on increasing distance to the mean location of the extent), 3) *Joint-Maxmin* (ordering in which each successive point is chosen to "maximize the minimum distance" to previously selected points), and 4) *Joint-Random* (randomly

ordering locations). Interested readers are encouraged to refer to Section S1, Supporting Information (SI) and Guinness (2018) for details on these permutations.

In addition to the above "Joint-" permutations, we introduce "Separate-" permutations where we first separate out the point and areal data and apply the above-mentioned four permutations separately to each. We then form the final permutation by sorting the "ordered" point data followed by the "ordered" areal data. This leads to four additional corresponding permutations: 5) *Separate-Coordinate*, 6) *Separate-Middleout*, 7) *Separate-Maxmin*, and 8) *Separate-Random*. The difference between "Joint-" and "Separate-" permutations is illustrated in Figure 4. We assume the centroid of an areal pixel as its location for applying the permutations.

Using the hypothetical example in Figure 2 (a), we illustrate the effect of these eight chosen permutations (Section S1, SI) on how the pixels and points are ordered in $\mathcal{A}$ and how it affects the evaluation of $\hat{f}(z(\mathcal{A})|\theta)$. To see which permutation performs better for the *Vecchia-multiscale* in general, we use simulated data in two (e.g, a variable varying across latitude and longitude) and three (e.g., a variable varying across latitude, longitude and time) dimensions. The details of the simulations and the corresponding results are given in Section S2, SI.

For both two and three dimensions, in general, the *Separate-Maxmin* and *Separate-Random* perform the best while the *Coordinate-based* orderings perform the worst. This is important because many approximation schemes use *Coordinate-based ordering* as their default (e.g. Datta et al., 2016; Sun & Stein, 2016) and it should be used with caution

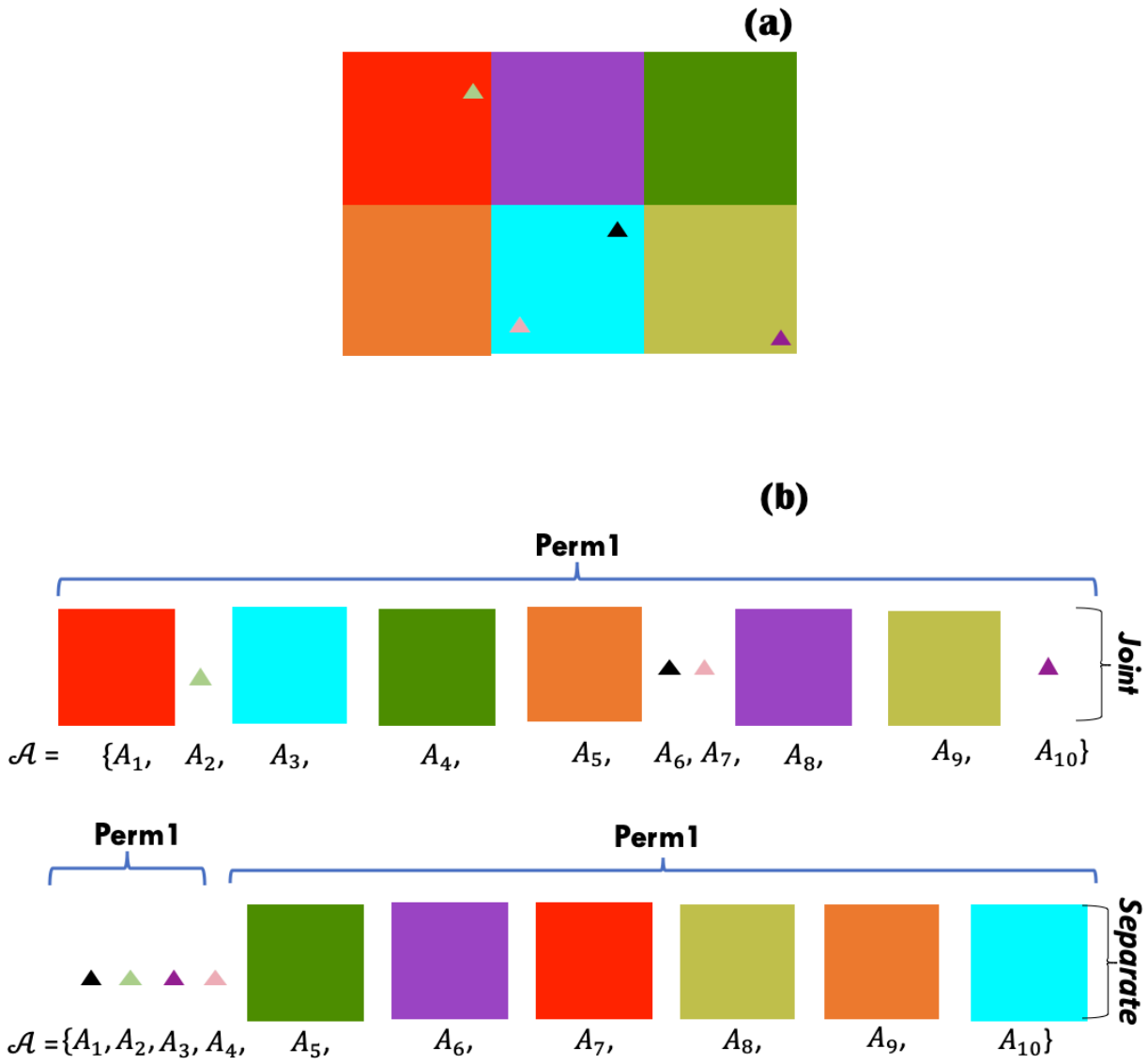


Figure 4. Illustration of "Joint-" and "Separate-" permutations for Vecchia-multiscale. (a) Hypothetical example comprising six aggregate pixels and four point data. Different colors are used to distinguish between different pixels and points. (b) The "Joint-" permutation results in both the pixels and points getting permuted together following a given permutation "Perm1". For "Separate-" ordering, we first separate the point and aggregate data, apply the permutation "Perm1" separately to each, and then form the final permutation by sorting the permuted point data followed by the permuted aggregate data. In this figure we choose a random permutation as "Perm1" and the resulting permutations of the pixels/points are shown. The "Joint-" and "Separate-" permutations can lead to different ordering of the pixels/points in $\mathcal{A} = \{A_1, \ldots, A_{10}\}$ resulting in different values of the approximate likelihood computed using Vecchia-multiscale. In this paper, we explore "Coordinate", "Middleout", "Maxmin" and "Random" as possible permutations for "Perm1". The centroid of an aggregate pixel is chosen as its location for permutations.

when using *Vecchia-multiscale*. The subvector $\boldsymbol{A_{m_i}}$ (equation 11) consists of a good mix of both far and near pixels as well as nearby point data for *Separate-Maxmin* (Figure S2 (i)-(l), SI) and *Separate-Random* (Figure S2 (m)-(p), SI). We hypothesize that conditioning a pixel/point on both near and far pixels help in better approximation of the exact likelihood. Additionally, the "Separate-" permutations lead to the subvector $\boldsymbol{A_{m_i}}$ consist of nearby point data which is potentially helpful because 1) for a given study domain, point data are generally sparse for any environmental variable and are generally (but not always) considered more accurate than remote sensing data, and 2) we define our model at the point scale (equation 1), and it is thus potentially helpful to condition pixels/points on nearby point data.

We therefore suggest adopting *Separate-Maxmin* or *Separate-Random* when using *Vecchia-multiscale.* Since, our aim is to propose a general algorithm, we only use location information for permuting $\{A_1, \ldots, A_n\}$. A promising area of future research is exploring physically-based permutation of pixels based on the environmental variable to be fused. In the next Section, we apply the *Vecchia-multiscale* to fuse multiscale SM data for CONUS.

4 Results and Discussion

4.1. Case Study : Soil moisture

We fuse fifteen days of SMOS, SMAP, and point (USCRN and SCAN) SM data across CONUS from July 06-20, 2017. We randomly hold-out 27 point stations ($\approx$ 20%) for validation leaving 116 station data for training. Since SM observations are theoretically bounded between 0 and 1 and exhibit considerable skewness, the Gaussian assumption

becomes untenable. We thus use a logit transform $SM' = log(\frac{SM}{1-SM})$ which transforms the SM values to lie between $-\infty$ to ∞ and also make the distribution less skewed (Figure S4, SI). Overlapping data from SMOS and SMAP during the analyzed period also exhibit slightly better correlation on the transformed scale (Figure S5, SI).

4.2.1 Mean, covariance and bias

Numerous studies (Cosh & Brutsaert, 1999; Crow et al., 2012; Entin et al., 2000; Gaur & Mohanty, 2013, 2016; Joshi et al., 2011; Joshi & Mohanty, 2010; Kathuria et al., 2019a; Ryu & Famiglietti, 2006; Teuling & Troch, 2005; Vereecken et al., 2014) have found that SM distribution across space and time is affected primarily by precipitation, soil texture, topography and vegetation. Therefore, we model the spatio-temporal SM distribution as a function of these physical covariates. For SMAP, since we only consider pixels where SM retrieval was successful (from flag data) and have a vegetation water content $\leq 5\ kg/m^2$, we assume that the SMAP data are of good quality and do not have any bias. As we did not pre-filter SMOS data, we assume a constant additive and multiplicative bias for SMOS. Exploratory analysis between overlapping SMOS-SMAP pixels at the logit scale (Figure S5, SI) also suggest a (additive and multiplicative) bias between the two platforms. We assume normally distributed measurement error (at the transformed scale) with mean zero and variance τ^2_{SMAP} and τ^2_{SMOS} for the two platforms respectively. Since the USCRN/SCAN data undergo rigorous quality control, we assume point data to be the ground truth with no bias/error.

We use exploratory analysis for determining the parametric forms for the mean function. Since we assume bias in SMOS data, we use only SMAP and point data for the

410 exploratory analysis. For the exploratory analysis, the covariates are linearly averaged
411 to the SMAP resolution. For rainfall, we assume 3-day antecedent mean rainfall as a
412 covariate. On the original scale (Figure 5 (a)), the relationship between SM and the
413 physical controls is non-linear. But after some non-linear transformations of the
414 covariates (and logit transform of SM), an approximate linear relationship between SM
415 and the covariates can be assumed (Figure 5 (b)). The mean trend of SM can be therefore
416 written as:
417

418 $$\mu\left(log\left(\frac{SM}{1-SM}\right)\right) = \mu(SM') = \beta_0 + \beta_1 log(LAI) + \beta_2 exp(-\frac{rain}{p^{\beta}_{rain}}) + \beta_3 exp(-\frac{elevation}{p^{\beta}_{elevation}}). \quad (12)$$

419

420 We fix p^{β}_{rain} and $p^{\beta}_{elevation}$ as 3.3 mm and 342.6 m based on exploratory analysis. These
421 two parameters represent the range of the exponential functions in equation 12 for
422 which an approximate linear relationship holds between SM' and the transformed
423 covariates in Figure 5 (b). Note that the covariates are resampled only for exploratory
424 analysis and no resampling of (SM and covariate) data is required for implementing the
425 actual algorithm in Section 3. Since we use an equidistant grid to approximate
426 multiscale SM data, the grid points are assigned values according to the covariate pixels
427 in which they lie. Though this results in grid points lying in a covariate pixel getting the
428 same values, this allows us to work with covariate data at different resolutions and
429 avoid errors introduced due to resampling of covariate data.
430

431 The covariance between any two points (x_1, y_1, t_1) and (x_2, y_2, t_2), where x, y, t
432 represent the latitude, longitude and time respectively, will also vary based on the
433 underlying covariate heterogeneity and therefore the assumption of a stationary

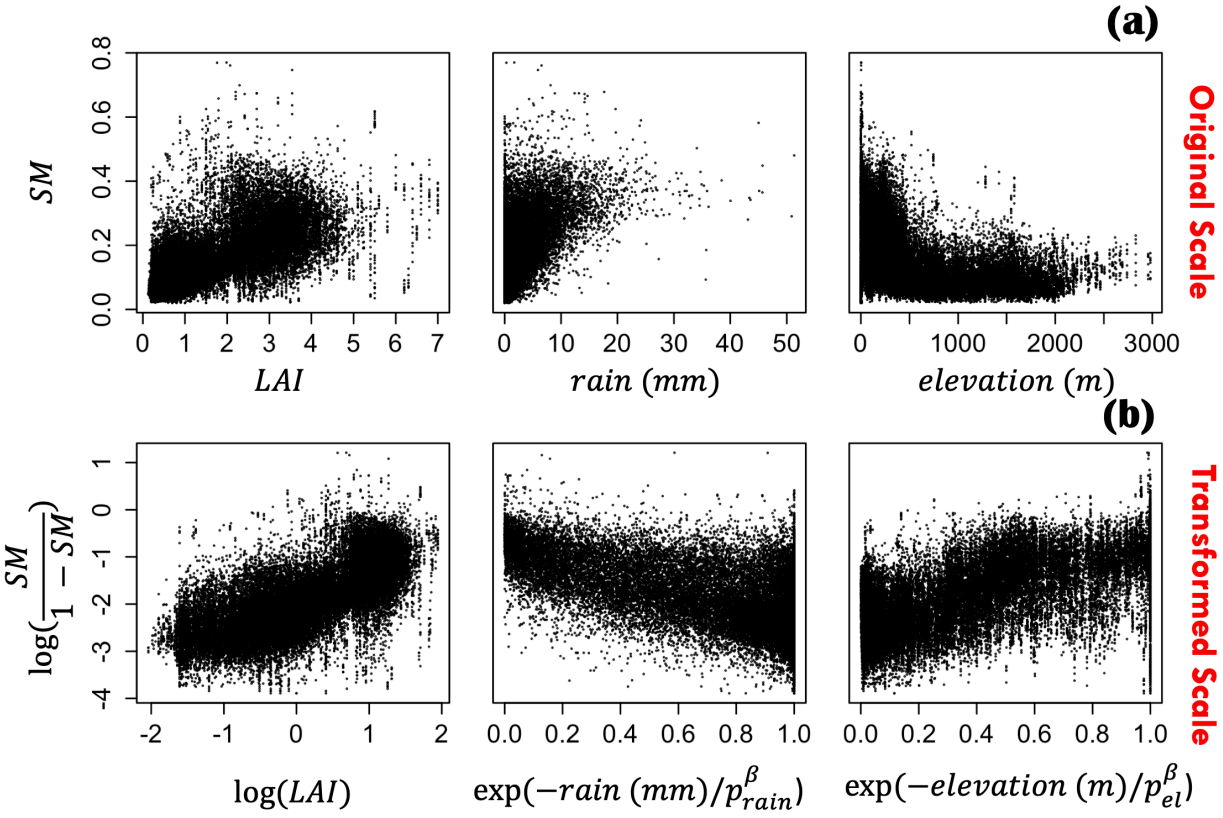


Figure 5 Exploratory analysis of soil moisture with physical covariates. (a) The relationship of soil moisture with the physical covariates is non-linear on the original scale. (b) Appropriate covariate transformation results in an approximate linear relationship of SM (on the logit scale) with the physical covariates. The values of p_{rain}^{β} *and* $p_{elevation}^{\beta}$ *are fixed as 3.3 mm and 342.6 m in the plots.*

