# MULTISCALE SPATIO-TEMPORAL BIG DATA FUSION OF HYDROLOGICAL VARIABLES FROM POINT TO SATELLITE FOOTPRINT SCALES

A Dissertation

by

DHRUVA KATHURIA

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

| | |
|---|---|
| Chair of Committee, | Binayak P. Mohanty |
| Committee Members, | Matthias Katzfuss |
| | Nicholas Duffield |
| | Ignacio Rodriguez-Iturbe |
| | Patricia K. Smith |
| Head of Department, | John Tracy |

August 2021

Major Subject: Biological and Agricultural Engineering

# ABSTRACT

Soil moisture (SM) and evapotranspiration (ET) are key climate variables governing environmental processes from local to global scales. The global burgeoning of SM and ET datasets holds a significant potential in improving our understanding of multiscale hydrological dynamics. The primary issues that hinder the fusion of SM and ET data are (1) different resolution of the data instruments, (2) inherent spatial variability in SM and ET caused due to atmospheric and land surface controls, (3) measurement errors caused due to imperfect retrievals of instruments, and (4) massive size of the datasets. This dissertation aims to develop data fusion algorithms to combine multiscale data and improve understanding of multiscale SM and ET dynamics while accounting for the above-mentioned challenges. The research questions answered in this dissertation include 1) determining the effects of surface and atmospheric controls on the spatio-temporal mean and co-variance of SM using a non-stationary geostatistical algorithm; 2) predicting SM across multiple scales and quantifying the effects of surface physical controls (soil texture, vegetation, topography) and rainfall on SM distribution as well as their effect on retrieval errors of soil moisture platforms; 3) providing a novel framework to fuse SM data for continental scale analysis and 4) improving existing ET data fusion algorithms by accounting for uncertainty in retrievals and incorporating ancillary data/domain knowledge. It was found that the variance and correlation structure of SM varies significantly with spatial heterogeneity in land surface controls for a watershed in Winnipeg, Canada. For the same watershed, the proposed data fusion framework was applied to combine point, airborne and satellite SM data and it was adept at assimilating and predicting SM distribution across all three scales. The data fusion framework was then extended to combine point and satellite SM data across Contiguous US and the effects of physical controls on SM distribution were quantified. For ET data fusion, a state-space modeling framework was developed to combine daily ET satellite data for three agricultural sites in Texas and it was found that when compared with daily Eddy-Covariance ET data, the proposed approach outperformed the traditional fusion algorithm.

# DEDICATION

To Mom, Dad, Nanu and Nani

# ACKNOWLEDGMENTS

I want to express my sincere gratitude to all the people who supported me during my doctoral studies.

I thank my advisor, Dr. Binayak Mohanty who went beyond his role as an advisor and put considerable time and effort towards the successful completion of my PhD. His constant guidance and motivation helped me achieve my research objectives in a timely manner. I also thank my committee members, Dr. Matthias Katzfuss, Dr. Nicholas Duffield, Dr. Ignacio Rodriguez-Iturbe and Dr. Patricia Smith for their constant motivation and feedback which helped me finish my dissertation in time. I also want to thank Dr. Allen Berthold for his constant guidance over the past two years. I am also grateful to my Master's thesis advisor Dr. Pradeep Mujumdar for motivating me to pursue Phd at Texas AM University under Dr. Mohanty. I also want to thank Dr. Kerry-Cawse Nicholson for her guidance and providing me with an opportunity to present my research to the Carbon-Club at Jet Propulsion Laboratory.

I also want to thank all the support staff at Texas AM who tirelessly worked towards making my experience at Texas AM smooth. I especially want to thank David Riggs, Stormy Kretzschmar, Ashlea Schroeder, Cheryl Yeager and Amy Santoy for their selflessness and always helping me in time of need.

My deepest gratitude goes out to my parents Mrs. Geetanjali Kathuria and Mr. Parmod Kathuria for always putting my needs above theirs. My dissertation would not have been remotely possible without their constant support and sacrifices they made for my career and my happiness. I also want to thank my late paternal grandparents, Mr. Sundar Kathuria and Mrs. Yashodra Kathuria and maternal grandparents, Mr. Kishan Nayyar and Aprajita Nayyar for their love and their constant belief in me. I want to thank my brother, Arjun Kathuria and sister-in-law Shubhi Kathuria for their support and love for the last five years. I also want to thank my cousins Naina Nayyar and Kunal Nayyar for being there for me whenever I needed them.

My lifelong friends, Pavneet, Avidesh, Garv, Sahil, Mukul, Varun and Abhishek have greatly

influenced me and I can never thank them enough for always being a pillar of support. I also want to thank Shahina and Antik for being there whenever I needed them. I also want to thank Shubham for being an amazing flatmate and helping me out during my final semester. I also want to thank the past members of my research group Nandita, Neelam, Jonggun, Zhenlei, Samagra and Siddharth for their support and guidance during my initial years at Texas AM University. I am grateful to the current members Deanroy, Amir, Vinit, Minki, Leah and Evan for being a source of constant joy and inspiration during my time at Texas AM. I also want to thank Deepa Mohanty for her love and kindness and providing me with a family away from home. I express my gratitude towards my speech therapist Carl Herder, in helping me tremendously for the past two years.

Lastly, I thank God for giving me the opportunity and resources to pursue my doctoral studies at Texas AM University.

# CONTRIBUTORS AND FUNDING SOURCES

## Contributors

This work was supervised by a dissertation committee consisting of Dr. Binayak Mohanty, Dr. Matthias Katzfuss, Dr. Patricia Smith , Dr. Nicholas Duffield and Dr. Ignacio Rodriguez-Iturbe.

## Funding Sources

This work was made possible in part by the following grants: NASA-SMAP-ST: SMAP Science using Multi-Scale Data Fusion: Forecasting Flash Drought to Flash Flood; NASA Jet Propulsion Laboratory-SURP: Linking the Texas Water Observatory with ECOSTRESS; NASA-SUSMAP -SUSMAP: Root Zone Soil Hydraulic Property Estimation by SMAP.

# TABLE OF CONTENTS

Page

## LIST OF FIGURES

FIGURE Page

# LIST OF TABLES

TABLE Page

# 1. INTRODUCTION

## 1.1 Soil Moisture

Soil moisture (SM) is a critical variable modulating the water and energy fluxes between earth and atmosphere. At a local scale, SM is an essential variable for efficient agricultural and irrigation practices [3] while at a global scale it affects various environmental processes such as climate [4], biogeochemistry [5], carbon cycle [6], heat waves [7, 8], floods [9], droughts [10] and wildfires [11].

### 1.1.1 Spatio-temporal heterogeneity of Soil Moisture

SM exhibits spatio-temporal heterogeneity across scales governed by complex interactions between different surface and atmospheric controls. Precipitation, soil texture, topography and vegetation are considered as the dominant controls affecting SM distribution across space and time [12, 13, 14, 15, 16, 17, 18, 19, 20]. These physical controls interact dynamically to create SM patterns and it is often difficult to isolate their individual effects [21]. Spatio-temporal correlation in SM is a crucial factor to understand sub-grid SM dynamics and the mean and variance alone fail to provide an exhaustive characterization of SM variability [22, 23, 21, 24]. Geostatistics has traditionally been used to quantify this correlation structure in SM. The assumption of stationarity/isotropy, however, has been one of the chief reasons limiting the use of geostatistics in heterogeneous environments to determine the spatio-temporal variability of SM and its subsequent upscaling to the required support scales.

### 1.1.2 Soil Moisture Data Fusion

SM is measured at various resolutions ranging from a few centimeters to hundreds of kilometers and an SM measurement is defined using the scaling triplet: support, extent and spacing. Support refers to the representative area of the data sample, extent refers to the total coverage area of the data, and spacing refers to the distance between individual samples [25, 26]. SM is available from local to global scales from a variety of instruments [27] such as capacitance sensors, time domain reflectometers, active-passive airborne sensors, cosmic-ray neutron probes, Global Positioning Systems, and satellites such as Soil Moisture Ocean Salinity (SMOS) and Soil Moisture Active Passive (SMAP).

Data fusion of SM denotes the process of combining SM data from multiple platforms into a single holistic composite product such that the composite is more accurate and provides more information than any individual platform [28]. The abundance of multiscale SM data presents a never-before opportunity for the scientific community to improve its understanding of multi-scale SM dynamics. There is, thus, a need for a data-driven fusion scheme to combine remote sensing data with traditional point sensors.

### 1.1.3 Soil Moisture Big Data Fusion

The past few years has witnessed an explosive growth in SM (and other environmental) remote sensing data. This so called "Big" SM data are spatio-temporal (indexed by a spatial coordinate and a time stamp) and have the potential to uncover novel insights into multiscale SM dynamics from local to global scales. Usually, SM data are

- spatio-temporally dependent,
- available at multiple resolutions from various instruments, and
- 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 multiscale SM behaviour. Paradigms facilitating the fusion of disparate SM data while handling the sheer size of datasets are thus critical.

## 1.2 Evapotranspiration

Evapotranspiration (ET) is a key flux coupling the global water, energy and carbon cycle influencing a variety of processes and activities such as agricultural water management [29, 30], drought estimation [31, 32], water and energy balance closures [33, 34], water rights studies [35],

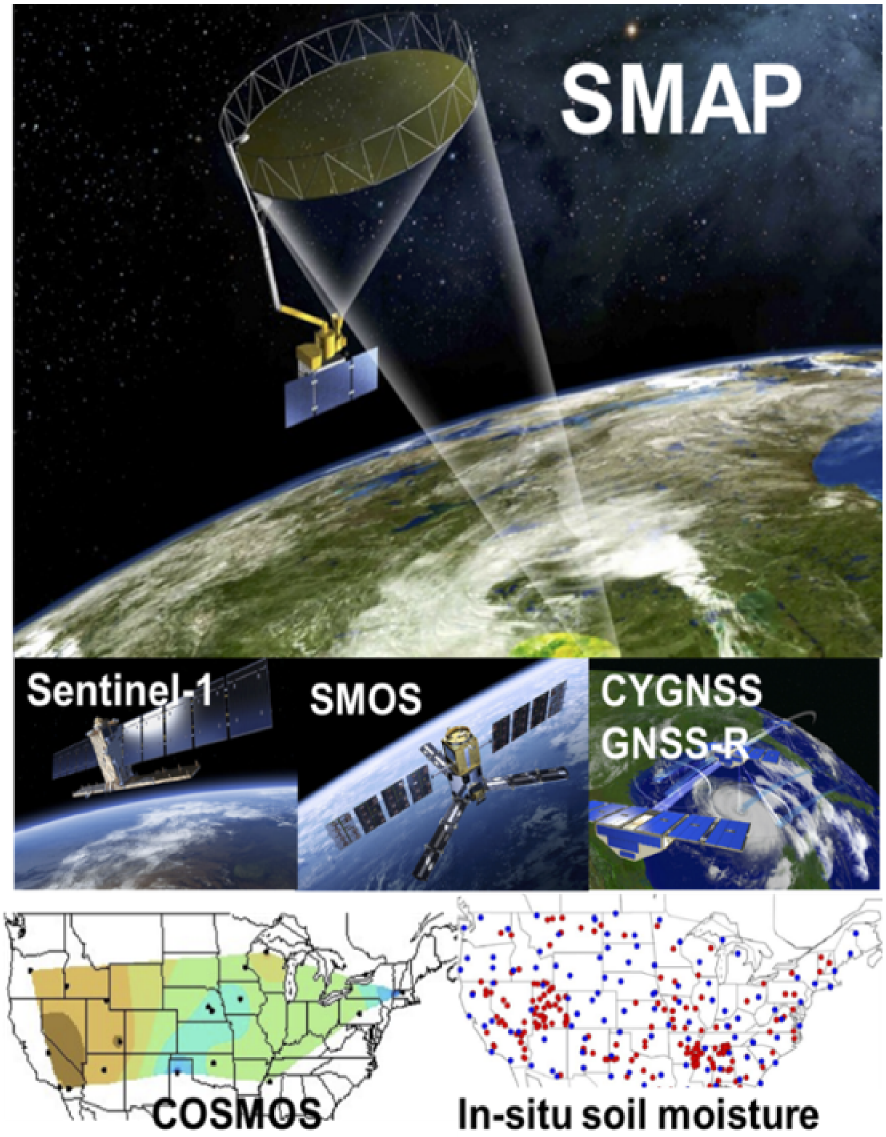

Figure 1.1: Different soil moisture platforms from point to satellite support scales

and atmospheric processes [36]. Remote sensing has emerged as an accurate and relatively inexpensive way to estimate ET over vast spatio-temporal domains and fusion of existing ET remote sensing platforms thus is critical.

### 1.2.1 Evapotranspiration data fusion

Current ET remote sensing platforms either provide observations at a fine spatial resolution ($\sim 70 - 100m$) but have extended revisit times or else provide data at a coarse spatial resolution albeit at a high temporal frequency. At sub-field scales, ET data at a fine spatial resolution is therefore available after prolonged time intervals. Combining different remote sensing ET products is therefore challenging because of the sparsity of data at a fine spatial resolution. Additionally, the data fusion scheme should account for measurement errors in the platforms and errors caused due to different retrieval methods as well as provide a framework to include additional physical/empirical knowledge of the underlying variable.

## 1.3 Motivation and Objectives

Due to the importance of SM and ET in water, energy and carbon cycles, the motivation of this dissertation is to increase our understanding of SM and ET dynamics at multiple scales as well as provide multiscale estimates of these variables typically required from local to regional levels for varied applications. This is achieved by developing and presenting data fusion algorithms combining SM and ET data available from disparate multi-resolution platforms while accounting for the challenges associated with remote sensing retrievals and incorporating the effect of different physical controls on the underlying spatio-temporal distribution of these variables. Thus, the objectives of this work are:

1. Present a framework to model SM in a nonstationary setting using a flexible geostatistical model whose variance/correlation structure varies with underlying atmospheric and surface heterogeneity.

2. Present a multiscale fusion algorithm to combine SM data from multiple instruments (in an optimal way) while partitioning the inherent spatio-temporal dynamics of SM distribution and the measurement errors induced by different data instruments.

3. Extend the data fusion scheme in Objective 2 to a Big Data setting when the size of the datasets becomes massive.

4. Propose a data fusion scheme for ET estimation when the fine resolution ET data are temporally sparse.

## 1.4 Hypotheses

The dissertation evaluates the following hypotheses:

1. Physical controls at a particular location will affect both the variance of SM at that location as well as its correlation with SM at other locations.

2. Geostatistics driven multiscale data fusion of SM platforms increases predictive capability of SM across multiple scales and can be efficiently scaled to combine Big data across vast spatio-temporal domains.

3. Existing unsupervised ET data fusion algorithms can be modified to include ancillary data and domain knowledge for improved ET estimation at fine spatial resolutions.

The dissertation addresses the objectives of this research in the five chapters that follow. Chapter 2 proposes a novel non-stationary geostatistical framework such that the spatial distribution of SM varies with the underlying physical controls. Chapter 3 presents a multiscale data fusion framework combining a hierarchical paradigm with the the non-stationary algorithm proposed in Chapter 2. Chapter 4 extends the data fusion to a "Big" data setting by proposing a novel numerical approximation such that the data fusion algorithm can be applied efficiently for Big data. Chapter 5 focuses on data fusion of ET platforms where existing data fusion schemes are modified such that they can include uncertainty in observations, ancillary data and domain knowledge. Chapter 6 concludes the dissertation by providing a summary as well as future directions of the research.

# 2. A NONSTATIONARY GEOSTATISTICAL FRAMEWORK FOR SOIL MOISTURE PREDICTION IN THE PRESENCE OF SURFACE HETEROGENEITY

## 2.1 Synopsis

[1]Soil moisture is spatially variable due to complex interactions between geologic, topographic, vegetation and atmospheric variables. Correct representation of sub-grid soil moisture variability is crucial in improving land surface modeling schemes and remote sensing retrievals. In addition to the mean structure, the variance and correlation of soil moisture are affected by the underlying land surface heterogeneity. This often violates the underlying assumption of stationarity/isotropy made by classical geostatistical models. The present study proposes a geostatistical framework to predict and upscale soil moisture in a non-stationary setting using a flexible spatial model whose variance/correlation structure varies with changing land surface characteristics. The proposed framework is applied to model soil moisture distribution using *in situ* data in the Red River watershed in Southern Manitoba, Canada. It is seen that both the variance and correlation structure exhibits spatial non-stationarity for the given surface heterogeneity driven primarily by vegetation and soil texture. At the beginning of the crop season, soil texture plays a critical role in the drying cycle by decreasing variance and increasing correlation as the soil becomes drier. Once the crops begin to mature, vegetation becomes the dominant driver, promoting spatial correlation and reducing SM variance. We upscale our point scale soil moisture predictions to the airborne extent ($\sim$ 1.5 km) and find that the upscaled soil moisture agrees well with the observed airborne data with RMSE values ranging from 0.04 to 0.08 (v/v). The proposed framework can be used to predict and upscale soil moisture in heterogeneous environments.

[1] Reprinted with permission from, ' A nonstationary geostatistical framework for soil moisture prediction in the presence of surface heterogeneity' by Kathuria et al. (2019) Water Resour. Res., 55, 729-753, https://doi.org/10.1029/2018WR023505, Copyright 2019 John Wiley and Sons, Inc.

## 2.2 Introduction

Although soil moisture (SM) comprises a minuscule percentage of the global liquid freshwater, it forms an essential link between the earth and the atmosphere by modulating the water and energy fluxes between the two systems. SM is critical for optimal agricultural and irrigation practices [3] at a local scale. On a global scale, it affects weather [37] and climate [4] predictions, drought forecasts [10] and future soil carbon predictions [6]. It plays a crucial role in characterizing biogeochemical processes [5] and quantifying terrestrial dust emissions [38, 39]. Through its interactions with the atmosphere, SM acts as a driver for the intensification of extreme events such as floods [9], heat-waves [7, 8] and wildfires [11] among others.

SM is highly variable in both space and time [21, 40, 41]. Correct representation of SM variability is crucial in improving land surface modeling schemes [42] and aids to reduce bias in predictions of water and energy fluxes [43]. Understanding SM variability at a fine scale is critical for planning ground-based sampling schemes, which in turn, plays a vital role in the validation of satellite SM retrievals [23].

SM variability across different spatial scales is largely controlled by complex interactions between geologic, topographic, vegetative and atmospheric variables. While large scale ($> 100$ km) SM patterns are mainly controlled by precipitation and evapotranspiration, sub-grid variability is affected by heterogeneity in surface controls [12]. Soil texture, topography and vegetation in general, are the dominant surface controls affecting SM distribution.

Soil texture is defined by the relative percentages of sand, silt and clay in a given soil type. Soil texture affects the water holding capacity of the soil and has a pronounced effect on the SM distribution. The same has been validated by several studies [15, 14]. [13] found that surface SM variability was strongly affected by soil texture using a semi-variogram analysis. For a semi-arid region in Nebraska, [44] inferred that an increase in spatial variability in soil texture led to a corresponding increase in variability of SM. [16] showed that soil texture affects the correlation lengths of SM at different support scales. [45] also observed soil texture to be the dominant physical control in Iowa and Oklahoma at both point and airborne support scales.

Topography usually affects the spatial distribution of SM in wet conditions and the same was validated by [45]. [17] found that topography affects the SM distribution during and immediately after a rainfall event while [18] inferred that topography influences SM distribution under uniform vegetation conditions.

Vegetation influences both the downward flow of water (via interception) and the upward flow (through transpiration) of water-vapor. Unlike texture and topography, the effect of vegetation is dynamic and can vary significantly throughout the year. Vegetation can stimulate an increase or decrease in spatial variability of SM based on the season and time of the year [19, 20].

Although many studies have found a hierarchy or dominance of one surface control over the other, the consensus is that soil texture, vegetation and topography interact dynamically to create SM patterns and it is often difficult to isolate their individual effects [21]. These interactions vary with diverse wetness conditions and with the dynamic stages during the wetting and drying cycles [46, 47]. The surface controls can act together to either create or destroy spatial variability [48] and also govern the relationship of the variance with the mean SM across climate zones [49]. In addition to influencing the mean and variance, heterogeneity in surface controls also affects the spatial correlation structure of SM [50, 51, 14] engendering non-stationarity in correlation lengths [13]. Nested correlation structures in SM distribution have also been attributed to both the heterogeneity of the surface controls and spatial variability of precipitation with smaller correlation lengths fostered by surface controls and larger correlation lengths by rainfall patterns [16].

Correlation in SM is a crucial factor to understand sub-grid SM dynamics [52], and the mean and variance alone fail to provide an exhaustive characterization of SM variability [22, 21]. Therefore, in addition to the mean and variance, spatial correlation should also be considered when modeling spatial distribution of SM. Classical geostatistics has been traditionally used to model SM and have generally relied on variogram-based approaches to quantify the above mentioned correlation structure in SM. Variogram techniques, though straight forward to implement, should be used with caution, as they generally assume some form of stationarity/isotropy in the SM distribution. The assumption of stationarity/isotropy has been one of the chief reasons limiting the use of geostatistics in heterogeneous environments to determine the spatial variability of SM and its subsequent upscaling to the required support scales. While there has been significant growth in varied SM estimation techniques such as wavelet analysis [53, 14], temporal stability analysis [54, 15, 55], thermal inertia [56], machine learning algorithms [57, 58], data assimilation [59, 60], etc., the application of geostatistics to the same has been waning in recent years. This is unfortunate, because parameters of geostatistical methods have direct physical interpretations enabling rich scientific inference and increased potential for transferability to data-scarce regions. Recent state of the art work in big data statistics [61, 62] has also enabled geostatistics techniques to be implemented on a global scale, making them attractive candidates to do large-scale SM inference. Therefore, the motivation of the present work is to present a geostatistical framework to predict and upscale SM in a non-stationary setting. The hypothesis of this study is that local surface controls at a particular location will affect both the variance of SM at that location as well as its correlation with SM at other locations. For example, the variance/correlation characteristics for a region with predominantly bare sandy soil will be different than a clay soil in an agricultural field. In such cases, stationarity no longer holds and classical geostatistical methods tend to fail. The objective of the present work is to present a framework in such a scenario using a flexible stochas-

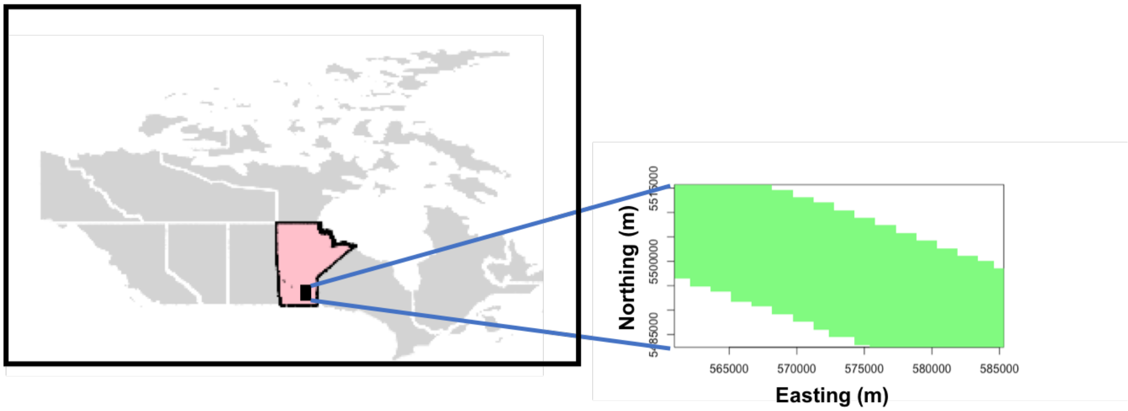

Figure 2.1: Study site located in the Red River watershed at Winnipeg in the state of Manitoba in Canada [1].

tic geostatistical model whose variance/correlation structure varies with changing surface controls. To achieve this, we use a class of non-stationary spatial models adapted from [63] and [64]. The point scale spatial distribution of SM is then upscaled to an areal (block) support following [65]. This approach is formally optimal and physically interpretable, as opposed to the ad-hoc upscaling techniques usually employed in practice.

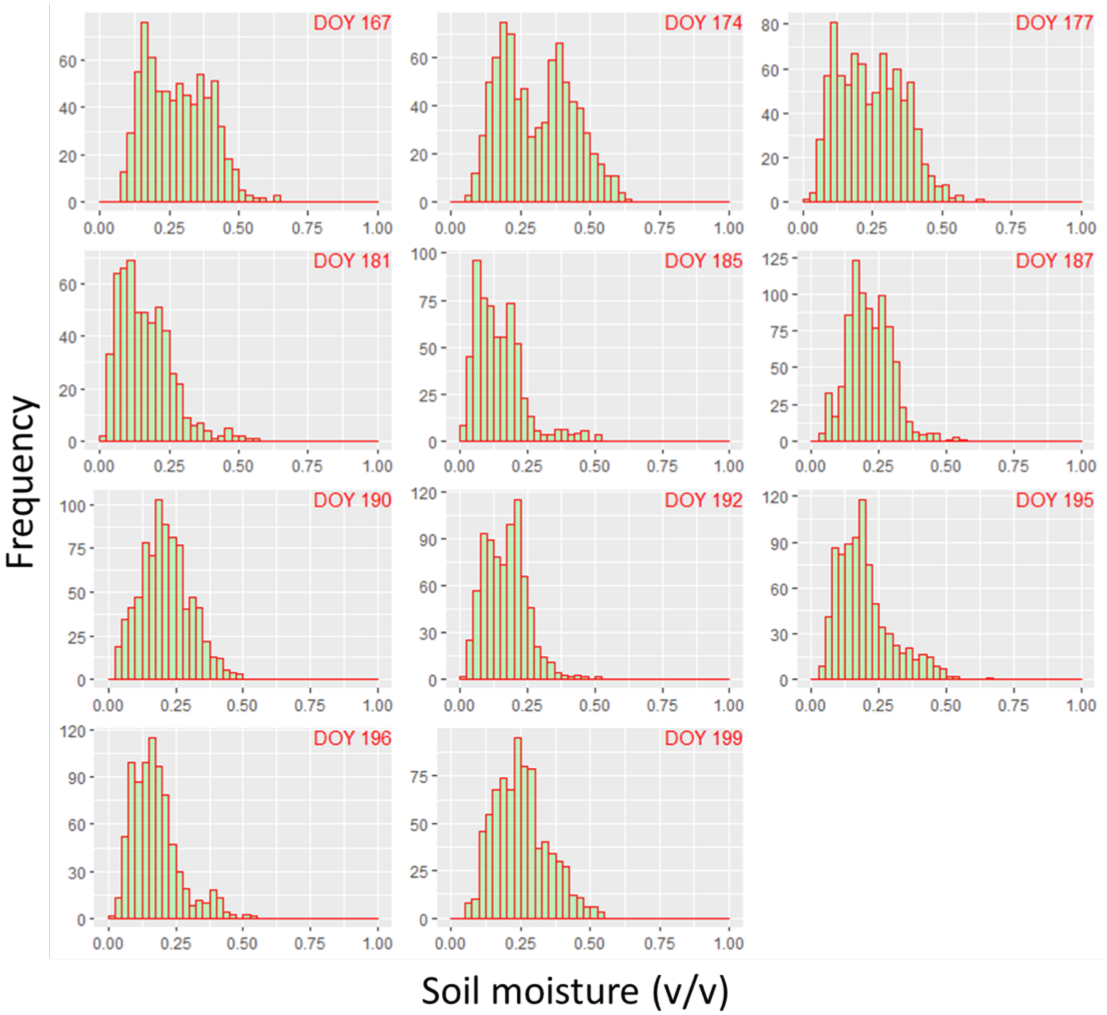

Figure 2.2: Histogram of point scale soil moisture distribution during the SMAPVEX12 campaign. Soil moisture exhibits a large range of wetness conditions due to underlying land surface heterogeneity [1].

The proposed framework is applied to a watershed exhibiting large variability in both wetness conditions and surface heterogeneity in Winnipeg, Canada. Though various non-stationary covariance functions have been proposed in literature [66, 67, 68], to the best of the authors' knowledge, this is the first study which explicitly models and aggregates the spatially varying variance and correlation structure of SM in terms of the underlying surface heterogeneity using a geostatistical framework. In the following sections of this chapter, vectors and matrices have been denoted by bold-faced letters.

## 2.3 Study Area

The Soil Moisture Active Passive Validation Experiment 2012 (SMAPVEX12) was conducted from June 6th to July 17th 2012 in Winnipeg, Manitoba (Canada) (Figure 2.1) which is classified as having fully humid climate (Dfb classification) according to the Köppen-Geiger climate classification [69], with an average annual precipitation of 521 mm. This study area was chosen because the region experiences significant spatial variability in SM arising due to heterogeneity in soil texture and vegetation and as a result exhibits a large range of wetness conditions in the study domain (Figure 2.2). There is a sharp contrast of soil texture across the site, with heavy clays in the east to fine loamy sands in the west. The land cover is dominated by perennial agricultural crops (cereals, soybeans, canola and corn) with approximately 15% of land occupied by wetlands and forests in the northwest region. The campaign was carried out during the crop-growing season with low biomass contents ($<0.1$ kg/m$^2$) at the start of the campaign and reaching biomass contents of 1-2 kg/m$^2$ for bean crops and 4 kg/m$^2$ for corn in the final week.