436 covariance function is too simplistic. Thus, for the covariance function C (equation 1),
437 we use a non-stationary covariance function (Kathuria et al., 2019a; Reich et al., 2011)
438 such that:

439 $$C\big(SM'(x_1, y_1, t_1), SM'(x_2, y_2, t_2)\big) = C(s_1, s_2)$$

440 $$= \sum_{j=1}^{M} w_j(X_{cov}(s_1)) w_j(X_{cov}(s_2)) C_j(|s_1 - s_2|). \quad (13)$$

441 The covariance function in equation 13 is a weighted sum of M isotropic covariance
442 functions $\{C_j; j = 1,2, \ldots\ldots, M\}$ where the weights $\{w_j; j = 1,2, \ldots\ldots, M\}$ are a function of the
443 underlying physical covariates $X_{cov}(s)$ affecting the covariance. The weighting
444 functions w_js are modeled using a multinomial logistic function of the underlying
445 covariates: $w_j(s) = \frac{exp(X_{cov}(s)^T \alpha_j)}{\sum_{l=1}^{M} exp(X_{cov}(s)^T \alpha_l)}$. The details of the covariance function can be found
446 in Reich et al. (2011) and Kathuria et al. (2019a). For our analysis, we choose exponential
447 covariance functions (Matern with smoothness = 0.5) for individual C_js (equation 13)
448 with different range parameters for space (r_{xy}^j) and time (r_t^j) (e.g., Guinness, 2018):
449

450 $$C_j(s_1, s_2) = \sigma_j^2 exp\left(-\sqrt{\frac{||(x_1,y_1)-(x_2,y_2)||^2}{(r_{xy}^j)^2} + \frac{|t_1-t_2|^2}{(r_t^j)^2}}\right). \quad (14)$$

451 We chose the exponential covariance functions for individual C_js as changing the
452 smoothness parameter for Matern resulted in insignificant change in the estimated
453 maxmimum likelihood, and exponential functions are computationally faster to
454 evaluate than Matern due to the added cost of evaluating the Bessel functions for the
455 Matern function. We fix $M = 3$ to keep the number of parameters to be estimated
456 relatively low. We include LAI, three-day mean antecedent rain, clay and elevation in

457 X_{cov}. As mentioned in Section 3, both the mean and covariance functions are defined at
458 point scale with computations at areal supports done as outlined in Section 3.1. In this
459 work, since point data are sparse, the parameter estimates of the mean and covariance
460 functions are expected to be mainly driven by SMAP and SMOS data. Note that we do
461 not include latitude, longitude or time as covariates in either the mean or covariance
462 function to make the fusion scheme more general and transferable.
463
464 **4.2.2 Parameter estimation and inference**
465 We assume a numerical grid $\mathcal{G}$ (Section 3.1) spaced approximately 0.09 degrees apart
466 across the CONUS for each of the fifteen days resulting in close to 100,000 grid points
467 per day ($n_{\mathcal{G}} \approx 15 \times 100{,}000 = 1{,}500{,}000$). The total number of observations n from all
468 platforms (SMAP, SMOS, and USCRN/SCAN) for fifteen days equal 100,386. Parameter
469 estimation and subsequent predictions by computing exact likelihood is
470 computationally intractable for such a big dataset and thus requires an approximation.
471 We use the approximation detailed in Section 3 using the *Separate-Maxmin* orderings.
472 Since SMAP and SMOS observe SM at an interval of 3-7 days, we compute the *Separate-*
473 *Maxmin* ordering only considering the spatial coordinates (latitude and longitude) of
474 the data so that the temporal information of SM is also adequately represented in the
475 conditioning vector $\boldsymbol{z}_{\boldsymbol{m}_i}$ in equation 11. We fix the number of neighbors as $m = 60$; the
476 choice of m was taken to balance the predictive accuracy and computational speed. We
477 carry out parameter estimation using a global optimization algorithm called
478 Generalized Simulated Annealing (Xiang et al., 2013), a generalized and improved form
479 of simulated annealing, to find the parameter estimates that maximize the likelihood.
480

481 On the logit scale, the estimated mean parameters (equation 12) are $\beta = \{\beta_0, \beta_1, \beta_2, \beta_3\}$ =
482 {-1.71, 0.08, -0.35, 0.17} thus showing a good correlation of mean SM with the controls
483 especially antecedent rainfall. The additive and multiplicative bias for SMOS are $\delta =$
484 -0.003 and $\kappa = 0.15$ respectively, while the measurement error variance for SMAP
485 and SMOS are $\tau^2_{SMAP} = 0.026$ and $\tau^2_{SMOS} = 0.023$. To quantify the effect of covariates on
486 the spatio-temporal covariance of SM, we first transform the covariance to the original
487 scale. For a specified covariance between two points (from the covariance function in
488 equation 13) on the logit scale, we use the well-known Cholesky-Decomposition
489 method to simulate (50,000) pairs of values for these two points (Gong et al., 2013). We
490 then back-transform these values to the original scale and use the empirical covariance
491 of the pairs as an approximation of the covariance at the original scale.
492
493 For the non-stationary covariance function, since the covariance between any two
494 points depends on the lag-distance in space and time as well as the covariates(X_{cov}), the
495 effect of an individual covariate on the covariance is nontrivial. We thus quantify the
496 effect of a covariate by comparing the covariance for different lags (in space and time)
497 when the control is at the mean value (of the study domain) to when the control is at
498 extreme value (5^{th} and 95^{th} percentile) while keeping the other controls at their mean
499 values (Kathuria et al., 2019a; Reich et al., 2011). The resulting correlation plots are
500 given in Figure 6. We find that all four covariates affect the correlation in space with
501 higher values of rainfall, LAI, percent clay and lower values of elevation associated with
502 increase in spatial correlation. For the temporal correlation, we found only a slight effect
503 of the covariates on the correlation. Note, however, inclusion of other physical
504

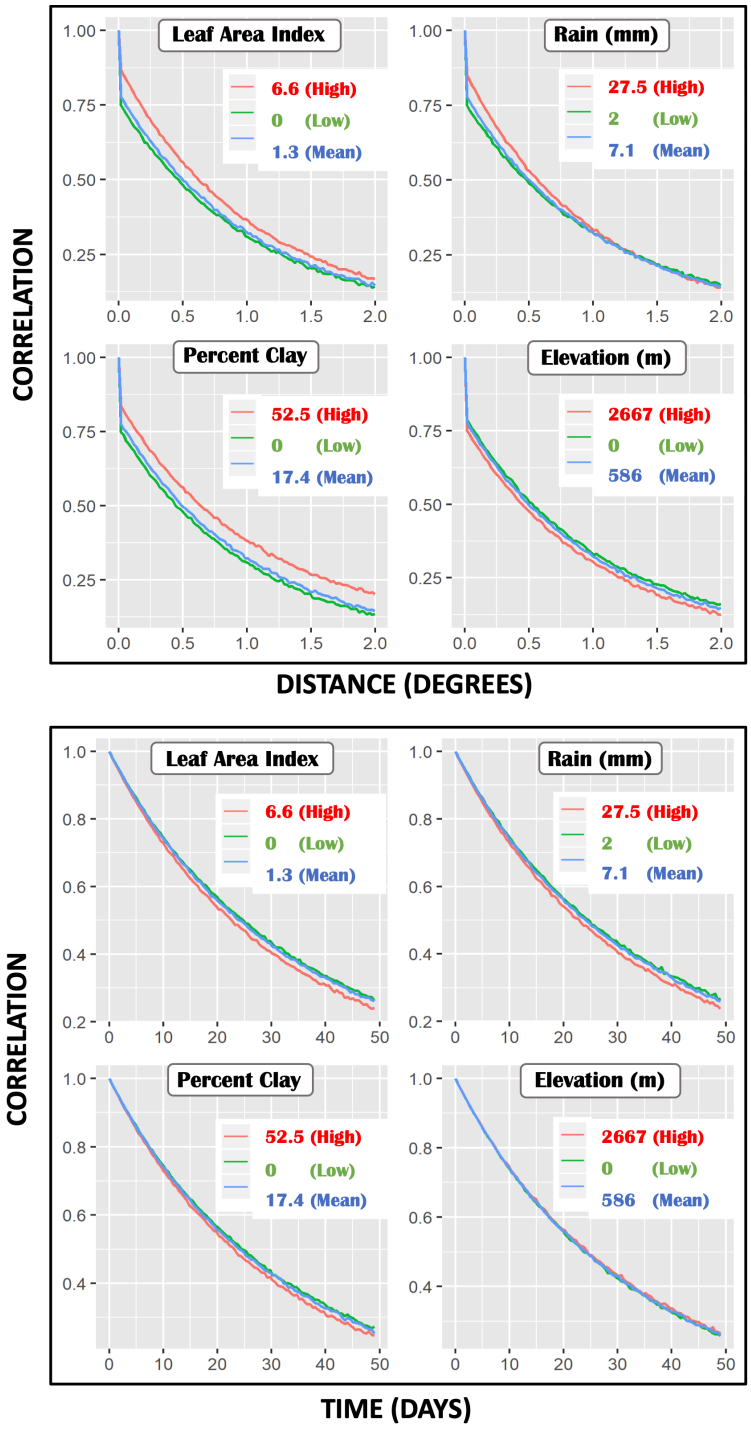
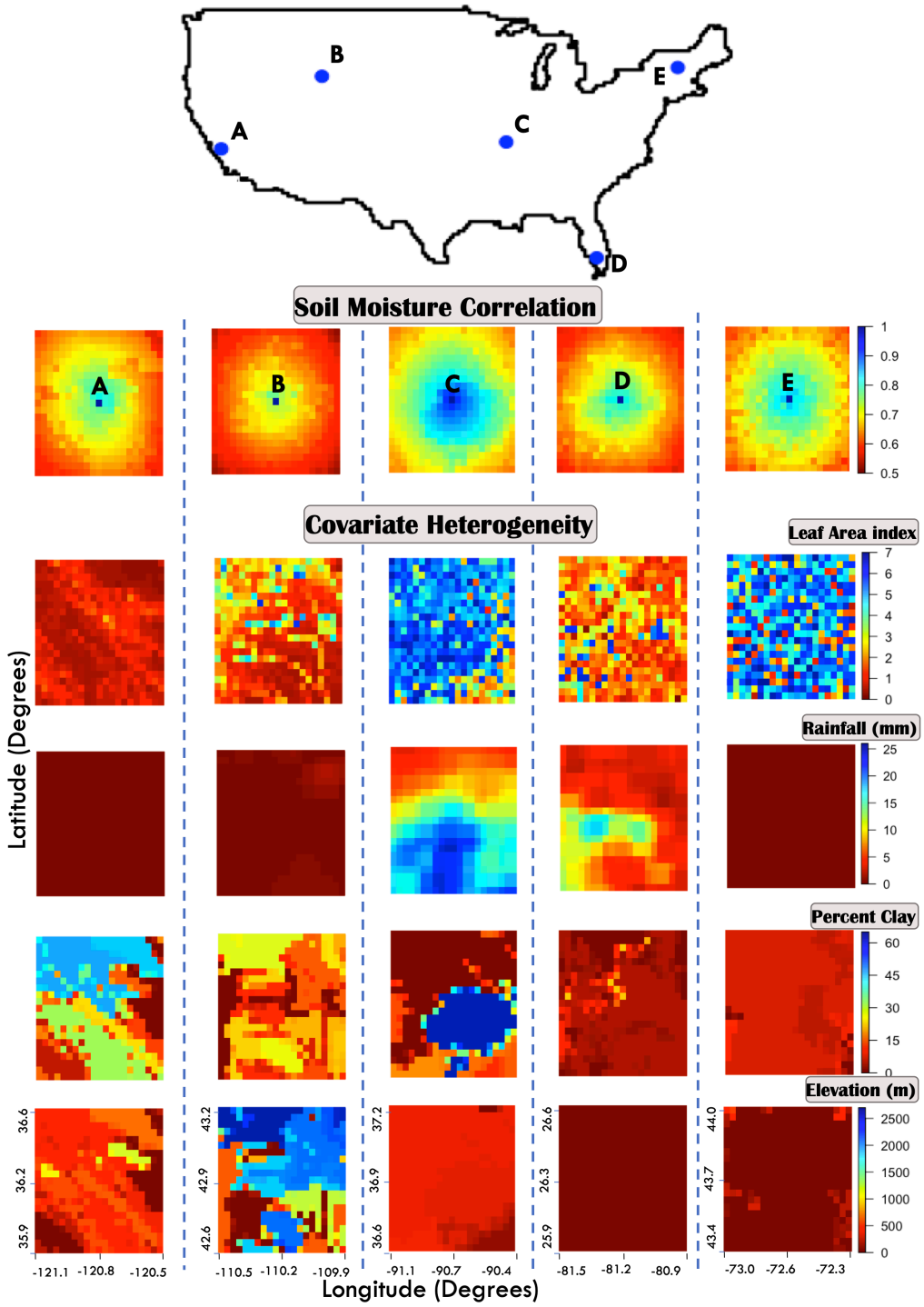


Figure 6. Spatial (top) and temporal (bottom) correlation plots when one of the physical covariates (Leaf Area Index, rainfall, clay and elevation) is changed from the mean value (of the study domain) to high (95^{th} percentile) and low values (5^{th} percentile). For each of the plots, the blue curve is the same representing the spatial and temporal correlation when all the covariates are at their mean values (LAI =1.3, Rain = 7.1 mm, Clay =17.4% and Elevation = 586 m) The red (green) curve refers to the correlation when one covariate is changed to a high (low) value keeping the other three covariates at the mean value.