During the campaign, airborne data were collected for 15 days using a passive-active L-band sensor (PALS). *In situ* SM data were also collected during the flight days using Stevens Water Hydra Probe and Delta-T Theta probes. The sampling was primarily done in the agricultural part of the watershed and we restrict ourselves to the subset of the watershed for which these point scale observations are available. For each agricultural field, SM was measured at 16 sampling locations and three replicate readings were taken at each location. For complete details of the study site and the SMAPVEX12 campaign, interested readers are encouraged to refer to [70]. The SM data at both *in situ* and airborne scales as well as the soil texture and rainfall data were

Table 2.1: Summary statistics for soil texture and vegetation [1].SD = Standard Deviation

| Surface control | Minimum | $10^{th}$ percentile | Mean | $90^{th}$ percentile | Max | SD |
|---|---|---|---|---|---|---|
| Percent clay | 5 | 5 | 34.61 | 65 | 65 | 24.58 |
| Percent silt | 5 | 5 | 19.79 | 33 | 42 | 9.47 |
| LAI | | | | | | |
| DOY 167 | 0.43 | 0.65 | 1.67 | 3.50 | 4.58 | 1.04 |
| DOY 174 | 1.00 | 1.34 | 2.70 | 4.71 | 6.26 | 1.27 |
| DOY 177 | 1.15 | 1.54 | 3.10 | 5.47 | 6.70 | 1.50 |
| DOY 181 | 1.11 | 1.49 | 2.56 | 4.42 | 6.38 | 1.19 |
| DOY 185 | 0.32 | 1.83 | 2.85 | 4.32 | 5.97 | 0.97 |
| DOY 187 | 1.38 | 2.17 | 3.25 | 4.44 | 5.83 | 0.88 |
| DOY 190 | 1.63 | 2.29 | 3.79 | 5.26 | 6.27 | 1.08 |
| DOY 192 | 1.77 | 2.65 | 4.00 | 5.42 | 6.07 | 1.03 |
| DOY 195 | 1.59 | 2.10 | 3.85 | 5.09 | 6.03 | 1.10 |
| DOY 196 | 1.53 | 1.97 | 3.74 | 5.16 | 6.13 | 1.17 |
| DOY 199 | 1.06 | 2.05 | 3.63 | 5.30 | 6.01 | 1.18 |

accessed from the National Snow and Ice Data Center (NSIDC) website. The Leaf Area index (LAI) data were extracted from the four-day composite MODIS (MCD15A3H, version 6) product at a 500 meters resolution available at National Aeronautics and Space Administration (NASA) Land Processes Distributed Active Archive Center (LPDAAC). The summary statistics for soil texture and vegetation are given in Table 2.1.

## 2.4 Theory

Geostatistics has traditionally been used to model and predict SM patterns across different spatial scales [18, 26, 71, 16, 46] where it is assumed that the observed SM is an incomplete sample from a single realization of a continuous spatial stochastic process $\{y(s) : \boldsymbol{s} \in \mathcal{D}\}$ or $y(.)$ in a spatial domain or region $\mathcal{D}$. $y(s)$ is typically structured as:

$$y(s) = \mu(s) + e(s) \tag{2.1}$$

where $\mu(s) \equiv E(y(s))$ is the deterministic mean function and, $e(s)$ is the zero-mean spatially dependent stochastic process. The most commonly used specification for the mean function is

$\mu(s) = \boldsymbol{X}(s)^T\boldsymbol{\beta}$, where $\boldsymbol{X}(s)$ is the vector of covariates at $s$ and $\boldsymbol{\beta}$ represents the vector of regression coefficients. In traditional geostatistics, the $\boldsymbol{\beta}$ vector is sometimes first provisionally estimated as $\boldsymbol{\beta}_{ols}$ from the data ignoring $e(.)$, using ordinary least squares (OLS). The estimated mean function $\boldsymbol{X}^T\boldsymbol{\beta}_{ols}$ is then subtracted from the data resulting in residuals. Empirical variograms are then generated for the residuals and parameteric variogram models are "matched" to the empirical variograms to get an estimate of parameters in $e(.)$ using method of moments [72, 73, 74].

Though easy to implement, the above mentioned technique for parameter estimation, also referred to as classical geostatistics [75] or two-point statistics [21] is largely informal and not optimal, suffering from biased parameter estimates due to provisional estimation of the mean [76, 77, 75]. Further, empirical variograms are sensitive to outliers and become unstable as the lag distance increases [77]. It has been seen in studies that the parameters estimated from empirical variograms are highly uncertain even with more than 200 SM measurements [24]. Therefore, it is suggested that direct parameter estimation using empirical variograms should be avoided and instead, they should be used as graphical exploratory tools for formal statistical procedures such as maximum likelihood (ML) estimation [77].

Parameter estimation based on the likelihood function is statistically formal and is optimal under mild regularity conditions [75]. The only additional assumption it requires is a fully specified joint distribution for $y(.)$ in equation (2.1) by explicitly defining parameteric forms for $\mu$ and for the covariance of $e$ [75]. A popular choice for the above is assuming the mean $\mu = \boldsymbol{X}^T\boldsymbol{\beta}$ (as before), and the spatially dependent stochastic process $e(.)$ as a Gaussian Process with mean zero and covariance $C$ such that $e(.) \sim GP(0, C)$. For a finite set of $n$ locations $\mathcal{S} = \{s_1, ..., s_n\}$ with $s_i \in \mathcal{D}$, $\mathbf{y} = \mathbf{y}(\mathcal{S}) = (y(s_1), ..., y(s_n))$ has a multivariate normal distribution given as:

$$\mathbf{y} = (y(s_1), ..., y(s_n)) \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{C}) = (2\pi)^{-n/2}|\boldsymbol{C}|^{-1/2}exp(-(1/2)(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{C}^{-1}(\mathbf{y} - \boldsymbol{\mu})), \quad (2.2)$$

where $\boldsymbol{\mu}$ is an $n$-dimensional vector= $(Ey(s_1), ..., Ey(s_n)) = (\boldsymbol{X}(s_1), ..., \boldsymbol{X}(s_n))^T\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{\beta}$ and $\boldsymbol{C}$ is a $n \times n$ matrix with $(i, j)^{th}$ element equal to $C(s_i, s_j)$ for $i, j = 1, .., n$. Let $C$ be a function of a parameter vector $\boldsymbol{\theta}$. The log-likelihood of the joint distribution is then given as:

$$log\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\theta}) = -\frac{1}{2}(log|\boldsymbol{C}(\boldsymbol{\theta})| + (\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta}))^T\boldsymbol{C}^{-1}(\boldsymbol{\theta})(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})) + nlog(2\pi)) \tag{2.3}$$

Then, the ML estimates are the parameter values $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ which maximize equation (2.3).

It is possible to find the parameters which maximize equation (2.3) by maximizing the profile log-likelihood which depends only on $\boldsymbol{\theta}$ and is therefore computationally cheaper to maximize. The corresponding profile log-likelihood of equation (2.3) is given by:

$$l(\boldsymbol{\theta}) = -\frac{1}{2}(log|\boldsymbol{C}(\boldsymbol{\theta})| + \mathbf{y}'P(\boldsymbol{\theta})\mathbf{y} + nlog(2\pi)), \tag{2.4}$$

where $P(\boldsymbol{\theta}) := \boldsymbol{C}^{-1} - \boldsymbol{C}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{C}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C}^{-1}$. The corresponding value of $\hat{\boldsymbol{\beta}}$ is a function of $\hat{\boldsymbol{\theta}}$ given by:

$$\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}) = (\boldsymbol{X}'\boldsymbol{C}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{y}. \tag{2.5}$$

Note that $\hat{\boldsymbol{\beta}}$ is the generalized least squares estimator of $\boldsymbol{\beta}$ if the covariance of $\mathbf{y}$ is equal to $\boldsymbol{C}(\hat{\boldsymbol{\theta}})$ [75]. In practice, the covariance function $C$ is often assumed to be isotropic, such that the covariance of the process $e(.)$ between any two points $s_1$ and $s_2$ depends only on the lag-distance between them, or $Cov[e(s_1), e(s_2)] = C(|s_1 - s_2|)$. Then $C$ is modelled using a parametric isotropic covariance function such as exponential, spherical, Matérn, etc. Sometimes, a nugget effect, representing microscale variability or observational error, is added to $C$ such that the resulting covariance function between any two points $s_1$ and $s_2$ becomes $C(s_1, s_2) + \tau^2 1_{[s_1=s_2]}(s_1, s_2)$, where $\tau^2$ represents the nugget variance and $1_{\boldsymbol{R}}(\boldsymbol{x})$ is the indicator function defined as $1_{\boldsymbol{R}}(\boldsymbol{x}) := \begin{cases} 1 & \boldsymbol{x} \in \boldsymbol{R} \\ 0 & \boldsymbol{x} \notin \boldsymbol{R} \end{cases}$ [77]. $\tau^2$ is assumed to be a constant and can either be known or be estimated from the data. The resulting covariance function is isotropic and as mentioned before this assumption is not always valid. In the next section we define the controls-driven non stationary covariance function for such situations.

## 2.5 Methodology

In this section, we present the statistical framework for SM prediction and upscaling under a non stationary setting. The utility of empirical variograms as exploratory tools is discussed in Section 2.5.1. The controls-driven non stationary covariance function is then presented in Section 2.5.2. We next detail tools for performing global optimization to find the ML parameter estimates and discuss procedures to do physical interpretation from the parameters. This is followed by spatial prediction at point scale and the upscaling algorithm in Section 2.5.4. The proposed geostatistical framework is summarized in Figure 2.3.

### 2.5.1 Exploratory variogram analysis

We use empirical variograms to perform exploratory analysis for SM data and check whether the assumption of stationarity holds. To do this, we divide the study area into sub-regions and draw variograms for each of the sub-regions after subtracting the mean. If the variograms for all the sub-regions are similar, we have some evidence that the process is spatially stationary. If the variograms appear significantly different, we test for non-stationarity. The variograms also give an indication of the structure of the non-stationary models.

### 2.5.2 Surface controls-driven non stationary covariance function

We adapt the spatial model of [63] and spatio-temporal model of [64] to develop the non-stationary covariance function. Since the basic motivation of our model is that different combinations of surface controls will govern the spatially dependent process $e(.)$, following [63], we model $e(.)$ in equation (2.1) as:

$$e(s) = \sum_{j=1}^{M} w_j(s) e_j(s) \tag{2.6}$$

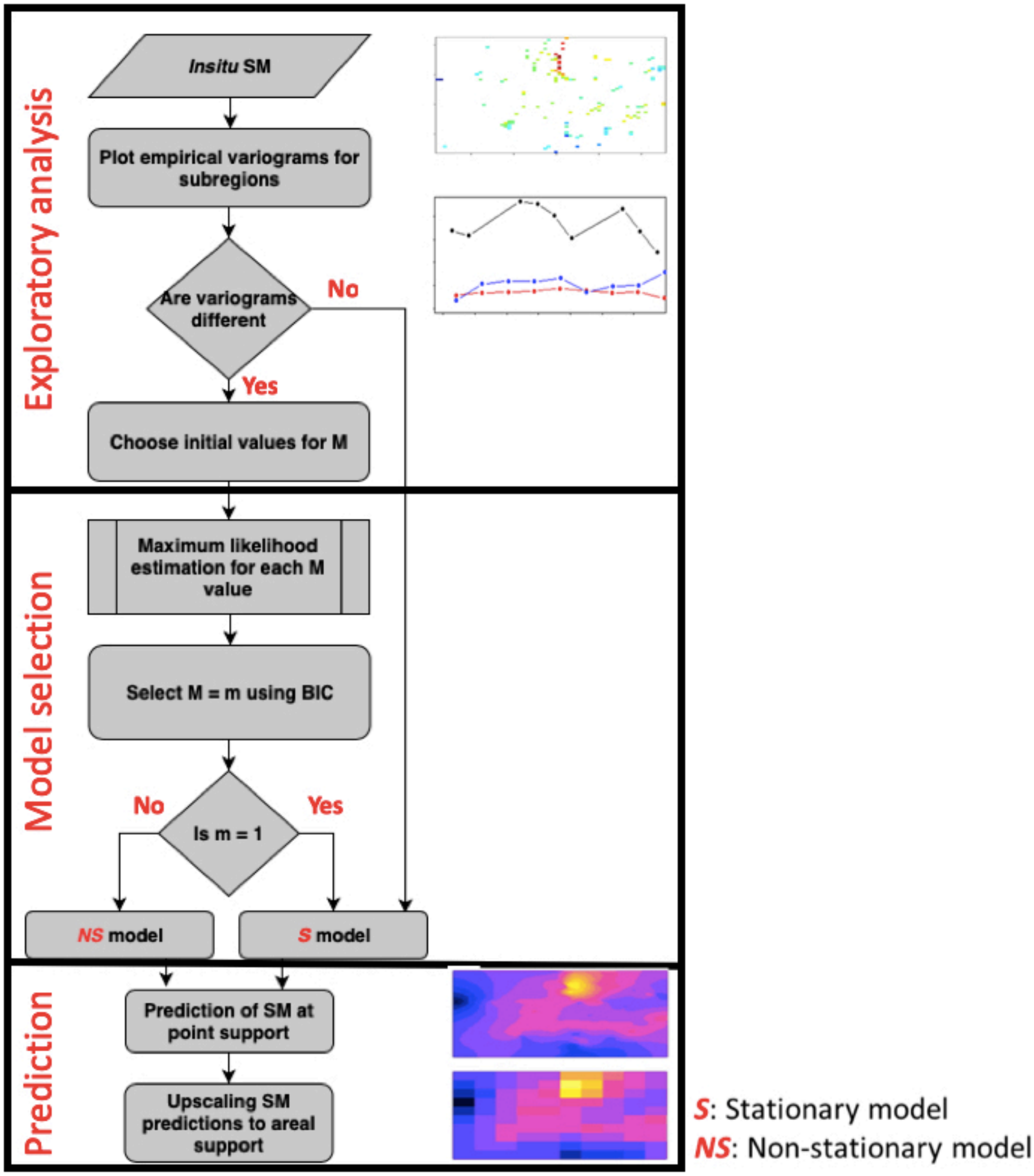
Exploratory analysis
Insitu SM
Plot empirical variograms for subregions
Are variograms different
No
Yes
Choose initial values for M
Model selection
Maximum likelihood estimation for each M value
Select M = m using BIC
Is m = 1
No
Yes
NS model
S model
Prediction
Prediction of SM at point support
Upscaling SM predictions to areal support
S: Stationary model
NS: Non-stationary model

Figure 2.3: Flowchart of the geostatistical framework [1].

that is, the spatially dependent process at a point $s$ is defined as the weighted sum of $M$ stochastic terms, $e_j(s)$, for $j = 1, .., M$. We take the individual $e_j$ to be independent zero-mean Gaussian processes with isotropic covariance functions $C_j$. Further, following [64], we model the weights $w_j$ as functions of surface controls $\boldsymbol{X}$. This ensures that the underlying surface heterogeneity governed by different combinations of surface controls determines the variance and correlation structure of SM distribution. Since the individual $e_j$s are independent, the covariance function for $e(.)$ is given as:

$$\begin{aligned} Cov(e(s_1), e(s_2)) &= Cov(\sum_{j=1}^{M} w_j(\boldsymbol{X}(s_1))e_j(s_1), \sum_{j=1}^{M} w_j(\boldsymbol{X}(s_2))e_j(s_2)) \\ &= \sum_{j=1}^{M} w_j(\boldsymbol{X}(s_1))w_j(\boldsymbol{X}(s_2))C_j(|s_1 - s_2|) \\ &= C(s_1, s_2, \boldsymbol{X}(s_1), \boldsymbol{X}(s_2)) \end{aligned} \tag{2.7}$$

The covariance for $e(.)$ can therefore be represented by the non-stationary covariance function $C$ in equation (2.7). $C$ is now a function of the surface controls $\boldsymbol{X}$, in addition to the lag-distance between the two points, and therefore serves as a non stationary extension to spatial models using isotropic covariance functions. While there can be different ways to model the weighting functions $w_j$, we use a multinomial logistic function given as:

$$w_j(s) = \frac{exp(\boldsymbol{X}(s)^T \boldsymbol{\alpha}_j)}{\sum_{l=1}^{M} exp(\boldsymbol{X}(s)^T \boldsymbol{\alpha}_l)} \tag{2.8}$$

where $\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_M$ are vectors of regression coefficients that describe the effect of surface controls on the covariance function. Note that if the number of controls is $p$, then the length of each $\boldsymbol{\alpha}_i$ vector is $p + 1$ (including the intercept). Using a multinomial function is advantageous because it ensures that the weighting functions $w_j$ always sum to one, making the isotropic covariance function a special case of the above (M =1, makes $w_1 = 1$ and $C = C_1$). This makes the above non-stationary model flexible, as it can be used to model both stationary and non-stationary processes. Since all the weights sum to one, we fix all the components of vector $\boldsymbol{\alpha}_1$ equal to zero as is usually done in logistic regression, which also reduces the number of parameters to be estimated.

For each of the individual covariance functions $C_j$, we choose the Matérn as it belongs to a class of flexible covariance functions and can model a wide variety of spatial processes. Moreover, other widely used covariance functions such as the exponential and Gaussian functions are special cases of the Matérn [75]. The Matérn function is defined as:

$$C(h) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \Big(\frac{h}{\lambda}\Big)^{\nu} K_{\nu}\Big(\frac{h}{\lambda}\Big), \qquad \lambda > 0, \quad \nu > 0 \tag{2.9}$$

where $\sigma^2$, $\lambda$ and $\nu$ are the variance, range and smoothness parameters, respectively, and $K_\nu$ is the modified Bessel function of the second kind of order $\nu$. Therefore, for M=m and $p$ controls, the total number of parameters to be estimated are the regression coefficients for mean, $\boldsymbol{\beta} = [\beta_1, ..., \beta_{(p+1)}]$, regression coefficients for the covariance function, $\boldsymbol{\alpha}_j = [\alpha_{j1}, ..., \alpha_{j(p+1)}]$ for $j = 2, .., m$ and the parameters for Matérn covariance functions $C_j$, $[\nu_j, \lambda_j, \sigma_j^2]$ for $j = 1, ..., m$. We denote all the parameters in $\boldsymbol{\alpha}_j$ and $C_j$ as the vector $\boldsymbol{\theta}$. Note that for $\boldsymbol{\alpha}_j$, we start from j=2 because, as mentioned earlier, we fix all components of $\boldsymbol{\alpha}_1$ equal to zero.

### 2.5.3 Maximum likelihood parameter inference

#### *2.5.3.1 Parameter estimation*

To obtain estimates of the parameters of the non-stationary model defined in Section 2.5.2, we find the ML estimates using the log-likelihood function defined in equation (2.4). Since we do not have closed-form expressions for $\hat{\boldsymbol{\theta}}$, we rely on non-linear optimization techniques to find our parameter estimates. The likelihood surface for spatial models tends to be relatively flat and can also consist of multiple modes [78]. For the non-stationary model in Section 2.5.2, the optimization becomes more involved. In our analysis, we therefore employ stochastic global optimization algorithms to find the ML estimates. The performance of any global optimization algorithm to find optimum parameter estimates is highly dependent on the problem being solved. [79] tested 18 commonly used global optimization algorithms on 48 objective functions to explore the efficacy of the algorithms for a variety of problems. In terms of both speed and accuracy, it was found that the

Generalized Simulated Annealing (GenSA) algorithm [80] and the GENetic Optimization using Derivatives (Genoud) [81] were the most consistent optimization routines.

Simulated annealing is a stochastic optimization algorithm and and is widely used to find global optima for complex non-linear objective functions with multiple local optima. The GenSA method is a generalized and improved form of the classical simulated annealing method [80]. Genoud on the other hand, combines evolutionary search algorithms (such as genetic algorithm) with derivative-based methods (Newton or quasi Newton) in order to solve complex non-linear objective functions with multiple local optima and saddle points. It is relatively faster than pure genetic algorithms, as it also incorporates the derivative information of the objective function. In order to make sure that we arrive at the global ML estimates, we implement both optimization schemes in this study. After finding the ML estimates of $\hat{\boldsymbol{\theta}}$, we obtain $\hat{\boldsymbol{\beta}}$ using equation (2.5).

After finding the ML estimates, interest typically lies in parameter inference as well as spatial prediction at unknown locations. Unlike $\hat{\boldsymbol{\beta}}$, parameter inference for $\hat{\boldsymbol{\theta}}$ is not trivial and we discuss the same in the following section.

#### *2.5.3.2 Effect of surface controls on variance and correlation*

Since the covariance between any two points $s_1$ and $s_2$ is now a function of their lag distance in space, $|s_1 - s_2|$, as well as the vector of the controls at the two points, $\boldsymbol{X}(s_1)$ and $\boldsymbol{X}(s_2)$, the effects of individual covariates on the variance/covariance of SM is not straightforward. A simple way to demonstrate the effect of an individual control on the variance and correlation structure is to analyze the effect of changing the value of that surface control on the non-stationary covariance function while keeping the other controls constant [64]. To do this, we analyze the covariance functions when the surface control is at low (10$^{th}$ percentile), dominant (mean) and high (90$^{th}$ percentile) values while fixing the rest of the controls at the mean values. This will be a good indication of the change in variance/correlation structure when individual surface controls are varied. If each of the surface controls is standardized (mean = 0, variance=1) then this leads to comparing and plotting three covariance functions:

$$\begin{aligned} C_k^{low}(s) &= \sum_{j=1}^{M} \frac{exp(\alpha_{j1} + (x_k^{0.1})\alpha_{jk})}{\sum_{l=1}^{M} exp(\alpha_{l1} + (x_k^{0.1})\alpha_{lk})} C_j(s) \\ C_k^{dominant}(s) &= \sum_{j=1}^{M} \frac{exp(\alpha_{j1})}{\sum_{l=1}^{M} exp(\alpha_{l1})} C_j(s) \\ C_k^{high}(s) &= \sum_{j=1}^{M} \frac{exp(\alpha_{j1} + (x_k^{0.9})\alpha_{jk})}{\sum_{l=1}^{M} exp(\alpha_{l1} + (x_k^{0.9})\alpha_{lk})} C_j(s) \end{aligned} \tag{2.10}$$

where, for the $k^{th}$ surface control $x^k$, $x_k^{0.1}$ and $x_k^{0.9}$ are the $10^{th}$ and $90^{th}$ percentile values. Since $\alpha_{j1}$ is the intercept term in equation (2.8), it is always included. To get the variances, we find the values at $s = 0$, and to get the correlations, we divide the above equations by their corresponding variances.

### 2.5.4 Spatial prediction of SM

*2.5.4.1 Point support*

Having estimated the parameters $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ of our spatial model, spatial prediction is relatively straightforward. Given observations $\mathbf{y} = \mathbf{y}(\mathcal{S})$ at a set of locations $\mathcal{S}$ and our likelihood estimates for the mean function $\hat{\mu}$ and covariance function $\hat{C}$, predictions $\mathbf{y}(\mathcal{S}^P)$ at any set of locations $\mathcal{S}^P = \{s_1^P, ..., s_{n_P}^P\}$ can be derived from the joint distribution of $\mathbf{y}$ and $\mathbf{y}(\mathcal{S}^P)$ given as:

$$\begin{pmatrix} \mathbf{y}(\mathcal{S}^P) \\ \mathbf{y} \end{pmatrix} \sim \mathcal{N}_{n_p+n}\left( \begin{pmatrix} \hat{\boldsymbol{\mu}}(\mathcal{S}^P) \\ \hat{\boldsymbol{\mu}}(\mathcal{S}) \end{pmatrix}, \begin{pmatrix} \hat{\boldsymbol{C}}(\mathcal{S}^P, \mathcal{S}^P) & \hat{\boldsymbol{C}}(\mathcal{S}^P, \mathcal{S}) \\ \hat{\boldsymbol{C}}(\mathcal{S}, \mathcal{S}^P) & \hat{\boldsymbol{C}}(\mathcal{S}, \mathcal{S}) \end{pmatrix} \right) \tag{2.11}$$

From well known properties of multivariate normal distribution, the posterior predictive distribution (PPD) of $\mathbf{y}(\mathcal{S}^P)$ given $\mathbf{y}$ is:

$$\mathbf{y}(\mathcal{S}^P)|\mathbf{y} \sim \mathcal{N}_{n_P}(\boldsymbol{\mu}_{\mathcal{S}^P|\mathbf{y}}, \boldsymbol{C}_{\mathcal{S}^P|\mathbf{y}}) \tag{2.12}$$

where

$$\boldsymbol{\mu}_{\mathcal{S}^P|\mathbf{y}} = \hat{\boldsymbol{\mu}}(\mathcal{S}^P) + \hat{\boldsymbol{C}}(\mathcal{S}^P, \mathcal{S})(\hat{\boldsymbol{C}}(\mathcal{S}, \mathcal{S}))^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}(\mathcal{S})) \tag{2.13}$$

$$\boldsymbol{C}_{\mathcal{S}^P|\mathbf{y}} = \hat{\boldsymbol{C}}(\mathcal{S}^P, \mathcal{S}^P) - \hat{\boldsymbol{C}}(\mathcal{S}^P, \mathcal{S})(\hat{\boldsymbol{C}}(\mathcal{S}, \mathcal{S}))^{-1}\hat{\boldsymbol{C}}(\mathcal{S}, \mathcal{S}^P), \tag{2.14}$$

and $\boldsymbol{\mu}_{\mathcal{S}^P|\mathbf{y}}$ and $\boldsymbol{C}_{\mathcal{S}^P|\mathbf{y}}$ are the mean and covariance of the PPD $\mathbf{y}(\mathcal{S}^P)|\mathbf{y}$.

#### *2.5.4.2 Areal support*

We define SM as a random variable governed by its mean and covariance structure at point support scale. For distribution $y(.)$ at the point support scale, its distribution for an areal pixel $A_1^P$ at spatial support $|A|$ will be the stochastic integral of the distribution $y(.)$ over the pixel:

$$y(A_1^P) = \frac{1}{|A|}{}_{A_1^P} y(s^P) ds^P \tag{2.15}$$

The spatial integral in equation (2.15) is stochastic and can be viewed as an integral over each possible realizations of $y(.)$. The expectation of the PPD of $y(A_1^P)$ given observations $\mathbf{y}$ can be written as:

$$\begin{aligned} \mu^{|A|}_{A_1^P|\mathbf{y}} = E(y(A_1^P)|\mathbf{y}) = E\big(\frac{1}{|A|}{}_{A_1^P} y(s^P) ds^P|\mathbf{y}) &= \frac{1}{|A|}{}_{A_1^P} E(y(s^P)|\mathbf{y}) ds^P \\ &= \frac{1}{|A|}{}_{A_1^P} \mu_{s^P|\mathbf{y}} ds^P \end{aligned} \tag{2.16}$$

which is simply the average of the posterior mean (equation (2.13)) over the pixel $A_1^P$. Similarly, the covariance, $C^{|A|}_{(A_1^P, A_2^P)|\mathbf{y}}$ between any two pixels $A_1^P$ and $A_2^P$ is:

$$C^{|A|}_{(A_1^P, A_2^P)|\mathbf{y}} = Cov(y(A_1^P), y(A_2^P)|\mathbf{y}) = \frac{1}{|A||A|}{}_{A_1^P}{}_{A_2^P} C_{(s_1^P, s_2^P)|\mathbf{y}} ds_2^P ds_1^P \tag{2.17}$$

where $C_{(s_1^P, s_2^P)|\mathbf{y}}$ is given by equation (2.14) for any two points $s_1^P$ and $s_2^P$. For a set of areal

Table 2.2: Summary statistics for *in situ* soil moisture [1].