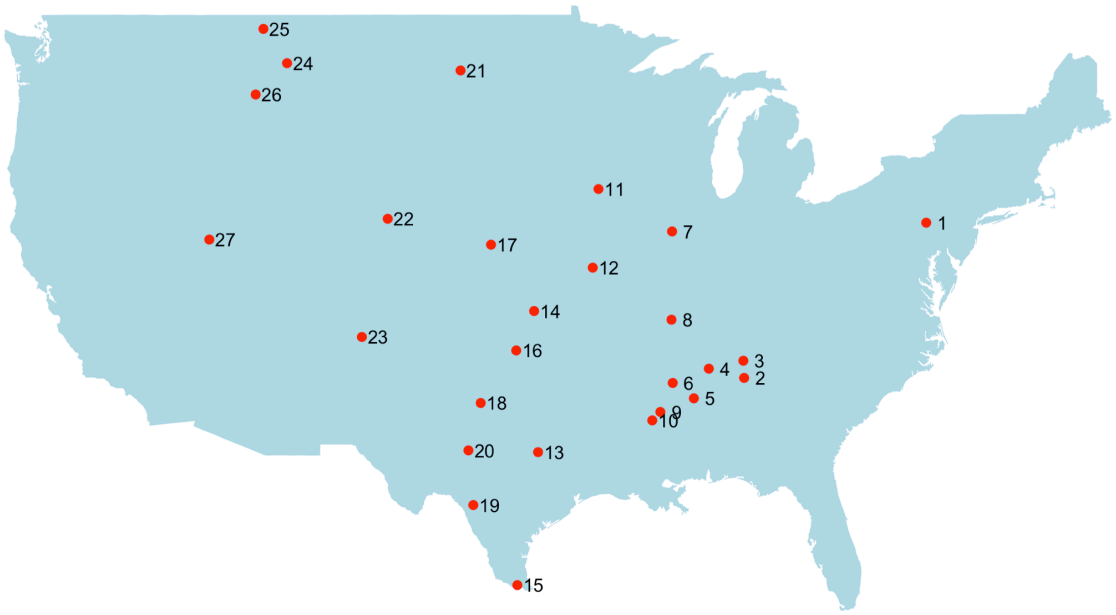
505

506 covariates as well as analysis of a longer time-period might show the effect of certain
507 covariates on the temporal SM correlation.
508
509 Of course, individual plots in Figure 6 represent only three combinations of the physical
510 covariates. In reality, all the covariates exhibit considerable heterogeneity across
511 CONUS (Figure S6, SI) and act together to give vastly different correlation patterns. To
512 illustrate this effect, we choose 5 points (A-E, Figure 7) across CONUS under
513 contrasting covariate heterogeneity and look at the spatial correlation of these points
514 with surrounding points (~ 3 km apart) within an approximately $60\ km\ \times\ 60\ km$
515 region for July 06, 2017. We see that the correlation pattern differs significantly based on
516 the how the quartet of rainfall, LAI, clay and elevation vary in the surrounding region
517 of the respective points.
518
519 **4.2.3 Predictions at different Scales**
520 Once the parameters have been estimated, we compute multiscale SM predictions
521 (Appendix A1.2) across CONUS. As a final step, we back-transform the predictions i.e.,
522 $SM = exp(SM')/(1 + exp(SM'))$ to the original scale. We compare our SM predictions at
523 four support scales: point (USCRN and SCAN), 3 km (SMAP/Sentinel-1), 25 km
524 (SMOS) and 36 km (SMAP). We compute five-day SM forecasts from July 21-25, 2017 on
525 all four support scales.
526
527 **4.2.3.1 USCRN and SCAN Scale**
528 As mentioned before, we randomly held out 27 USCRN and SCAN stations across
529 CONUS (Figure 8) as test data. Figure 9 depicts the SM for the "observed"



530

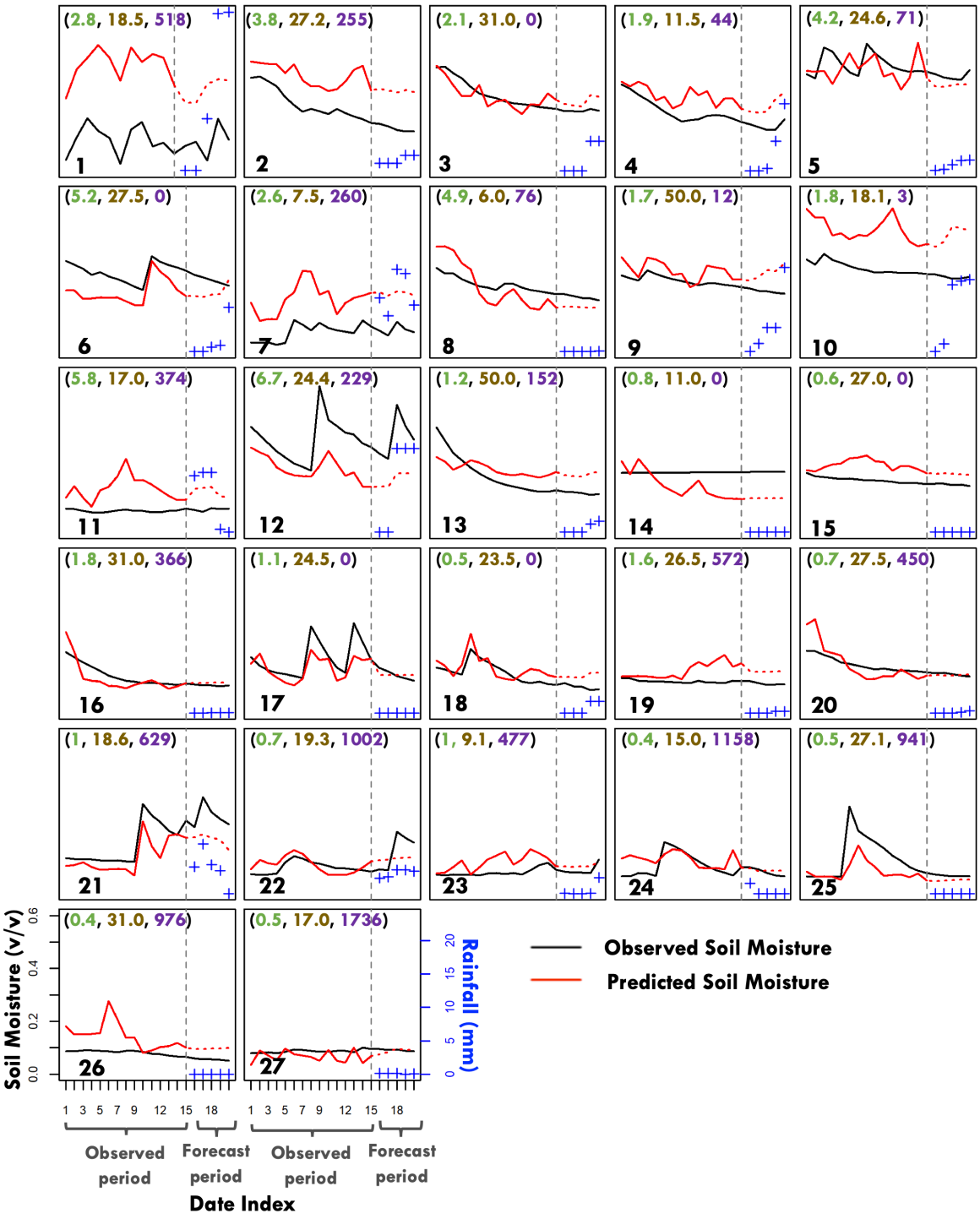
Figure 7. Spatial Correlation pattern of Soil moisture for five points (A-E) across Contiguous US for July 06, 2017. The correlation of the five points with their surrounding region varies considerably due to the covariate heterogeneity of the regions.



531

Figure 8. Location of the validation USCRN/SCAN stations across Contiguous US. We randomly hold out the 27 USCRN/SCAN stations to compare soil moisture predictions at the point scale across Contiguous US. The locations span different hydroclimates and surface heterogeneities.

532
533



534
535

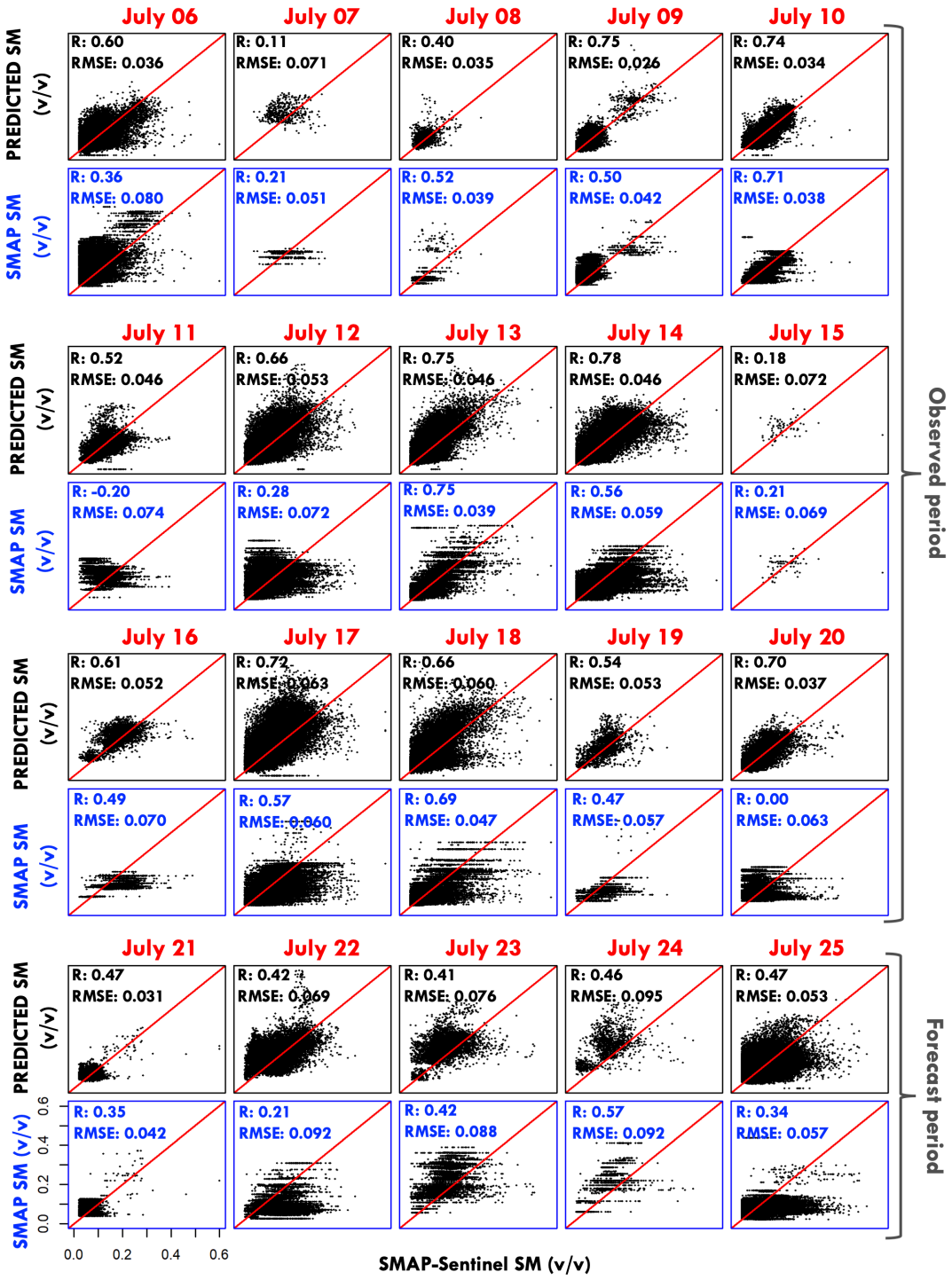
Figure 9. Comparison of soil moisture predictions with the observed SCAN/USRN data for the "observed" (July 06-20, 2017) and "forecast" period (July 21-25, 2017). The covariate values of LAI (averaged during the forecast period), percent clay and elevation (m) are denoted by green, brown and purple colors respectively. The three-day mean antecedent rainfall is also given in blue during the forecast period to demonstrate its effect on SM forecasts.