| Day of year | Min | $10^{th}$ %ile | Mean | $90^{th}$ %ile | Max | SD | No.of data points |
|---|---|---|---|---|---|---|---|
| DOY 167 | 0.084 | 0.137 | 0.274 | 0.425 | 0.629 | 0.112 | 735 |
| DOY 174 | 0.063 | 0.144 | 0.301 | 0.475 | 0.640 | 0.128 | 827 |
| DOY 177 | 0.019 | 0.097 | 0.240 | 0.392 | 0.642 | 0.115 | 864 |
| DOY 181 | 0.003 | 0.060 | 0.165 | 0.287 | 0.574 | 0.095 | 558 |
| DOY 185 | 0.006 | 0.051 | 0.144 | 0.236 | 0.518 | 0.089 | 606 |
| DOY 187 | 0.045 | 0.122 | 0.214 | 0.309 | 0.551 | 0.081 | 860 |
| DOY 190 | 0.029 | 0.094 | 0.213 | 0.335 | 0.499 | 0.092 | 827 |
| DOY 192 | 0.019 | 0.072 | 0.169 | 0.262 | 0.517 | 0.080 | 805 |
| DOY 195 | 0.029 | 0.082 | 0.188 | 0.314 | 0.669 | 0.095 | 832 |
| DOY 196 | 0.025 | 0.078 | 0.167 | 0.254 | 0.545 | 0.081 | 807 |
| DOY 199 | 0.054 | 0.133 | 0.245 | 0.370 | 0.547 | 0.091 | 789 |

pixels $\mathcal{S}^{ar} = \{A_1^P, ..., A_{n^{ar}}^P\}$, with $A_i^P \subset \mathcal{D}$, the joint distribution of $\mathbf{y}(\mathcal{S}^{ar})|\mathbf{y}$ is multivariate normal: $\mathbf{y}(\mathcal{S}^{ar})|\mathbf{y} \sim \mathcal{N}_{n^{ar}}(\boldsymbol{\mu}^{|A|}_{\mathcal{S}^{ar}|\mathbf{y}}, \boldsymbol{C}^{|A|}_{\mathcal{S}^{ar}|\mathbf{y}})$, where the elements of $\boldsymbol{\mu}^{|A|}_{\mathcal{S}^{ar}|\mathbf{y}}$ and $\boldsymbol{C}^{|A|}_{\mathcal{S}^{ar}|\mathbf{y}}$ are given by equations (2.16) and (2.17) [65]. Therefore, if we can evaluate the integrals in these equations, we can find the distribution of the PPD for pixels $\mathcal{S}^{ar}$. The integrals in equation (2.16) and (2.17) are generally not available in closed form. In such situations, we adopt a numerical approximation by assuming a fine numerical grid $\mathcal{S}^P$ over the entire region over which we want to make predictions and replacing the integrals with summations [65]. If $p_j^{A_i^P}$ for $j = 1, ..., n_{A_i^P}$ represents the grid points falling in pixel $A_i^P$, then the posterior mean for pixel $A_1^P$ and the posterior covariance for pixels $A_1^P$ and $A_2^P$ are respectively given by:

$$\mu^{|A|}_{A_1^P|\mathbf{y}} \approx \frac{1}{n_{A_1^P}} \sum_{j=1}^{n_{A_1^P}} \mu_{p_j^{A_1^P}|\mathbf{y}} \tag{2.18}$$

$$C^{|A|}_{(A_1^P, A_2^P)|\mathbf{y}} \approx \frac{1}{n_{A_1^P} \times n_{A_2^P}} \sum_{j_1=1}^{n_{A_1^P}} \sum_{j_2=1}^{n_{A_2^P}} C_{(p_{j_1}^{A_1^P}, p_{j_2}^{A_2^P})|\mathbf{y}} \tag{2.19}$$

The posterior variance of any pixel $A_1^P$, then becomes $C^{|A|}_{(A_1^P, A_1^P)|\mathbf{y}}$.

## 2.6 Results and Discussion

For the SMAPVEX12 campaign, out of the 15 days, we select 11 days on which no rainfall occurred during sampling. We also leave out DOY 201 due to insufficient number of *in situ* data points. For the *in-situ* data, we average the three replicate readings measured at each point. Table 2.2 gives the summary statistics of the SM distribution for each day. We standardize the SM data for each day prior to our analysis, so that the effect of the controls on the variance/correlation structure of SM can be compared for different wetness conditions.

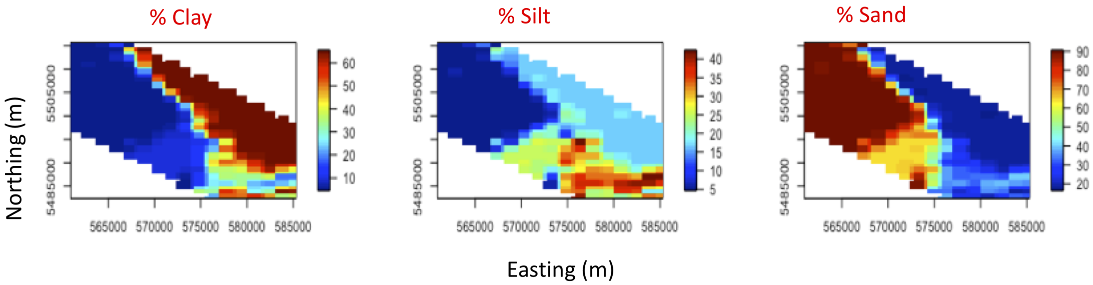

Figure 2.4: Spatial variation of soil texture in the study domain. There is a sharp change in soil texture from north-west to south-east [1].

Soil texture (Figure 2.4) primarily comprises of three variables: percent sand, percent clay and percent silt. Out of these, we select percent clay and percent silt for our analysis. Our study domain has a gentle rolling topography with less than 2% change in elevation. The elevation also has a high correlation with percent clay ($\sim 0.7$). We, therefore, exclude elevation from the analysis to avoid multi-collinearity. As the MODIS derived LAI is available at an interval of four days, we assume a linear change in LAI for the intervening days. The resulting LAI plots are given in Figure 2.5. We also standardize each of the surface controls by their mean and standard deviations so that they have the same range and scale. This is paramount in order to compare the importance of the individual controls in our study. The daily cumulative precipitation during the campaign recorded by 7 rain gauges lying within 5 km of the study domain is given in Figure 2.6.

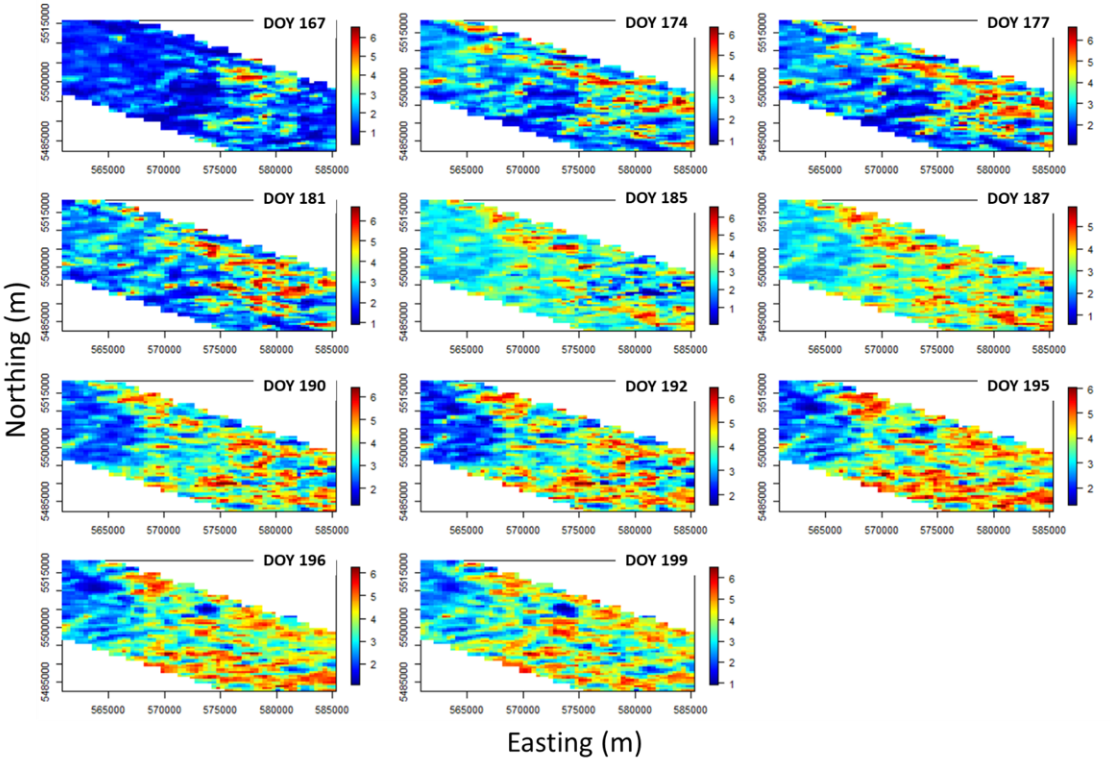

Figure 2.5: Evolution of vegetation (Leaf Area index) during the campaign. The crops reach full maturity by the end of the campaign depicted by high LAI values at the later stages [1].

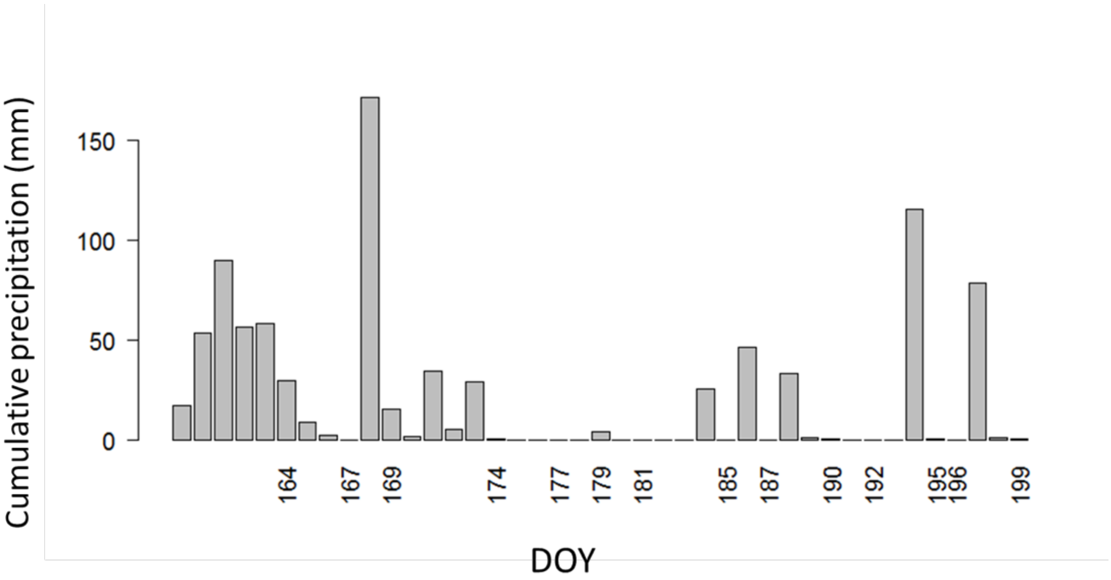

Figure 2.6: Bar plot representing the daily cumulative rainfall in the study domain [1].

### 2.6.1 Exploratory variogram analysis

Though there are multiple ways to choose the subregions, we divide our study area into three equal subregions from west to east, since both soil and LAI vary in that direction (Figures 2.4 and 2.5). We chose to take three subregions as it ensured that every bin of each of the empirical variograms had at least 30 observations. We subtract the mean $\boldsymbol{X}^T\boldsymbol{\beta}$ prior to analyzing the variograms. The variogram plots for each day are given in Figure 2.8. We find that for each of the days, the variogram for subregion 1 is significantly different from the other two subregions. Hence, there is an indication of non-stationarity in the variance/correlation structure. This can be attributed to the fact that subregion 1 is a predominantly sandy area and also has lower values of LAI than the other two regions (Figures 2.4 and 2.5). These two surface controls affect the SM distribution resulting in different empirical variograms. As the spatial structure of the other two subregions is similar to each other, we explore for M=1 (isotropic case), 2 and 3 in equation (2.7) and model the SM distribution as a weighted mixture of independent Gaussian processes with isotropic covariance functions. As stated earlier, though variograms are sub-optimal for formal parameter estimation, they can prove to be useful tools for selecting the structure of the parametric stationary/non-stationary model estimated by the ML method.

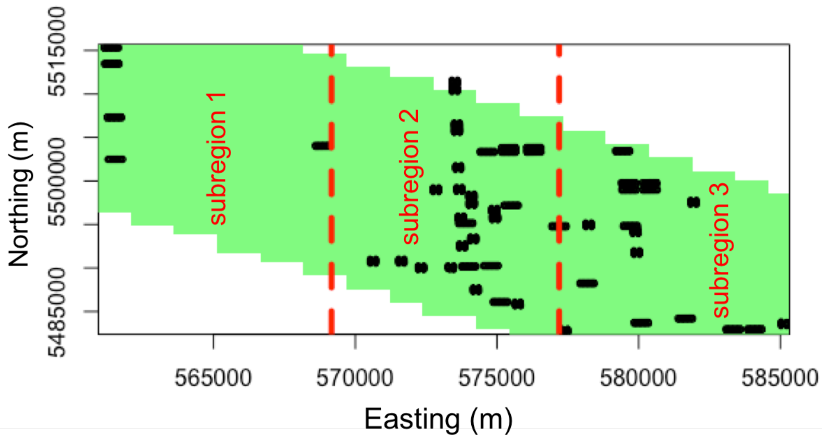

Figure 2.7: Subregions for exploratory variogram analysis. The black rectangles represent the agricultural fields in which sampling was carried out [1].

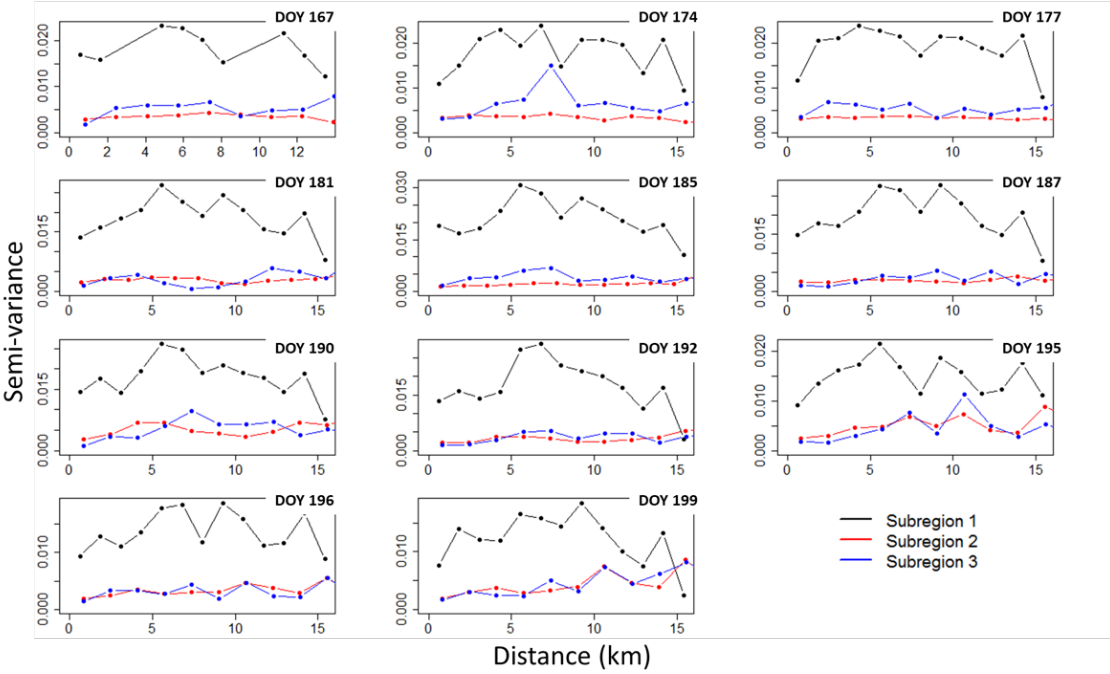

Figure 2.8: Empirical variograms of different subregions after accounting for the mean wetness conditions. There are 2 distinct variograms for each of the 11 days indicating the existence of non-stationarity in the variance and correlation structure of soil moisture in the study domain [1].

### 2.6.2 Parameter estimation

We fit the model in equation (2.1) on *in situ* SM data for individual days using ML estimation as outlined in Section 2.4 for M=1 (single isotropic covariance function) and M=2 and 3 (non-stationary covariance functions). As mentioned earlier, for our surface controls $\boldsymbol{X}$, we use LAI, percent clay and percent silt for both the mean $\boldsymbol{X}^T\boldsymbol{\beta}$, and the covariance function $C$ (equation (2.7)). We do not include higher order polynomial terms for $\boldsymbol{X}$ in $\boldsymbol{X}^T\boldsymbol{\beta}$ as it is not recommended to take complex mean structures (unless there is a scientific justification) by just looking at the data as data with strong spatial correlation frequently appear to have trends [75]. Following Section 2.5.2, for M=1 (single isotropic covariance function), M=2 and M=3, we estimate 7, 14 and 21 parameters, respectively. We maximize the profile log-likelihood (equation (2.4)) for each of the three models and find the ML estimates for $\hat{\boldsymbol{\theta}}$. We then substitute the values of $\hat{\boldsymbol{\theta}}$ in equation (2.5) to find $\hat{\boldsymbol{\beta}}$. Due to the large number of parameters (in $\hat{\boldsymbol{\theta}}$) to be estimated for M=2 (10) and M=3 (17)

Table 2.3: BIC comparison for isotropic (M=1) and non stationary models (M=2 and M=3) [1].

| | without $\tau^2$ | | | with $\tau^2$ | | |
|---|---|---|---|---|---|---|
| | M=1 | M=2 | M=3 | M=1 | M=2 | M=3 |
| Parameters | 7 | 14 | 21 | 8 | 15 | 22 |
| DOY 167 | -164.97 | -562.09 | -532.36 | -160.16 | -560.96 | -534.08 |
| DOY 174 | -380.66 | -716.23 | -688.19 | -376.90 | -712.14 | -681.49 |
| DOY 177 | -189.02 | -501.24 | -480.73 | -182.88 | -496.42 | -473.97 |
| DOY 181 | 192.99 | -197.98 | -184.52 | 199.31 | -192.16 | -178.19 |
| DOY 185 | 317.37 | -145.99 | -131.92 | 322.78 | -144.31 | -125.51 |
| DOY 187 | 358.25 | -130.06 | -129.26 | 363.86 | -125.82 | -122.48 |
| DOY 190 | 135.44 | -347.76 | -355.61 | 138.60 | -344.15 | -352.77 |
| DOY 192 | 223.1 | -253.11 | -243.29 | 226.46 | -247.02 | -236.6 |
| DOY 195 | -61.49 | -559.57 | -534.9 | -72.38 | -569.31 | -553.82 |
| DOY 196 | 166.87 | -300.32 | -264.95 | 168.21 | -303.15 | -268.84 |
| DOY 199 | -127.55 | -459.11 | -457.65 | -135.61 | -453.61 | -450.96 |

models, the convergence to the global optimum was very slow and took a fairly large number of iterations for both GenSA and Genoud. Another reason for slow convergence is that the parameters in equation (2.7) can switch between each other with no change in the covariance function C. e.g., switching $[\boldsymbol{\alpha_1}, \nu_1, \lambda_1, \sigma_1^2]$ with $[\boldsymbol{\alpha_2}, \nu_2, \lambda_2, \sigma_2^2]$ in equation (2.7) will not change C [64]. Since the weights in equation (2.7) are relative to each other and the *in situ* data is standardized, we found that bounding the values of individual components of $\alpha_j$ between -20 to 20, $\nu_j$ and $\sigma_j^2$ between 0 to 1.5, and $\lambda_j$ between 0 to 1000 km led to relatively faster convergence times without much loss in accuracy.

### 2.6.3 Model selection

Since the three models have different number of parameters, to choose between the models, we penalize the maximum log-likelihood value (equation (2.3)) for each of the models based on the number of parameters following the Schwartz or Bayesian Information Criterion (BIC) given as:

$$BIC = -2 \times l(\boldsymbol{\theta}) + log(n) \times q \tag{2.20}$$

where $q$ is the number of parameters in the model and n are the number of observations. Since

Table 2.4: $\hat{\beta}$ values for the mean SM for M=2 [1].

| Day of year | Intercept | LAI | % Clay | % Silt |
|---|---|---|---|---|
| DOY 167 | 0.36 | 0.01 | 0.10 | 0.05 |
| DOY 174 | -0.03 | 0.06 | 0.13 | 0.01 |
| DOY 177 | -0.05 | 0.11 | 0.08 | -0.02 |
| DOY 181 | 0.07 | -0.01 | 0.00 | 0.04 |
| DOY 185 | 0.00 | -0.05 | 0.21 | 0.14 |
| DOY 187 | -0.01 | 0.00 | 0.16 | 0.07 |
| DOY 190 | 0.16 | 0.05 | 0.15 | 0.04 |
| DOY 192 | 0.39 | 0.10 | 0.02 | 0.08 |
| DOY 195 | 0.07 | 0.02 | 0.11 | 0.01 |
| DOY 196 | 0.05 | 0.04 | 0.14 | 0.14 |
| DOY 199 | 0.58 | 0.04 | 0.27 | 0.15 |

the BIC includes the negative log-likelihood, lower values of BIC are preferred in model selection. The BIC selects the model with the highest posterior model probability [82] and is a rough approximation of the Bayes factor widely used in hypothesis testing [83]. The values of BIC for the three models for each of the 11 days are given in Table (2.3). We select the model with more parameters only if the evidence against selecting a model with fewer parameters is very strong or the difference in BIC between the two models is greater than 10 [83]. This leads us to always select the model with M=2 for all days, which is also consistent with the exploratory variogram analysis (Figure 2.8). We also include the nugget variance term, $\tau^2$ (Section 2.4), for each of three models and estimate it along with parameters in $\hat{\boldsymbol{\theta}}$. We calculate the BIC for each of the three models, but all of them are also rejected in favor of M=2 (without measurement error) using the criterion above. As we averaged the three replicate readings for each *in situ* observation, the measurement error becomes insignificant. Also note that the difference in BIC values between M=1 and M=2 is quite large for each day, implying an overwhelming evidence of non-stationarity. Since we assume the same mean structure ($\boldsymbol{X}^T\boldsymbol{\beta}$) for all three models, we conclude that there is a spatial non-stationarity in the variance/correlation structure of SM driven by surface controls.

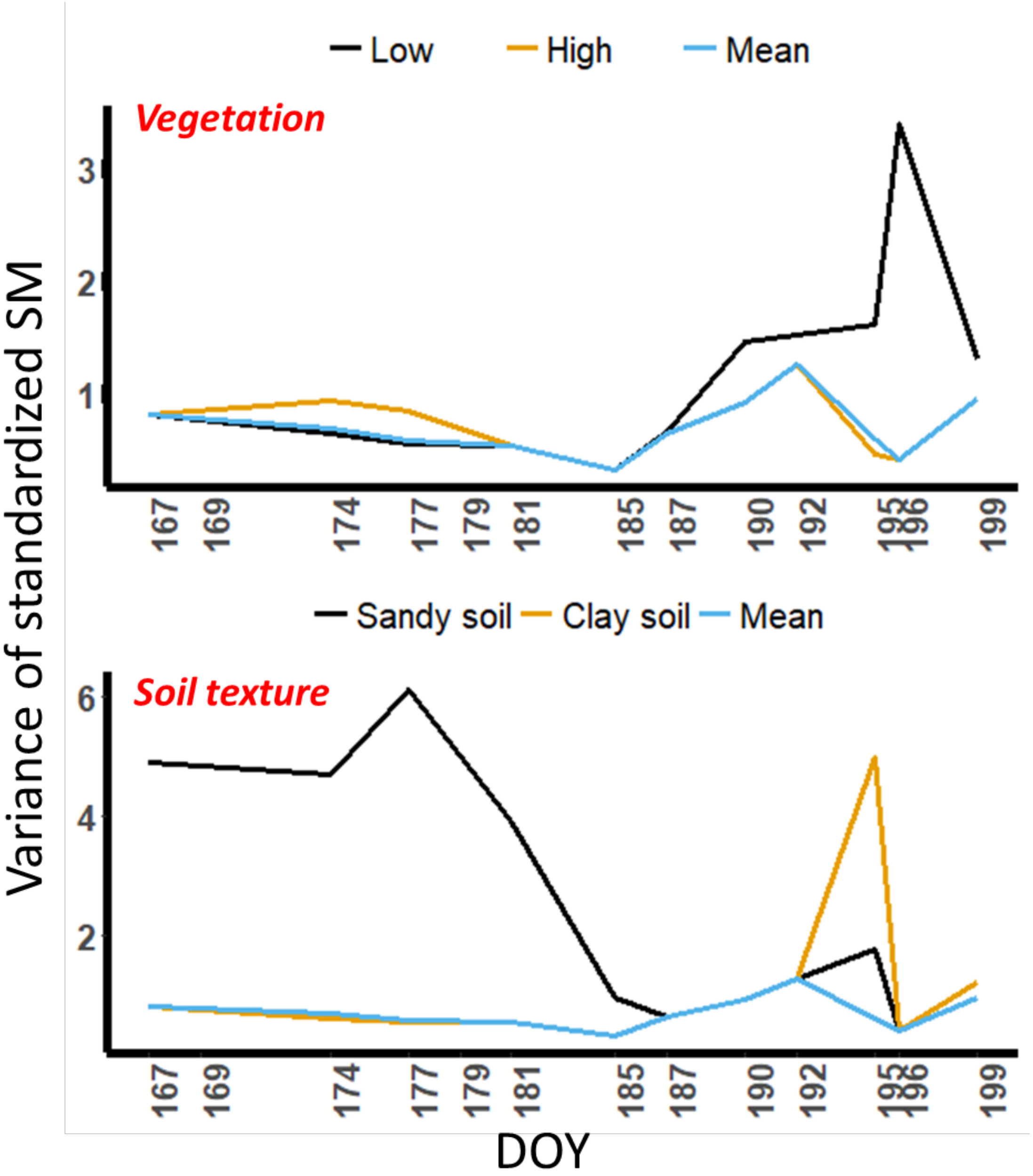

Figure 2.9: Effect of vegetation and soil texture on soil moisture variance. Soil texture causes a significant spatial non-stationarity in soil moisture variance during the initial phase of the crop growing stage with sandy soils exhibiting more variance in general. The effect of vegetation on soil moisture variance is dominant as the crops reach maturity [1].

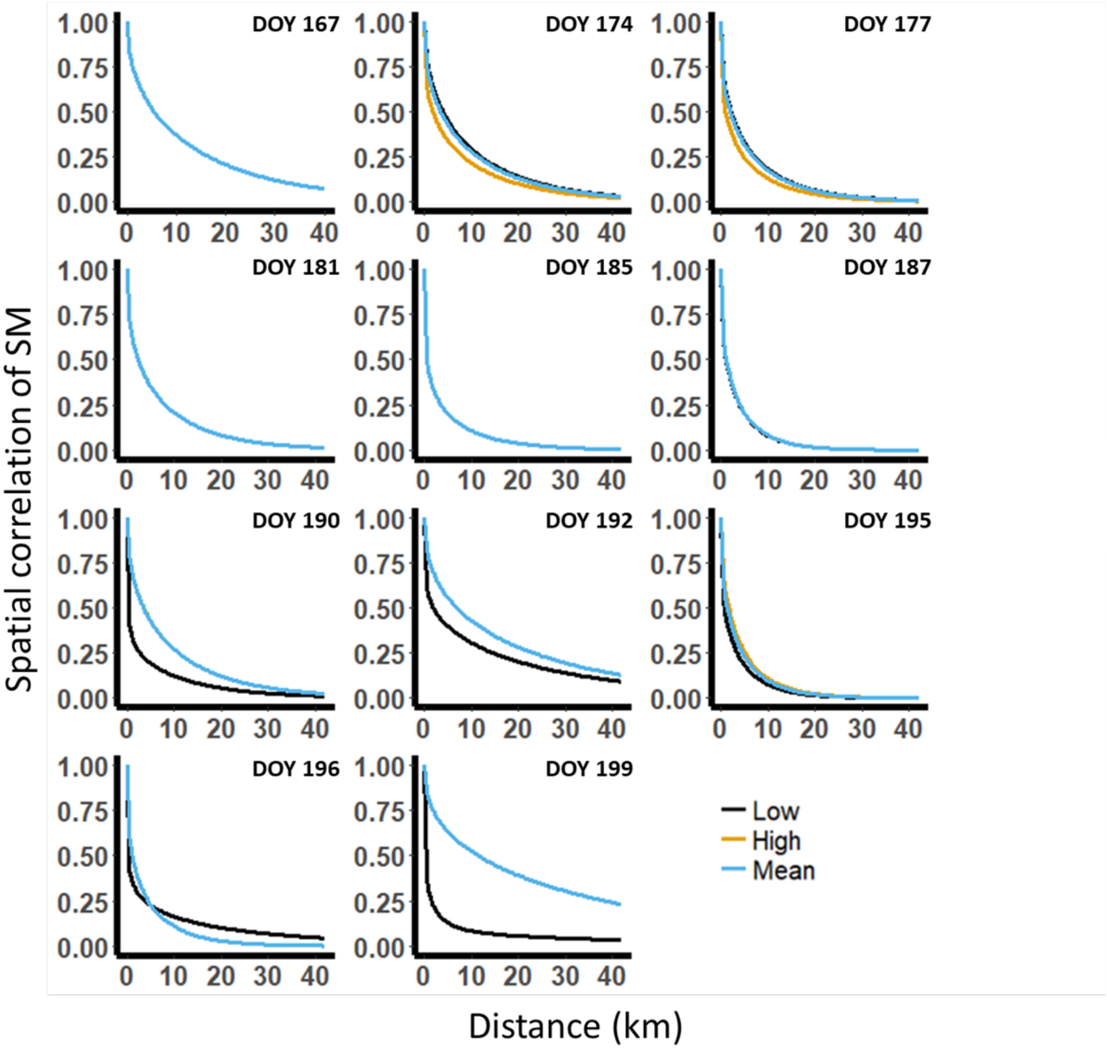

Figure 2.10: Effect of vegetation on correlation structure of soil moisture [1].