(July 06-20, 2017) and "forecast" (July 21-25, 2017) period. For the observed period, the correlation (R) and root mean squared error (RMSE) are 0.67 and 0.087 v/v respectively. The slightly high value of the overall RMSE can be attributed to some point station data where there is high bias between the predictions and observation (such as Site 1, 2, 7 and 10) and some stations where the observed SM does not change much during the 20-day period (such as Site 14) possibly resulting from sensor malfunction. Though the SM predictions during the observed period will be mainly influenced by SMAP and SMOS, the predictions serve to fill in important gaps left by these platforms which observe SM at a time interval of 3-7 days.

For the forecast period, R and RMSE of the sites are 0.57 and 0.086 v/v respectively. The forecast period is especially important because it allows us to forecast five-day SM at the point scale in the absence of any observed SM data. We plot the three-day mean antecedent rainfall (from 4km PRISM data) during the forecast period to demonstrate the wetting of SM in response to rainfall. The degree of wetting of SM in our predictions varies not only with rainfall amount but also with the underlying land-surface covariates. Overall, the forecasts for July 21-25, 2017 at point scale are satisfactory given that we utilize only SMAP, SMOS and 116 point station (training) data across CONUS during July 06-20, 2017. Better bias characterization driven by underlying surface heterogeneity for both SMOS and SMAP can help to reduce the bias occurring at some sites.

4.2.3.2 SMAP/Sentinel-1 Scale

The SMAP/Sentinel-1 L2 SM (Das et al., 2018) product uses concurrent 36 km SMAP T_b

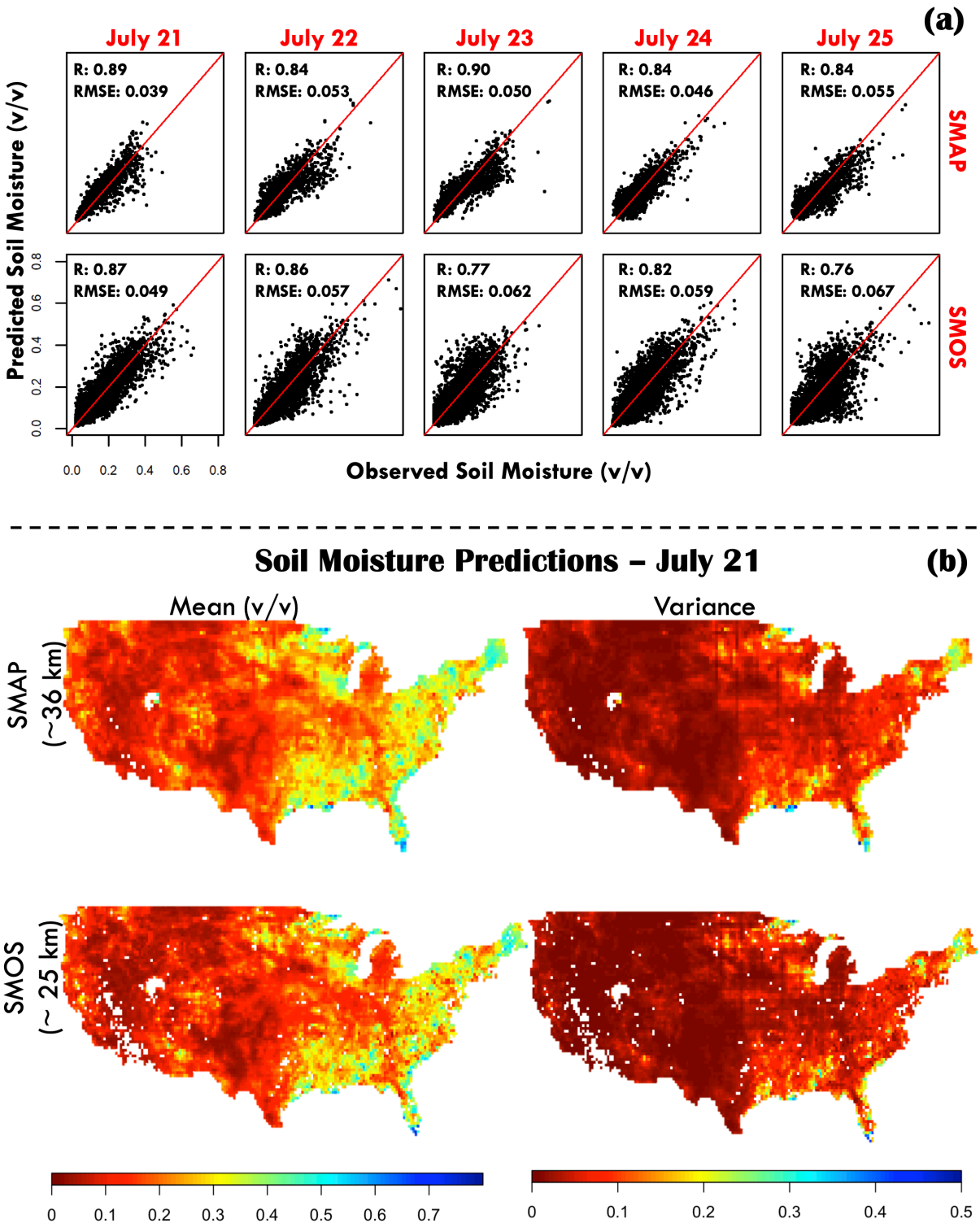


560

Figure 10. Comparison of soil moisture predictions and SMAP soil moisture with the observed SMAP/Sentinel-1 soil moisture at 3 km scale. For the majority of the days, the predicted soil moisture using the fusion approach outperforms the original base SMAP product (even for the forecast period). The red line denotes the 1:1 line.

561 measurements and 3 km backscatter measurements from Sentinel-1 radars to give 3 km
562 SM in the overlapping regions of the two platforms. The Sentinel-1 radars have a much
563 narrower swath (~250 km) however, compared with the relatively wide swath (1,000
564 km) of SMAP which significantly reduces the spatial coverage of the SMAP/Sentinel-1
565 product. The average temporal revisit time of Sentinel-1 radars is 6 days and due to
566 different revisit times of SMAP and Sentinel-1 radars, the temporal resolution of the
567 SMAP/Sentinel-1 SM product varies from 6-12 days. Therefore, for any given day, the
568 coverage of the SMAP/Sentinel-1 product across CONUS is quite limited.
569
570 We compute SM predictions at 3 km (assuming the equidistant grid points $\mathcal{G}$ to be 1 km
571 apart) for the observed SMAP/Sentinel-1 pixels during the 20-day period and compare
572 with the observed SMAP/Sentinel-1 product (Figure 10). We also compare the
573 SMAP/Sentinel-1 observations with the SMAP product from which it is derived. We
574 see that for the majority of the days the SM predictions agree well with the
575 SMAP/Sentinel-1 product outperforming the original SMAP product even for the
576 forecast period. This shows that fusing SMAP SM with SMOS (and USCRN-SCAN data)
577 and accounting for the effects of physical covariates on SM distribution results in better
578 predictive accuracy at 3 km support scale than just using the SMAP SM. Since the
579 spatio-temporal coverage of SMAP/Sentinel-1 is extremely limited, predictions using
580 the data fusion scheme are useful as they help predict SM across the entire CONUS at a
581 daily scale.
582
583 **4.2.3.3 SMAP and SMOS Scale**

584 Since we use all of SMAP and SMOS data for the "observed" period (July 06-20, 2017)
585 for estimating our parameters, we compare SM predictions with observed SMOS and
586 SMAP data for the forecast period (Figure 11 (a)). We make predictions assuming an
587 equidistant numerical grid spaced approximately 9 km apart and remove pixels which
588 have less than 7 grid points lying inside the pixels. We find that the predictions
589 satisfactorily agree with the observed SM with RMSE ranging from 0.039 v/v to 0.055
590 v/v for SMAP, and 0.049 v/v to 0.067 v/v for SMOS while R ranging from 0.84 to 0.90
591 for SMAP, and 0.76 to 0.87 for SMOS. As an illustration, the mean SM predictions as
592 well as the prediction variance for July 21, 2017 are given in Figure 11 (b). It should be
593 noted that since the multiscale predictions are derived from both SMOS and SMAP,
594 their accuracy is affected by how well the two platforms agree with each other. To get a
595 rough estimate of this, we bilinearly interpolated the SMOS pixels which overlap with
596 the SMAP pixels for July 21-25, 2017 and found an RMSE of 0.051 v/v to 0.076 v/v
597 while R varied from 0.74 to 0.86.
598
599 The proposed data fusion scheme thus shows good potential for improving SM
600 predictions across scales. Future research efforts should focus on applying the
601 algorithm for bigger time periods and across different seasons using high performance
602 computing systems. Improved formulations of the mean, bias and covariance functions
603 as well as the inclusion of other physical covariates should be explored. The accuracy of
604 the data fusion scheme at multiple scales can be improved by fusing SM estimates from
605 other platforms such as the Cyclone Global Navigation Satellite System (CYGNSS) and
606 the highly anticipated NASA–ISRO Synthetic Aperture Radar (NISAR) mission. The
607 data fusion allows seamless integration of any number of platforms at varied scales;
608 appropriate parametrization of the bias and error for individual platforms, however, is



609

Figure 11. (a) Comparison of soil moisture predictions and SMAP and SMOS observed soil moisture for July 21-25, 2017. The red line denotes the 1:1 line. (b) Soil moisture predictions across Contiguous US along with the prediction variance. Predictions are unavailable for certain regions due to absence of covariate data.

610 necessary. As mentioned earlier, the proposed algorithm is general and can be
611 potentially used to fuse other spatio-temporally correlated environmental variables
612 which have measurements available from multiple platforms.

613 **5 Conclusions**

614 In this work, we propose a geostatistical framework called *Vecchia-multiscale* for fusing
615 multiscale big data. Using simulated data, we found that certain orderings work better
616 in approximating the exact likelihood at a fraction of the computational cost. We then
617 apply *Vecchia-multiscale* to fuse real SM datasets and compute multiscale SM predictions
618 and forecast five-day SM across scales.

619

620 As the volume of environmental data are expected to dramatically increase in the
621 future, further research into finding better orderings becomes critical. We chose our
622 orderings based only on space and time; future work will focus on proposing
623 physically-based orderings where, in addition to the mean and covariance, the ordering
624 will also be covariate-driven. We applied *Vecchia-multiscale* to simulated data and real
625 SM observations; further application to diverse (spatio-temporally correlated)
626 environmental variables will vet the widespread utility of the algorithm. An advantage
627 of the proposed approach is that it is not a "black-box" and its components can be
628 readily modified based on the underlying physical variable and expert-knowledge.
629 Note that this algorithm can only be applied under a Gaussian Process assumption. In
630 cases where such an assumption is untenable, recent research indicates that the
631 approximation can be further extended using Generalized Gaussian Processes (Zilber &
632 Katzfuss, 2019).

633

We live in an exciting era where a deluge of environmental data presents an unprecedented opportunity for uncovering hidden patterns existing in nature and ultimately achieving the elusive mass and energy balance in Earth-System processes. Data-fusion algorithms harnessing the combined utility of RS and insitu data are critical to advance our understanding of global environmental processes at multiple scales and make data-driven predictions. Moreover, since the breakthrough in numerical modeling occurred when satellite data were assimilated in physical models, fusing multi-platform satellite data can enhance the utility of existing physical models and help take the next leap forward in understanding and predicting environmental processes.

Acknowledgments

We acknowledge the funding support from NASA grants NNX16AQ58G and 80NSSC20K1807, and Texas A&M University. Katzfuss was partially supported by National Science Foundation (NSF) Grants DMS-1654083 and DMS-1953005. Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing. The data used in this study can be accessed from the links provided below: http://bec.icm.csic.es; https://nsidc.org/data/SPL3SMP/versions/6; https://nsidc.org/data/SPL2SMAP_S; http://doi.org/10.5067/MODIS/MCD15A3H.006; https://prism.oregonstate.edu; https://data.nal.usda.gov/dataset/gridded-soil-survey-geographic-database-gssurgo; https://catalog.data.gov/dataset/united-states-climate-reference-network-uscrn-

processed-data-from-the-version-2-uscrn-database;

https://www.wcc.nrcs.usda.gov/scan.

Appendix

A1. Parameter estimation and prediction for *Vecchia-multiscale*

A1.1 Parameter estimation

$$f(z(\mathcal{A})) = f(z(A_1)|\theta) \times \prod_{i=2}^{n} f(z(A_i)|\boldsymbol{z}(\boldsymbol{A_{1:i-1}}), \theta) \quad \text{(A1)}$$

where $\boldsymbol{A_{1:i-1}} = \{A_1, A_2, \ldots . A_{i-1}\}$. If $z(\mathcal{A}) \sim N(\mu_z, \Sigma_z)$, then it can be shown that the $(i,j)^{th}$ element of Λ—the inverse of Cholesky factor of Σ ($\Lambda^T\Lambda = \Sigma^{-1}$) — can be written as

$$\lambda_{ij} = -\frac{w_{ij}}{\sigma^2_{z_i|z_{1:i-1}}} \quad \text{(A2)}$$

where w_{ij} equals $\Sigma^{-1}_{1:i-1}C(\boldsymbol{A_{1:i-1}}, A_i)$ for $j = 1, \ldots, i-1$, equals -1 for $j = i$, and equals 0 for $j > i$. Here $\Sigma_{1:i-1} = C(\boldsymbol{A_{1:i-1}}, \boldsymbol{A_{1:i-1}}) + \tau^2 I_{i-1}$ and $\sigma^2_{z_i|z_{1:i-1}} = C(A_i, A_i) - C(A_i, \boldsymbol{A_{1:i-1}})\Sigma^{-1}_{1:i-1}C(\boldsymbol{A_{1:i-1}}, A_i)$. Here I_{i-1} represents the identity matrix of size $i-1$. We write $C(\boldsymbol{A_{1:i-1}}, A_i)_j \approx (h^{\kappa}_{A_j})^T C\left(\mathcal{G}_{A_j}, \mathcal{G}_{A_i}\right) h^{\kappa}_{A_i}$ and $C(\boldsymbol{A_{1:i-1}}, \boldsymbol{A_{1:i-1}})_{jk} \approx (h^{\kappa}_{A_j})^T C\left(\mathcal{G}_{A_j}, \mathcal{G}_{A_k}\right) h^{\kappa}_{A_k}$ for $j, k = 1, \ldots, i-1$ where $\mathcal{G}_{A_l}$ denotes the subset of the total grid points $\mathcal{G}$ lying inside the pixel A_l and $h^{\kappa}_{A_l}$ is given by equation 7.