### 2.6.4 Effect of surface controls on the mean of SM

For M=2, the $\hat{\boldsymbol{\beta}}$ values for each day are given in Table 2.4. Since the controls are standardized, a higher absolute value of $\hat{\beta}_j$ denotes a stronger effect of the $j^{th}$ control on the mean SM. We find that the $\hat{\beta}$ values for LAI are insignificant for all the days with the highest value of 0.11 on DOY 177. For percent clay, $\hat{\beta}$ values for all the days are positive, as higher clay content generally results in more water holding capacity of soil. The highest $\hat{\beta}$ value for clay occurs on DOY 199 (0.27). The effect of percent silt on mean SM is also insignificant on most of the days with a maximum

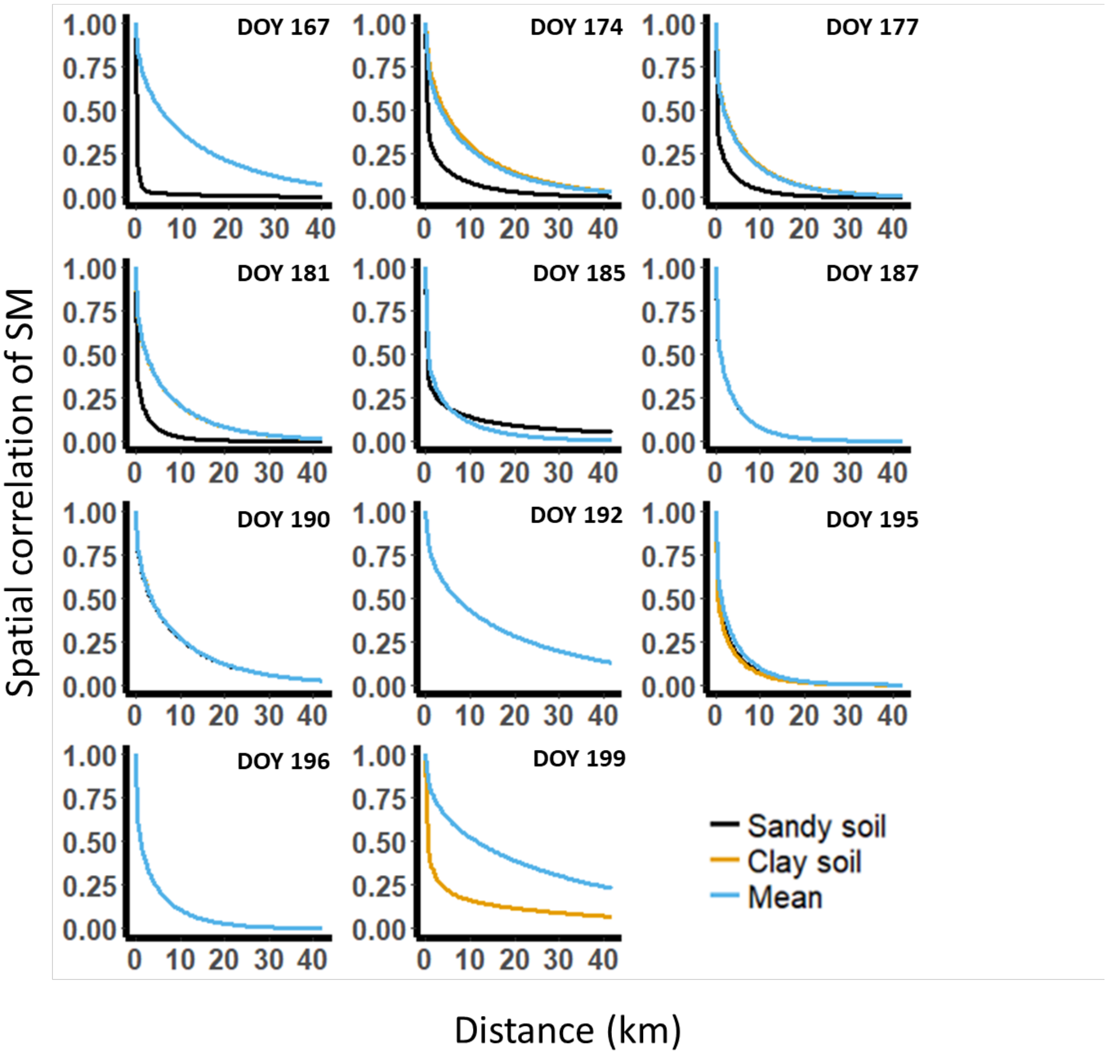

Figure 2.11: Effect of soil texture on correlation structure of soil moisture [1].

value of 0.15 on DOY 199.

### 2.6.5 Effect of surface controls on the variance/correlation of SM

As mentioned before, the effect of the controls on the covariance of SM is not straightforward. For M =2, the effect of surface controls on SM variance/correlation structure is inferred by analyzing the change in the parametric covariance function for low ($10^{th}$ percentile), dominant (mean) and high ($90^{th}$ percentile) values of the surface controls as detailed in Section $2.5.3.2$.

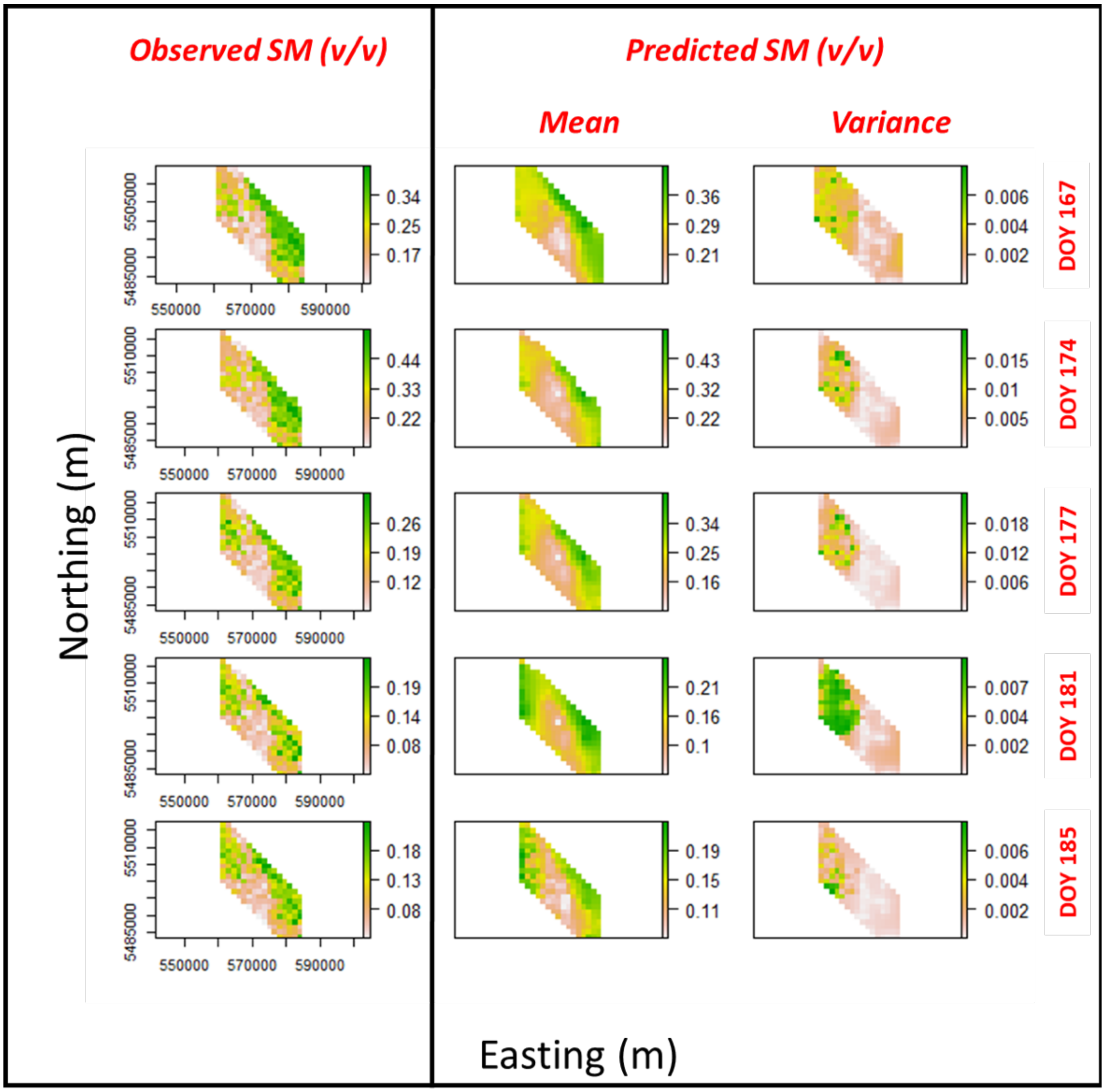

Figure 2.12: Comparison of PALS airborne soil moisture with soil moisture predictions (DOY 167 - DOY 185) using in situ data [1].

### *2.6.5.1 Vegetation*

The effect of the individual surface controls on the variance of SM is given in Figure 2.9. During the initial growing season, vegetation has a minimal effect on SM variance. But as the crops mature, as depicted by the increase in the mean LAI of the region, vegetation plays a dominant role in causing differences in SM variance across the region. The effect of vegetation on the spatial variance becomes apparent from DOY 190-199 having mean LAI values from 3.63-4.0 as opposed

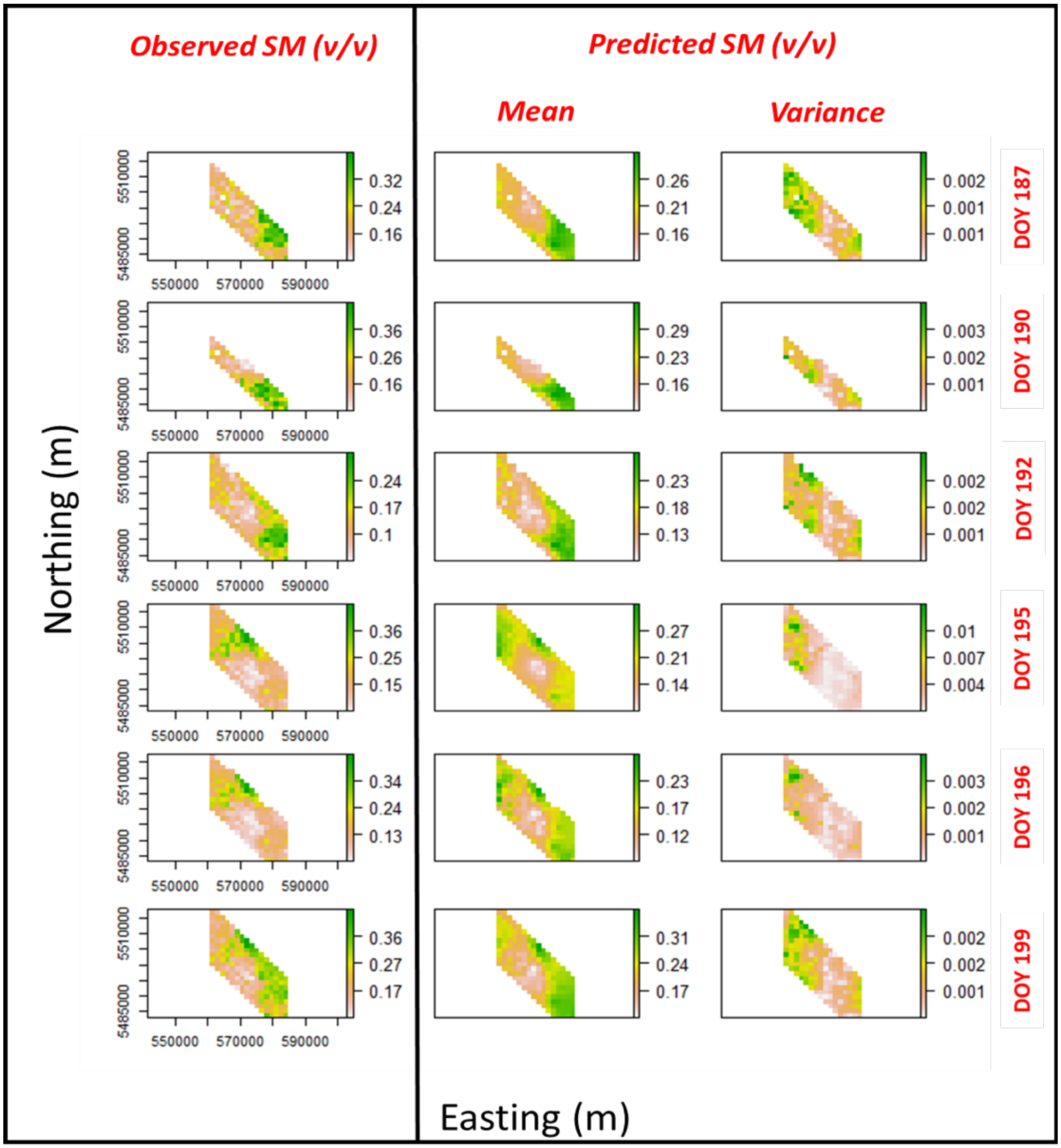

Figure 2.13: Comparison of PALS airborne soil moisture with soil moisture predictions (DOY 187 - DOY 199) using in situ data [1].

to the previous days which have mean LAI values less than 3.25. In particular, we find that areas with low LAI (1.97-2.65) have more SM variance than regions with dominant (3.63-4.0) and high (5-5.3) LAI contents. In fact, the presence of vegetation acts to reduce spatial variance, though the difference in regions with dominant and high LAI contents is minimal. Areas with low LAI values though have a much larger variance. We find that as the growing season progresses, the difference in SM variance between low and high vegetated areas increases until DOY 196. Two precipitation events occur on DOY 194 and DOY 197 leading to an increase in the mean SM content of 0.02 v/v and 0.08 v/v respectively. As a result, on DOY 195 and DOY 196, there is a decrease in the variance of SM in vegetated areas as opposed to low vegetated regions, which show an increase in SM variance. The precipitation event on DOY 197 on the other hand causes a sharp rise in the mean SM diminishing the effect of vegetation on SM variance.