We replace $\boldsymbol{A_{1:i-1}}$ with its subset $\boldsymbol{A_{m_i}}$ of maximum length m as defined in Section 3.2. This approximation leads to a sparse $\hat{\Lambda}$ because now $w_{ij} = 0$ for $j = 1, \ldots, i-1$ if $j \notin m_i$, leading to fast computation and low storage for $m \ll n$.

675

676 If $\mu(A_i) = X(A_i)^T \beta$ where $X(A_i) = \{X^1(A_i), \ldots, X^p(A_i)\}$ is a vector of covariates of length
677 p for pixel A_i. Then $\mu_{A_i} = X_{A_i}\beta$ in equation 9 where X_{A_i} is the matrix of covariates
678 associated with the points associated with y_{A_i} in equation 7. Then μ_z (equation 8) can be
679 written as $\tilde{X}\tilde{\beta}$ where the i^{th} row of $\tilde{X}$ is given as $\{h_{A_i}^{\kappa} X_{A_i}^1, \ldots, h_{A_i}^{\kappa} X_{A_i}^p, \delta(A_i)\}$. The parameter
680 vector $\tilde{\beta}$ can be profiled out by using the profile-likelihood:

681 $$-2log(f(z(\mathcal{A})|\theta) = -2log(det(\hat{\Lambda})) + (\hat{\Lambda}(z - \tilde{X}\tilde{\beta}))^T \hat{\Lambda}(z - \tilde{X}\tilde{\beta}) + nlog(2\pi) \quad \text{(A3)}$$

682

683 The maximum likelihood estimate for $\tilde{\beta}$ is given as (Guinness, 2018; Stein et al., 2004):

684 $$\tilde{\beta}_{MLE} = [(\hat{\Lambda}\tilde{X})^T(\hat{\Lambda}\tilde{X})]^{-1}(\hat{\Lambda}\tilde{X})^T(\hat{\Lambda}z) \quad \text{(A4)}$$

685

686 **A1.2 Prediction Algorithm**

687 We follow the prediction algorithm from Guinness (2018). Let $\mathcal{A}^{pred}$ denote a vector of
688 length n^{pred} comprising pixels where we want to want to make predictions y^{pred}. Form
689 the vector $\mathcal{A}^{comp} = (\mathcal{A}, \mathcal{A}^{pred})$ of length $n + n^{pred} = n^{comp}$. The corresponding
690 observation-prediction vector is $y^{comp} = (z, y^{pred})$. Let the covariance matrix of y^{comp}
691 be Σ^{comp}. Writing $\Sigma^{comp}(\Lambda^{comp})$ as a 2×2 block matrix $\{\Sigma_{ij}^{comp}\}_{i,j=1,2}$ ($\{\Lambda_{ij}^{comp}\}_{i,j=1,2}$) and
692 using standard rules of multivariate normality:

693 $$E[y^{pred}|z] = X^{pred}\beta + \Sigma_{21}^{comp}(\Sigma_{11}^{comp})^{-1}(z - \tilde{X}\tilde{\beta})$$

694 $$= -(\Lambda_{22}^{comp})^{-1}\Lambda_{21}^{comp} z \approx -(\hat{\Lambda}_{22}^{comp})^{-1}\hat{\Lambda}_{21}^{comp}(z - \tilde{X}\tilde{\beta}), \quad \text{(A5)}$$

695 where $\hat{\Lambda}^{comp}$ is the sparse approximation of Λ^{comp} calculated following A1.1.

To find the prediction variance $Var(y^{pred}|z)$, we first simulate uncorrelated standard normals of length n^{comp}; $w^* \sim \mathcal{N}(0, I_{n^{comp}})$ where $I_{n^{comp}}$ is the identity matrix of size n^{comp}. We then simulate $y^{comp*} = \{z^*, y^{pred*}\} = (\hat{\Lambda}^{comp})^{-1}w$ which is computationally fast since $\hat{\Lambda}^{comp}$ is a sparse triangular matrix. Then, $-\hat{\Lambda}_{22}^{-1}\hat{\Lambda}_{21}(z - z^*) + y^{pred*}$ approximately has a covariance matrix $\Sigma_{22}^{comp} - \Sigma_{21}^{comp}(\Sigma_{11}^{comp})^{-1}\Sigma_{12}^{comp}$, which is equal to $Var(y^{pred}|z)$ based on the well-known properties of multivariate normality. We simulate $-\hat{\Lambda}_{22}^{-1}\hat{\Lambda}_{21}(z - z^*) + y^{pred*}$ five thousand times to approximate the prediction variance.

References

Akbar, R., Short Gianotti, D. J., Salvucci, G. D., & Entekhabi, D. (2019). Mapped Hydroclimatology of Evapotranspiration and Drainage Runoff Using SMAP Brightness Temperature Observations and Precipitation Information. *Water Resources Research*, *55*(4), 3391–3413. https://doi.org/10.1029/2018WR024459

Barré, H. M. J. P., Duesmann, B., & Kerr, Y. H. (2008). SMOS: The mission and the system. *IEEE Transactions on Geoscience and Remote Sensing*, *46*(3), 587–593. https://doi.org/10.1109/TGRS.2008.916264

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. https://doi.org/10.1038/nature14956

Camps-valls, G., Tuia, D., & Bruzzone, L. (2013). Advances in hyperspectral image classification. *IEEE Signal Processing Magazine*, *January*, 45–54.

Cecinati, F., Rico-Ramirez, M. A., Heuvelink, G. B. M., & Han, D. (2017). Representing radar rainfall uncertainty with ensembles based on a time-variant geostatistical error modelling approach. *Journal of Hydrology*, *548*, 391–405. https://doi.org/10.1016/j.jhydrol.2017.02.053

Colliander, A., Jackson, T. J., Bindlish, R., Chan, S., Das, N., Kim, S. B., Cosh, M. H., Dunbar, R. S., Dang, L., Pashaian, L., Asanuma, J., Aida, K., Berg, A., Rowlandson, T., Bosch, D., Caldwell, T., Caylor, K., Goodrich, D., al Jassar, H., … Yueh, S. (2017). Validation of SMAP surface soil moisture products with core validation sites. *Remote Sensing of Environment*, *191*, 215–231. https://doi.org/10.1016/j.rse.2017.01.021

Cosh, M. H., & Brutsaert, W. (1999). Aspects of soil moisture variability in the Washita '92 study region. *Journal of Geophysical Research: Atmospheres*, *104*(D16), 19751–19757. https://doi.org/https://doi.org/10.1029/1999JD900110

Cressie, N. (1990). The origins of kriging. *Mathematical Geology*, *22*(3), 239–252. https://doi.org/10.1007/BF00889887

Crow, W. T., Berg, A. A., Cosh, M. H., Loew, A., Mohanty, B. P., Panciera, R., de Rosnay, P., Ryu, D., & Walker, J. P. (2012). Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products. *Reviews of Geophysics*, *50*(2). https://doi.org/https://doi.org/10.1029/2011RG000372

Daly, C., Neilson, R. P., & Phillips, D. L. (1994). A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *Journal of Applied Meteorology*.

https://doi.org/10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2
Das, N. N., Entekhabi, D., Dunbar, R. S., Kim, S., Yueh, S., Colliander, A., O'Neill, P. E., & Jackson, T. (2018). SMAP/Sentinel-1 L2 radiometer/radar 30-second scene 3 km EASE-grid soil moisture, version 2. *NASA National Snow and Ice Data Center DAAC*.
Datta, A., Banerjee, S., Finley, A. O., & Gelfand, A. E. (2016). Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2015.1044091
Diamond, H. J., Karl, T. R., Palecki, M. A., Baker, C. B., Bell, J. E., Leeper, R. D., Easterling, D. R., Lawrimore, J. H., Meyers, T. P., Helfert, M. R., Goodge, G., & Thorne, P. W. (2013). U.S. climate reference network after one decade of operations status and assessment. *Bulletin of the American Meteorological Society*, *94*(4), 485–498. https://doi.org/10.1175/BAMS-D-12-00170.1
Entekhabi, D., Njoku, E. G., O'Neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., Entin, J. K., Goodman, S. D., Jackson, T. J., Johnson, J., Kimball, J., Piepmeier, J. R., Koster, R. D., Martin, N., McDonald, K. C., Moghaddam, M., Moran, S., Reichle, R., Shi, J. C., … Zyl, J. Van. (2010). The Soil Moisture Active Passive (SMAP) Mission. *Proceedings of the IEEE*, *98*(5), 704–716. https://doi.org/10.1109/JPROC.2010.2043918
Entin, J. K., Robock, A., Vinnikov, K. Y., Hollinger, S. E., Liu, S., & Namkhai, A. (2000). Temporal and spatial scales of observed soil moisture variations in the extratropics. *Journal of Geophysical Research: Atmospheres*, *105*(D9), 11865–11877. https://doi.org/https://doi.org/10.1029/2000JD900051
Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, *99*(C5). https://doi.org/10.1029/94jc00572
Fan, X., Liu, Y., Gan, G., & Wu, G. (2020). SMAP underestimates soil moisture in vegetation-disturbed areas primarily as a result of biased surface temperature data. *Remote Sensing of Environment*, *247*(November 2019), 111914. https://doi.org/10.1016/j.rse.2020.111914
Fang, K., Shen, C., Kifer, D., & Yang, X. (2017). Prolongation of SMAP to Spatiotemporally Seamless Coverage of Continental U.S. Using a Deep Learning Neural Network. *Geophysical Research Letters*, *44*(21), 11,030-11,039. https://doi.org/10.1002/2017GL075619
Gaur, N., & Mohanty, B. P. (2013). Evolution of physical controls for soil moisture in humid and subhumid watersheds. *Water Resources Research*. https://doi.org/10.1002/wrcr.20069
Gaur, N., & Mohanty, B. P. (2016). Land-surface controls on near-surface soil moisture dynamics: Traversing remote sensing footprints. *Water Resources Research*. https://doi.org/10.1002/2015WR018095
Gelfand, A. E. (2001). On the change of support problem for spatio-temporal data. *Biostatistics*, *2*(1), 31–45. https://doi.org/10.1093/biostatistics/2.1.31
Gelfand, Alan E., Diggle, P. J., Fuentes, M., & Guttorp, P. (2010). *Handbook of Spatial Statistics*. CRC Press.
Ghil, M., & Malanotte-Rizzoli, P. (1991). Data Assimilation in Meteorology and Oceanography. *Advances in Geophysics*. https://doi.org/10.1016/S0065-2687(08)60442-2
Girotto, M., De Lannoy, G. J. M., Reichle, R. H., Rodell, M., Draper, C., Bhanja, S. N., & Mukherjee, A. (2017). Benefits and pitfalls of GRACE data assimilation: A case study of terrestrial water storage depletion in India. *Geophysical Research Letters*, *44*(9), 4107–4115. https://doi.org/10.1002/2017GL072994
Goovaerts, P., AvRuskin, G., Meliker, J., Slotnick, M., Jacquez, G., & Nriagu, J. (2005). Geostatistical modeling of the spatial variability of arsenic in groundwater of southeast Michigan. *Water Resources Research*, *41*(7), 1–19. https://doi.org/10.1029/2004WR003705
Guinness, J. (2018). Permutation and Grouping Methods for Sharpening Gaussian Process Approximations. *Technometrics*, *60*(4), 415–429. https://doi.org/10.1080/00401706.2018.1437476
Hengl, T., De Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0169748
Hu, G., Jia, L., & Menenti, M. (2015). Comparison of MOD16 and LSA-SAF MSG evapotranspiration products over Europe for 2011. *Remote Sensing of Environment*, *156*, 510–526. https://doi.org/10.1016/j.rse.2014.10.017
Joshi, C., & Mohanty, B. P. (2010). Physical controls of near-surface soil moisture across varying spatial scales in an agricultural landscape during SMEX02. *Water Resources Research*, *46*(12), 1–21. https://doi.org/10.1029/2010WR009152

Joshi, C., Mohanty, B. P., Jacobs, J. M., & Ines, A. V. M. (2011). Spatiotemporal analyses of soil moisture from point to footprint scale in two different hydroclimatic regions. *Water Resources Research*, *47*(1). https://doi.org/https://doi.org/10.1029/2009WR009002

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S. I., Sheffield, J., Goulden, M. L., Bonan, G., Cescatti, A., Chen, J., De Jeu, R., Dolman, A. J., Eugster, W., Gerten, D., Gianelle, D., Gobron, N., Heinke, J., Kimball, J., Law, B. E., Montagnani, L., … Zhang, K. (2010). Recent decline in the global land evapotranspiration trend due to limited moisture supply. *Nature*, *467*(7318), 951–954. https://doi.org/10.1038/nature09396

Kathuria, D., Mohanty, B. P., & Katzfuss, M. (2019a). A Nonstationary Geostatistical Framework for Soil Moisture Prediction in the Presence of Surface Heterogeneity. *Water Resources Research*, *55*(1). https://doi.org/10.1029/2018WR023505

Kathuria, D., Mohanty, B. P., & Katzfuss, M. (2019b). Multiscale Data Fusion for Surface Soil Moisture Estimation: A Spatial Hierarchical Approach. *Water Resources Research*, *55*(12). https://doi.org/10.1029/2018WR024581

Katzfuss, M., & Guinness, J. (2017). A general framework for Vecchia approximations of Gaussian processes. In *arXiv*. https://doi.org/10.1214/19-sts755

Katzfuss, M., Guinness, J., Gong, W., & Zilber, D. (2020). Vecchia Approximations of Gaussian-Process Predictions. *Journal of Agricultural, Biological, and Environmental Statistics*. https://doi.org/10.1007/s13253-020-00401-7

Kaufman, C. G., Schervish, M. J., & Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*. https://doi.org/10.1198/016214508000000959

Klees, R., Zapreeva, E. A., Winsemius, H. C., & Savenije, H. H. G. (2007). The bias in GRACE estimates of continental water storage variations. *Hydrology and Earth System Sciences*, *11*(4), 1227–1241. https://doi.org/10.5194/hess-11-1227-2007

Koster, Randal D., Brocca, L., Crow, W. T., Burgin, M. S., & De Lannoy, G. J. M. (2016). Precipitation estimation using L-band and C-band soil moisture retrievals. *Water Resources Research*. https://doi.org/10.1002/2016WR019024

Koster, Randal D., Crow, W. T., Reichle, R. H., & Mahanama, S. P. (2018). Estimating Basin-Scale Water Budgets With SMAP Soil Moisture Data. *Water Resources Research*, *54*(7), 4228–4244. https://doi.org/10.1029/2018WR022669

Koster, Randall D., Schubert, S. D., & Suarez, M. J. (2009). Analyzing the concurrence of meteorological droughts and warm periods, with implications for the determination of evaporative regime. *Journal of Climate*. https://doi.org/10.1175/2008JCLI2718.1

Krige, D. G. (1952). A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*. https://doi.org/10.2307/3006914

Lanfredi, M., Coppola, R., D'Emilio, M., Imbrenda, V., Macchiato, M., & Simoniello, T. (2015). A geostatistics-assisted approach to the deterministic approximation of climate data. *Environmental Modelling and Software*, *66*, 69–77. https://doi.org/10.1016/j.envsoft.2014.12.009

Lark, R. M. (2012). Towards soil geostatistics. *Spatial Statistics*, *1*, 92–99. https://doi.org/10.1016/j.spasta.2012.02.001

Li, S., Yu, Y., Sun, D., Tarpley, D., Zhan, X., & Chiu, L. (2014). Evaluation of 10 year AQUA/MODIS land surface temperature with SURFRAD observations. *International Journal of Remote Sensing*, *35*(3), 830–856. https://doi.org/10.1080/01431161.2013.873149

Lievens, H., Reichle, R. H., Liu, Q., De Lannoy, G. J. M., Dunbar, R. S., Kim, S. B., Das, N. N., Cosh, M., Walker, J. P., & Wagner, W. (2017). Joint Sentinel-1 and SMAP data assimilation to improve soil moisture estimates. *Geophysical Research Letters*. https://doi.org/10.1002/2017GL073904

Mao, H., Kathuria, D., Duffield, N., & Mohanty, B. P. (2019). Gap Filling of High-Resolution Soil Moisture for SMAP/Sentinel-1: A Two-Layer Machine Learning-Based Framework. *Water Resources Research*. https://doi.org/10.1029/2019WR024902

Mohanty, B. P., Ankeny, M. D., Horton, R., & Kanwar, R. S. (1994). Spatial analysis of hydraulic conductivity measured using disc infiltrometers. *Water Resources Research*, *30*(9), 2489–2498. https://doi.org/https://doi.org/10.1029/94WR01052

Mohanty, B. P., Famiglietti, J. S., & Skaggs, T. H. (2000). Evolution of soil moisture spatial structure in a mixed vegetation pixel during the Southern Great Plains 1997 (SGP97) Hydrology Experiment. *Water Resources Research*, *36*(12), 3675–3686. https://doi.org/https://doi.org/10.1029/2000WR900258

Mohanty, B. P., & Kanwar, R. S. (1994). Spatial variability of residual nitrate-nitrogen under two tillage systems in central Iowa: A composite three-dimensional resistant and exploratory approach. *Water Resources Research*, *30*(2), 237–251. https://doi.org/https://doi.org/10.1029/93WR02922

Mohanty, B. P., Kanwar, R. S., & Horton, R. (1991). A Robust-Resistant Approach to Interpret Spatial Behavior of Saturated Hydraulic Conductivity of a Glacial Till Soil Under No-Tillage System. *Water Resources Research*, *27*(11), 2979–2992. https://doi.org/https://doi.org/10.1029/91WR01720

Myneni, R., Knyazikhin, Y., Park, T. (2015). *MCD15A3H MODIS/Terra+Aqua Leaf Area Index/FPAR 4-day L4 Global 500m SIN Grid V006*. NASA EOSDIS Land Processes DAAC.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., & Sain, S. (2015). A Multiresolution Gaussian Process Model for the Analysis of Large Spatial Datasets. *Journal of Computational and Graphical Statistics*. https://doi.org/10.1080/10618600.2014.914946

O'Neill, P. E., Njoku, E. G., Jackson, T. J., Chan, S., & Bindlish, R. (2018). SMAP Algorithm Theoretical Basis Document: Level 2 & 3 Soil Moisture (Passive) Data Products. *Revision D*.

Pablos, M., Vall-llossera, M., Piles, M., Camps, A., Gonzalez-Haro, C., Turiel, A., Herbert, C. J., Chaparro, D., & Portal, G. (2019). *Influence of Quality Filtering Approaches in BEC SMOS L3 Soil Moisture Products*. https://doi.org/10.1109/igarss.2019.8900273

Reich, B. J., Eidsvik, J., Guindani, M., Nail, A. J., & Schmidt, A. M. (2011). A class of covariate-dependent spatiotemporal covariance functions for the analysis of daily ozone concentration. *Annals of Applied Statistics*. https://doi.org/10.1214/11-AOAS482

Reichle, R. H., & Koster, R. D. (2004). Bias reduction in short records of satellite soil moisture. *Geophysical Research Letters*, *31*(19), 2–5. https://doi.org/10.1029/2004GL020938

Reichle, Rolf H., McLaughlin, D. B., & Entekhabi, D. (2002). Hydrologic data assimilation with the ensemble Kalman filter. *Monthly Weather Review*, *130*(1), 103–114. https://doi.org/10.1175/1520-0493(2002)130<0103:HDAWTE>2.0.CO;2

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Risser, M. D., & Calder, C. A. (2015). Regression-based covariance functions for nonstationary spatial modeling. *Environmetrics*. https://doi.org/10.1002/env.2336

Ryu, D., & Famiglietti, J. S. (2006). Multi-scale spatial correlation and scaling behavior of surface soil moisture. *Geophysical Research Letters*, *33*(8). https://doi.org/https://doi.org/10.1029/2006GL025831

Schaefer, G. L., Cosh, M. H., & Jackson, T. J. (2007). The USDA Natural Resources Conservation Service Soil Climate Analysis Network (SCAN). *Journal of Atmospheric and Oceanic Technology*, *24*(12), 2073–2077. https://doi.org/10.1175/2007JTECHA930.1

Shen, C. (2018). A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. *Water Resources Research*, *54*(11), 8558–8593. https://doi.org/10.1029/2018WR022643

Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2017). Deep learning for precipitation nowcasting: A benchmark and a new model. *Advances in Neural Information Processing Systems*.

Soil Survey Staff. (2020). *Gridded Soil Survey Geographic (gSSURGO) Database for the Conterminous United States*. United States Department of Agriculture, Natural Resources Conservation Service. https://gdg.sc.egov.usda.gov/

Stein, M. L., Chi, Z., & Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*. https://doi.org/10.1046/j.1369-7412.2003.05512.x

Sun, Y., & Stein, M. L. (2016). Statistically and Computationally Efficient Estimating Equations for Large Spatial Datasets. *Journal of Computational and Graphical Statistics*. https://doi.org/10.1080/10618600.2014.975230

Teuling, A. J., & Troch, P. A. (2005). Improved understanding of soil moisture variability dynamics. *Geophysical Research Letters*, *32*(5). https://doi.org/https://doi.org/10.1029/2004GL021935

Vecchia, A. V. (1988). Estimation and Model Identification for Continuous Spatial Processes. *Journal of the Royal Statistical Society: Series B (Methodological)*. https://doi.org/10.1111/j.2517-6161.1988.tb01729.x

Velpuri, N. M., Senay, G. B., Singh, R. K., Bohms, S., & Verdin, J. P. (2013). A comprehensive evaluation of two MODIS evapotranspiration products over the conterminous United States: Using point and gridded FLUXNET and water balance ET. *Remote Sensing of Environment*, *139*, 35–49. https://doi.org/10.1016/j.rse.2013.07.013

Vereecken, H., Huisman, J. A., Pachepsky, Y., Montzka, C., van der Kruk, J., Bogena, H., Weihermüller, L., Herbst, M., Martinez, G., & Vanderborght, J. (2014). On the spatio-temporal dynamics of soil moisture at the field scale. *Journal of Hydrology*, *516*, 76–96. https://doi.org/https://doi.org/10.1016/j.jhydrol.2013.11.061

Wang, B., Zou, X., & Zhu, J. (2000). Data assimilation and its applications. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(21), 11143–11144. https://doi.org/10.1073/pnas.97.21.11143

Westermann, S., Langer, M., & Boike, J. (2012). Systematic bias of average winter-time land surface temperatures inferred from MODIS at a site on Svalbard, Norway. *Remote Sensing of Environment*, *118*, 162–167. https://doi.org/10.1016/j.rse.2011.10.025

Wikle, C. K., Zammit-Mangion, A., & Cressie, N. (2019). *Spatio-Temporal Statistics with R*. CRC Press.

Xiang, Y., Gubian, S., Suomela, B., & Hoeng, J. (2013). Generalized simulated annealing for global optimization: The GenSA package. *R Journal*. https://doi.org/10.32614/rj-2013-002

Zhong, Z., & Carr, T. R. (2019). Geostatistical 3D geological model construction to estimate the capacity of commercial scale injection and storage of CO2 in Jacksonburg-Stringtown oil field, West Virginia, USA. *International Journal of Greenhouse Gas Control*, *80*(March 2018), 61–75. https://doi.org/10.1016/j.ijggc.2018.10.011

Zilber, D., & Katzfuss, M. (2019). *Vecchia-Laplace approximations of generalized Gaussian processes for big non-Gaussian spatial data*.

Supporting Information File for

A Multiscale Spatio-Temporal Big Data Fusion Algorithm from Point to Satellite Footprint Scales

Dhruva Kathuria[1], Binayak P. Mohanty[1], and Matthias Katzfuss[2]

[1]Biological and Agricultural Engineering, Texas A&M University, College Station, Texas, USA

[2]Department of Statistics, Texas A&M University, College Station, Texas, USA.

S1. Illustration of different permutations for *Vecchia-Multiscale*

We illustrate the effect of different permutations (Figure S1 and S2) by applying the eight permutations to the hypothetical example in Figure 2 (a) comprising three datasets: areal datasets R_1 (64 green pixels) and R_2 (36 purple pixels), and point dataset P_1 (40 blue triangles), making the total number of observations $n = 140$. The numbers in columns (I) to (III) in Figure (S1) represent the ordering number in $\mathcal{A} = \{A_1, \ldots, A_{140}\}$ assigned to individual data in P_1 (I), R_1 (II) and R_2 (III) for the different permutations. Column (IV) denotes the subvector $\boldsymbol{A_{m_i}}$ (color-filled blue triangles, and color-filled green and purple pixels) for a randomly chosen pixel A_i (color-filled red) for $m = 20$.

The *Joint-Coordinate* permutation (Figure S1 (a)-(c)) sorts the data based on the sum of coordinate values resulting in the data from the three platforms getting ordered from the lower-left to the upper right along the diagonal. For any pixel A_i, this results in $\boldsymbol{A_{1:i-1}}$ located close to A_i. The subvector $\boldsymbol{A_{m_i}}$ (selected from elements of $\boldsymbol{A_{1:i-1}}$ closest to A_i in space) is thus located in the immediate neighborhood of A_i (Figure S1 (d)). *Middleout* ordering is based on the same heuristic as *Coordinate* ordering and orders the locations based on increasing distance from the mean location of the study domain (Guinness, 2018). Thus, it also has $\boldsymbol{A_{m_i}}$ located in the neighborhood of A_i (Figure S1 (h)).

The *Joint-Maxmin* ordering (Figure S1 (i)-(l)) selects the first pixel/point which is closest to the mean location of the study domain and then sequentially selects a successive pixel/point which maximizes the "minimum distance" to previously selected pixels/points (Guinness, 2018). This results in the pixels/points getting permuted such that for any A_i, $\boldsymbol{A_{1:i-1}}$ now consist of a good mix of both far and near pixels/points (Figure S1 (i)-(k)). The subvector $\boldsymbol{A_{m_i}}$ now consist of both far and near data surrounding

A_i (Figure S1 (l)). Though *Joint-Random* (Figure S1 (m)-(p)) is not based on any heuristic, it can give similar results to *Joint-Maxmin* (Guinness, 2018).

The corresponding "*Separate-*" orderings for the four "*Joint-*" orderings are given in Figure S2. The "*Separate-*" orderings separate the point and areal data, apply the permutations separately to each and then form the final permutation by sorting the permuted point data followed by the permuted areal data (Figure 4, main text). Though the "*Separate-*" orderings retain the heuristic of the corresponding "*Joint-*" permutations separately for point and areal data, the "*Separate-*" permutations introduce a constraint that the point data always lie in the beginning of the vector $\mathcal{A}$. For instance, in Figure S2 (Column I) since we have 40 point data, $\{A_1, \dots, A_{40}\}$ always represent point data in "Separate-" permutations. Now for any areal pixel A_i (which for "Separate-" permutations in this example represent $\{A_{41}, \dots, A_{140}\}$), $\boldsymbol{A_{1:i-1}}$ will always consist of point data. This often leads to the subvector $\boldsymbol{A_{m_i}}$ consist of point data which are near to A_i (Figure S2, Column IV).

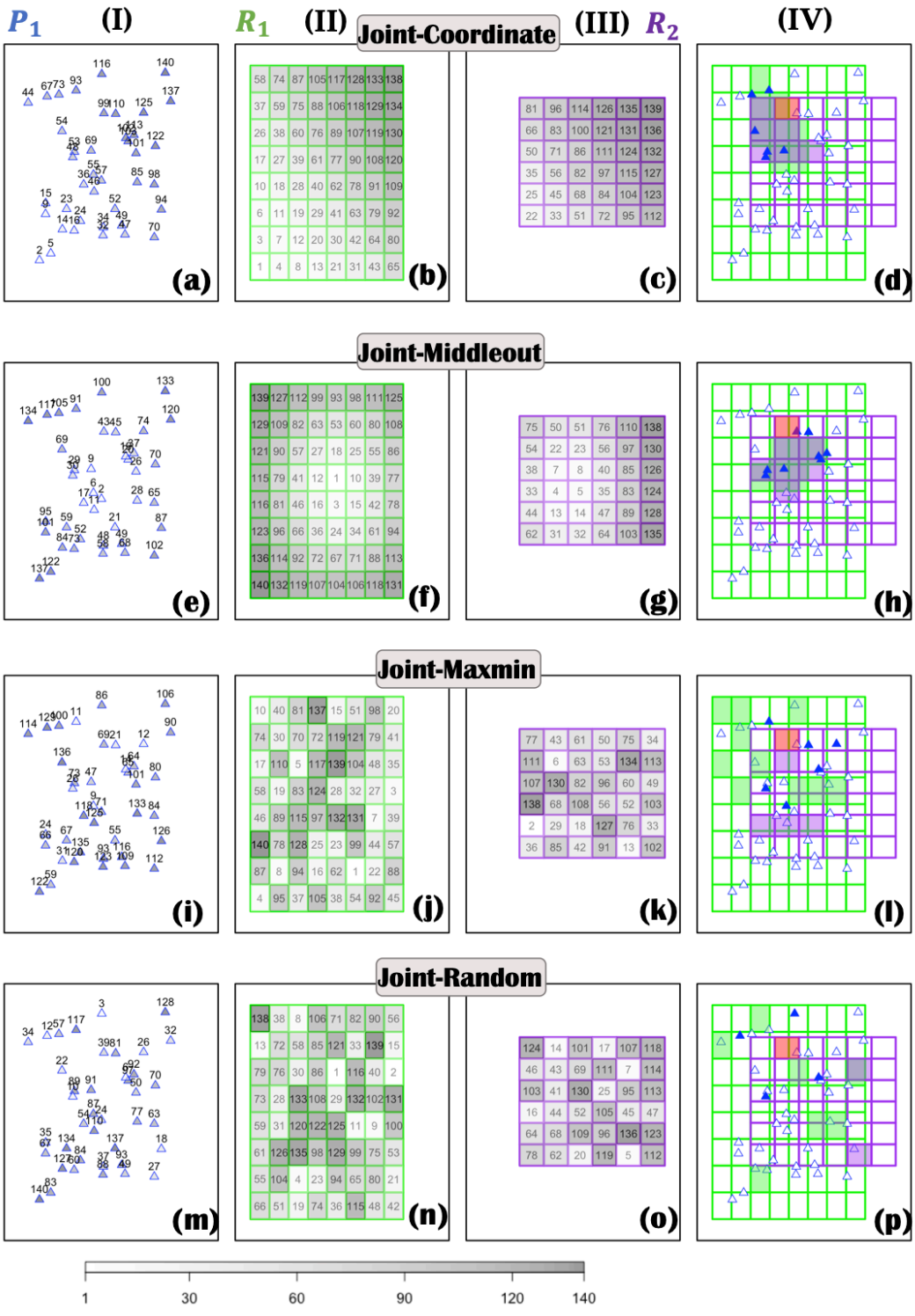


Figure S1. Illustration of the "Joint-" Permutations applied on the example from Figure 2 (a) in the main text consisting of 40 point data P_1 and 100 areal pixels in R_1 (64 pixels) and R_2 (36 pixels) . Numbers in columns (I) to (III) represent the ordering number in the vector $\mathcal{A} = \{A_1, \dots, A_{140}\}$ assigned to data in P_1 (I), R_1 (II) and R_2 (III) for the four different "Joint-" permutations. Column (d) denotes the subvector $\boldsymbol{A}_{\boldsymbol{m}_i}$ (equation 11, main text) comprising color-filled blue triangles, and color-filled green and purple pixels, for a randomly chosen pixel A_i (color-filled red) for $m = 20$.

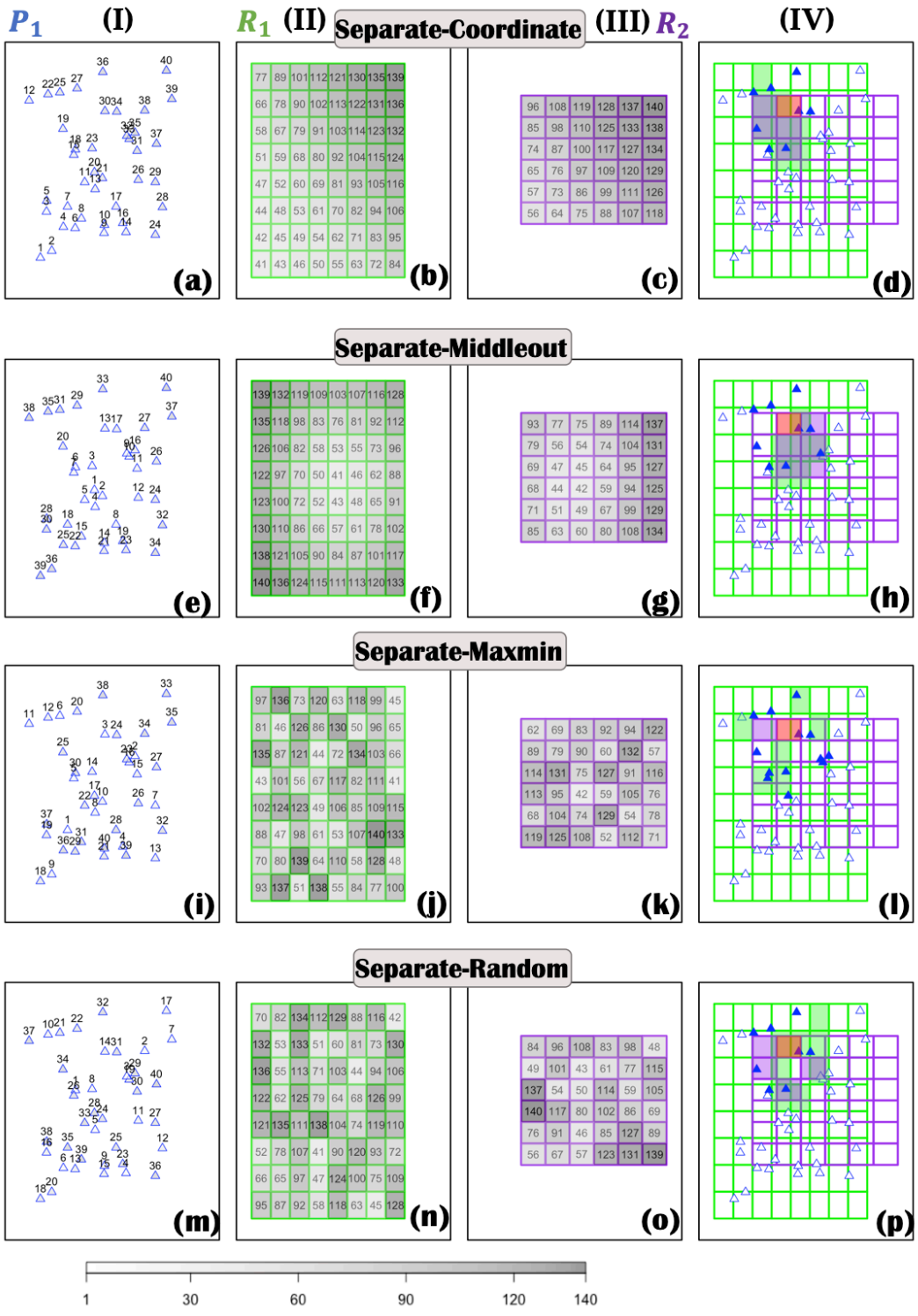


Figure S2. Illustration of the "Separate-" Permutations applied on the example from Figure 2 (a) in the main text consisting of 40 point data P_1 and 100 areal pixels in R_1 (64 pixels) and R_2 (36 pixels) . Numbers in columns (I) to (III) represent the ordering number in the vector $\mathcal{A} = \{A_1, \dots, A_{140}\}$ assigned to data in P_1 (I), R_1 (II) and R_2 (III) for the four different "Separate-" permutations. Column (d) denotes the subvector $\boldsymbol{A}_{\boldsymbol{m_i}}$ (equation 11, main text) comprising color-filled blue triangles, and color-filled green and purple pixels, for a randomly chosen pixel A_i (color-filled red) for $m = 20$.

S2. Simulation

We use simulations for two (e.g, a variable varying across latitude and longitude) and three (e.g., a variable varying across latitude, longitude and time) dimensions in space in a region $\mathcal{D} = [0,1] \times [0,1]$ and $[0,1] \times [0,1] \times [0,1]$ respectively. We fix each dimension between 0 and 1 for generality. The objective of the simulations is to investigate that for a given value of m, which approximation (equation 11) resulting out of the eight permutations better approximates the exact likelihood (equation 10). Similar to the hypothetical example in Figure 2 (a) in the main text, we assume three data sources for each setting—two aggregate datasets (R_1 and R_2) covering the entire region $\mathcal{D}$, and point dataset (P_1) in $\mathcal{D}$. The number of pixels in R_1 and R_2 along with their resolutions as well as the number of point data P_1 are given in Table S1. The number of point data are chosen as 1) 5% of the areal data to represent scenarios where the point data is sparse compared to areal data, and 2) 25% of the areal data to represent scenarios where point data are considerable in number compared to areal data. We assume an equidistant numerical grid $\mathcal{G}$ consisting of 11000 points for two dimensions and $1089 \times 11 = 11979$ points for three dimensions across $\mathcal{D}$.

As mentioned in the main text, evaluation of the exact likelihood requires quadratic complexity in the number of assumed grid points $n_{\mathcal{G}}$ and cubic complexity in the number of observations n. Therefore for the simulations, the number of observations of each platform and the size of the numerical grid are chosen so that the computation of actual likelihood $f(z(\mathcal{A})|\theta)$ is feasible.

We use a flexible class of covariance function called the Matern, with a range, smoothness and variance parameter, for simulating the covariance matrix. Other widely used covariance functions such as the Exponential and the Gaussian are special cases of the Matern. We do simulations for range $= \{0.2, 0.4, 0.6\}$, smoothness (nu) $= \{0.5, 1, 1.5\}$, variance $= 1$ and measurement error variance (in R_1 and R_2) $= \{0.05, 0.2\}$. This ensures that the simulations are carried out for a wide range of parameters resulting in a total of 72 simulations for each ordering. We perform 72 simulations for each of the eight orderings and take m = 5, 10, 20, 40, 60, 100, 120 and 180.

To control for simulation error, we use the Kullback-Leibler (KL) divergence, which measures how much information we lose using the approximation $\hat{f}(z(\mathcal{A})|\theta)$ (equation 11, main text) over the exact likelihood $f(z(\mathcal{A})|\theta)$ (equation 10, main text), both using the true value of the parameters. A lower KL-divergence between $\hat{f}(z(\mathcal{A})|\theta)$ and $f(z(\mathcal{A})|\theta)$ thus denotes a better approximation. Plots of eight representative simulations (out of 72) comparing the (log) KL-Divergence of the approximations over the true likelihood are given in Figure S3. For both 2D and 3D, in general, the *Separate-Maxmin* and *Separate-Random* perform the best while the *Coordinate-based* orderings perform the worst. There was no effect of measurement error on the relative performance of the orderings. Therefore, in general, we suggest adopting *Separate-Maxmin* or *Separate-Random* when using *Vecchia-multiscale*.

Table S1. Data setting for the simulations in Section S2.

Data	Resolution	Number of pixels/points	Grid points per pixel
	Two Dimensions		
R_1	0.09	$34 \times 34 = 1156$	9
R_2	0.06	$52 \times 52 = 2704$	4
P_1	-	$200(\approx 5\%)$ & $1000(\approx 25\%)$	-
Total	-	4060 & 4860	-
	Three Dimensions		
R_1	0.03	$11 \times 11 \times 11 = 1331$	9
R_2	0.02	$16 \times 16 \times 11 = 2816$	4
P_1		$20 \times 11 = 220(\approx 5\%)$ & $100 \times 11 = 1100(\approx 25\%)$	-
Total	-	4367 & 5247	-

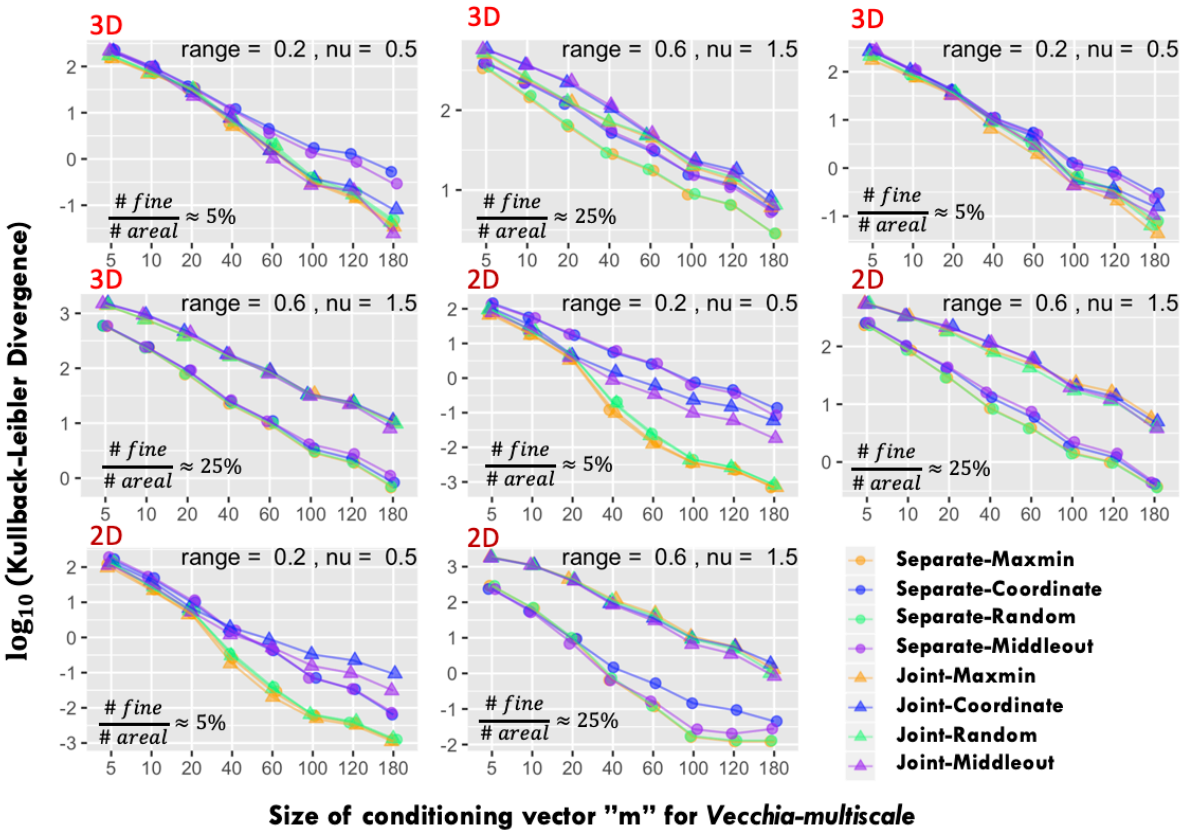


Figure S3 Representative simulations comparing the (log) KL-Divergence of the approximations over the true likelihood for measurement error variance equal to 0.05. A lower KL-Divergence denotes a better approximation. For the majority of the simulation settings, the Separate-Maxmin and the Separate-Random lead to better approximation of the exact likelihood.

S3 Supporting Information for Section 4 in the main text

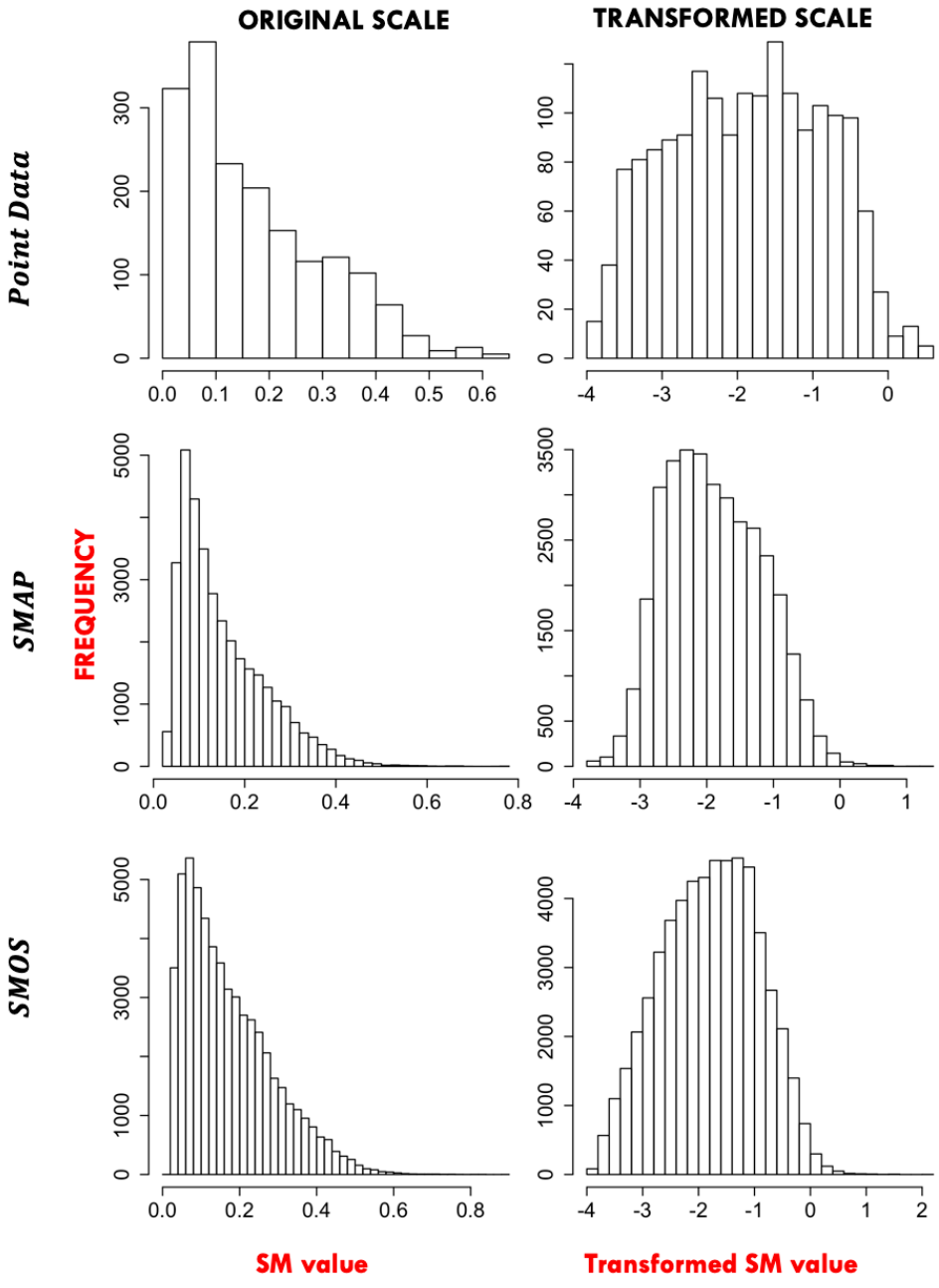


Figure S4 Histograms of point soil, SMAP and SMOS soil moisture data for July 06-20, 2017. On the original scale soil moisture exhibits considerable skewness but on the logit scale the soil moisture distribution becomes less skewed making the Gaussian assumption tenable.

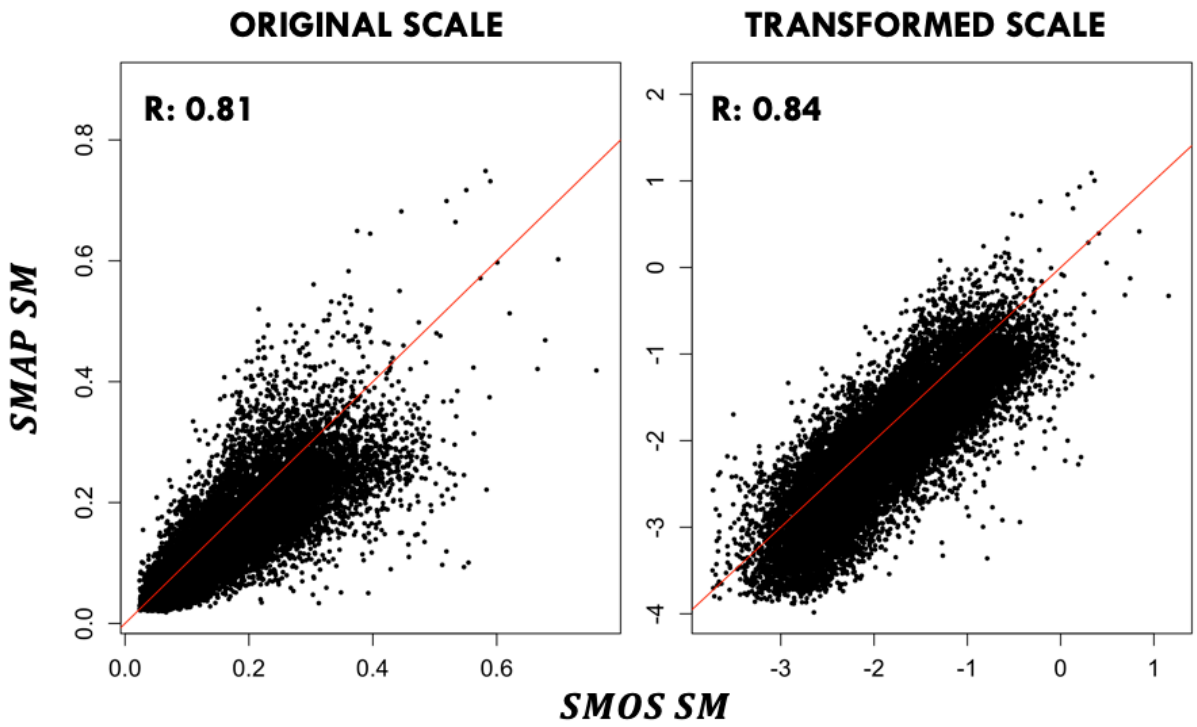


Figure S5 Overlapping SMOS and SMAP pixels for July 06-20, 2017. The SMOS pixels are bilinearly interpolated to the overlapping SMAP pixels for this exploratory analysis. The red line denotes the 1:1 line. The transformed scale results in a slightly better correlation (R) between the two datasets. On the transformed scale, it can also be seen that there is a bias between SMOS and SMAP datasets for the analyzed time period.

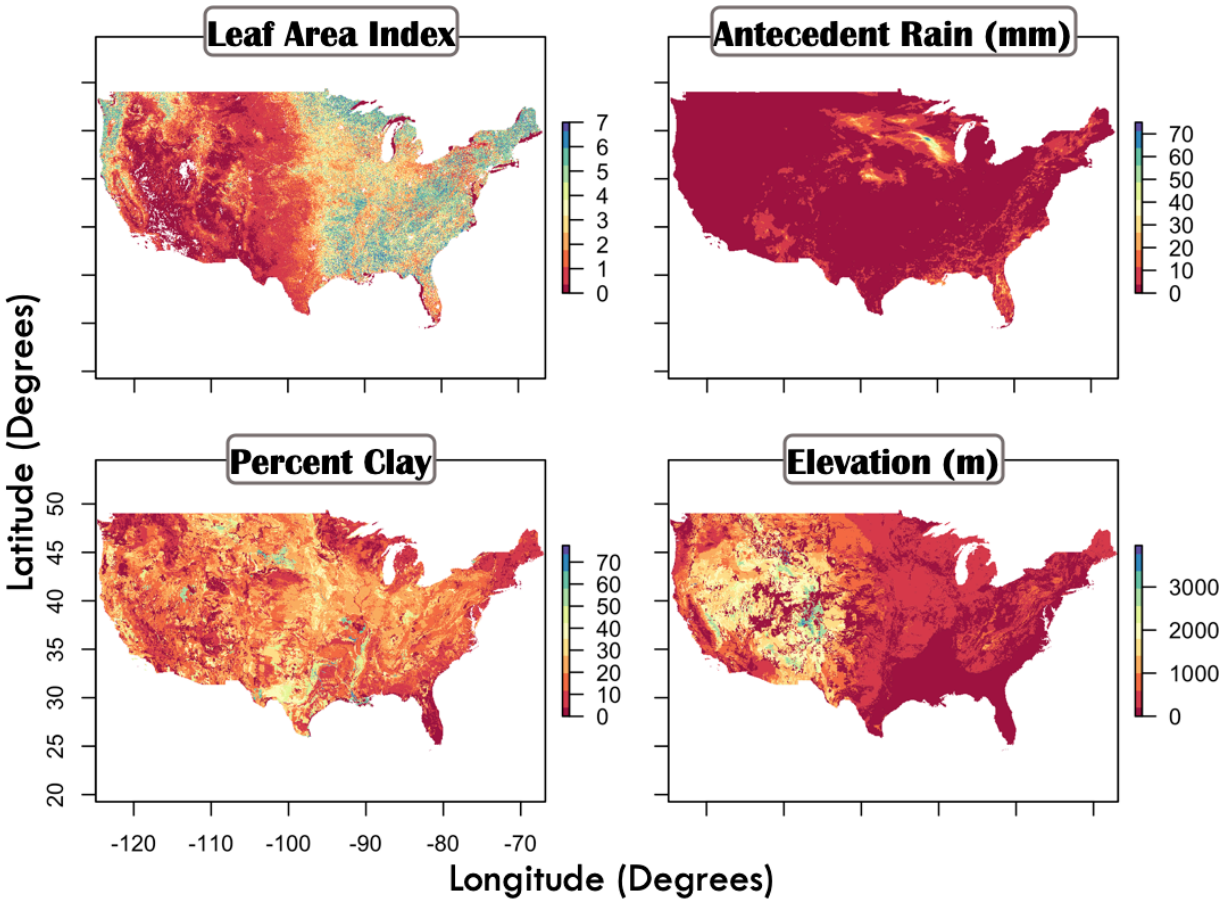


Figure S6 Covariate plots for July 06, 2020 for Contiguous US (CONUS). All the four covariates exhibit considerable heterogeneity across CONUS.