Similar to variance, vegetation exhibits a signature on the correlation structure of SM beginning DOY 190 (Figure 2.10) promoting spatial correlation in SM in wetter conditions. The effect of vegetation on SM correlation decreases (Figure 2.10) as the soil undergoes drying from DOY 190 (Mean SM = 0.21 v/v) to DOY 196 (Mean SM = 0.16 v/v). There is an increase in the mean SM on DOY 199 (Mean SM = 0.24 v/v) due to a precipitation event again causing a difference in the spatial correlation of SM.

We therefore conclude that vegetation plays a dominant role in reducing SM variance and promoting spatial correlation in SM distribution once the crops begin to mature, though its effect diminishes when there is a sharp increase in the SM content causing rainfall to be the dominant factor.

#### *2.6.5.2 Soil texture*

We classify soil texture as sandy (low silt, low clay), dominant soil texture (mean silt, mean clay) and clay soil (high clay, low silt). Soil texture plays a crucial role in SM distribution at the beginning of the crop growing season from DOY 167 to DOY 185 (Figures 2.9 and 2.11). We find that sandy soils exhibit larger spatial variance (Figure 2.9) than clay and dominant soil texture conditions. Regular rainfall showers occur from DOY 168 to DOY 173 (Figure 2.6) followed by a drying cycle until DOY 185 with slight rainfall on DOY 179 and DOY 184. The mean SM content during this drying period falls from 0.30 to 0.14 v/v. For sandy soils, there is a slight increase followed by a steep decrease in variability as drying occurs. For clay and dominant soil texture conditions, the change in the variance is not significant throughout the drying cycle. We therefore infer that, at the beginning of the crop season, SM variance in sandy soils, though significant in wet conditions, decreases as the soil dries. For dry conditions, soil texture has a minimal influence on SM variance. After DOY 185, vegetation governs SM variance. Though there is a spike on DOY 195, it can be attributed to the localized extreme precipitation event on DOY 194. For the correlation structure (Figure 2.11), there is a difference in the correlation between sandy and clay/dominant soil texture till DOY 185. Sandy soils, in general, are less correlated than clay and dominant soil texture soils but as drying occurs, the correlation increases. We therefore conclude that soil texture plays a dominant role on SM variance and correlation in the drying cycle (in the absence of significant vegetation) with a decrease in variance and increase in correlation as the soil becomes drier.

Table 2.5: RMSE and $R^2$ values between predicted and observed airborne soil moisture [1].

| Day of year | RMSE | $R^2$ |
|---|---|---|
| DOY 167 | 0.067 | 0.51 |
| DOY 174 | 0.081 | 0.49 |
| DOY 177 | 0.053 | 0.55 |
| DOY 181 | 0.045 | 0.34 |
| DOY 185 | 0.041 | 0.37 |
| DOY 187 | 0.060 | 0.45 |
| DOY 190 | 0.066 | 0.64 |
| DOY 192 | 0.041 | 0.59 |
| DOY 195 | 0.069 | 0.38 |
| DOY 196 | 0.069 | 0.24 |
| DOY 199 | 0.064 | 0.43 |

### 2.6.6 Upscaling SM predictions to airborne scale

We upscale our point scale predictions to the airborne support ($\sim$1500 m) using the steps outlined in Section 2.5.4.2 and validate the predictions using PALS SM. Note that the PALS data were only used to validate the predictions at the airborne support and were not used in parameter estimation, We assume a fine-scale grid over the entire domain such that the distance between the adjacent points on the grid is 200 m. The mean and variance of the PPD at the airborne support for each day are given in Figures 2.12 and 2.13. Note that the variance of the PPD is, in general, higher for the north-west region of the domain due to fewer *in situ* SM observations (Figure 2.7). Since the SM from airborne retrievals is typically subject to bias [84], we subtract the average of

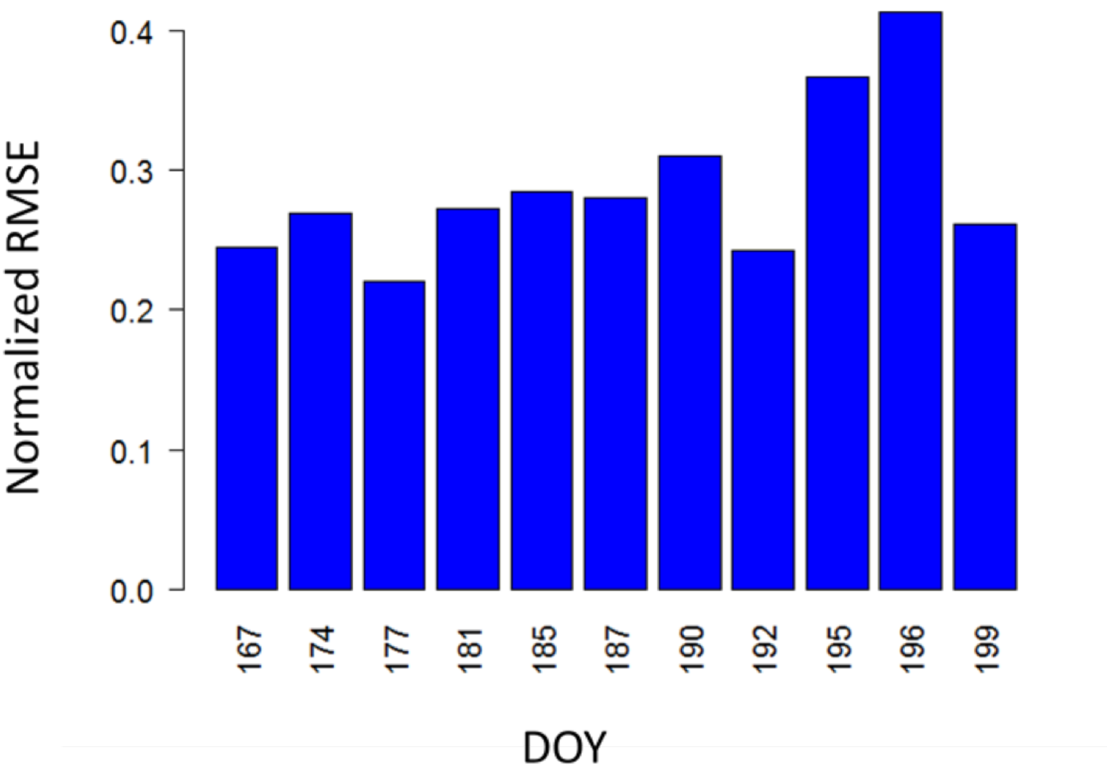

Figure 2.14: Comparison of normalized RMSE between predicted and observed airborne soil moisture [1].

the mean of the PPD with the mean of the PALS airborne SM for all pixels. If the total number of pixels for a given day is $n_{air}$, then for any airborne pixel $A_i^P$:

$$\mu^{1500}_{A_i^P|\mathbf{y}}(bc) = \mu^{1500}_{A_i^P|\mathbf{y}} - \sum_{j=1}^{n_{Air}} SM^{PALS}_{A_j^P} + \sum_{j=1}^{n_{Air}} \mu^{1500}_{A_j^P|\mathbf{y}} \tag{2.21}$$

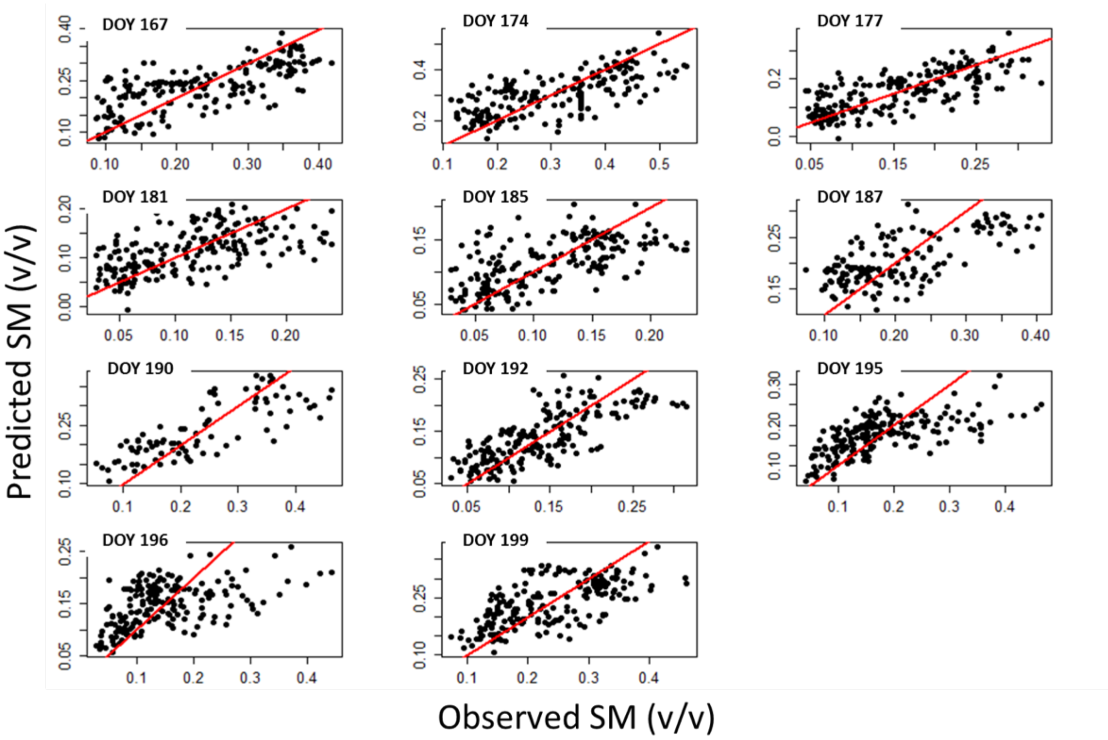

Figure 2.15: Pixel by pixel comparison of observed and predicted soil moisture at airborne scale [1].

where $\mu^{1500}_{A_i^P|\mathbf{y}}$, $SM^{PALS}_{A_i^P}$ and $\mu^{1500}_{A_i^P|\mathbf{y}}(bc)$ are respectively the mean of the PPD, the observed PALS SM and the bias-corrected mean of the PPD for the airborne pixel $A_i^P$.

The root mean square error ($RMSE$) and coefficient of determination ($R^2$) values for mean of the PPD and the observed PALS SM for individual days are given in Table 2.5. The $RMSE$ values lie between 0.041 to 0.081. To compare $RMSE$s for different wetness conditions, we find the normalized $RMSE$ ($nRMSE$) given by:

$$nRMSE = \frac{RMSE}{\sum_{j=1}^{n_{Air}} SM^{PALS}_{A_j^P}} \tag{2.22}$$

The $nRMSE$ values for each day are given in Figure 2.14. DOY 195 and DOY 196 in particular, have a higher $nRMSE$. We also perform a pixel by pixel comparison (Figure 2.15). The $R^2$ values lie between 0.24 to 0.64. In addition, we also plot empirical cumulative distribution functions ($ecdf$) for both observed PALS SM and mean of PPD in Figure 2.16. Analyzing both

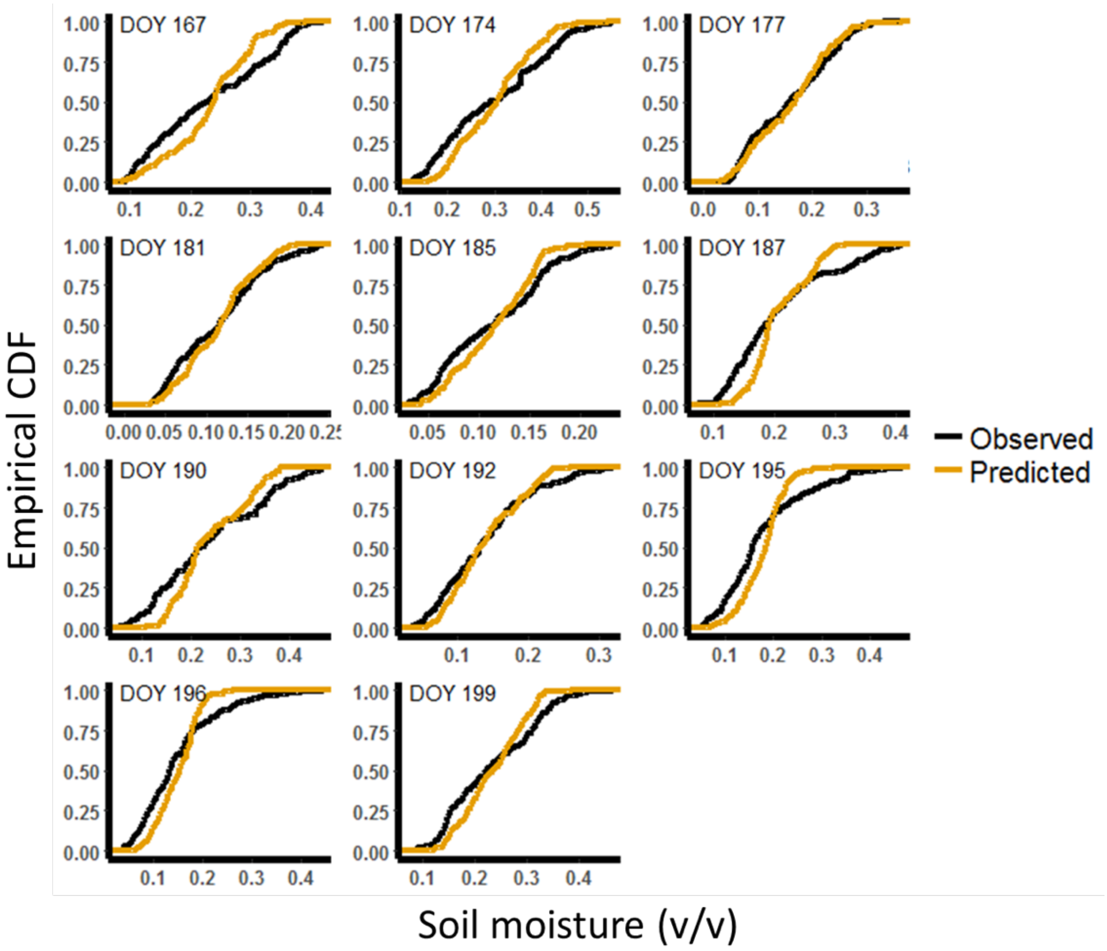

Figure 2.16: Comparison of empirical cumulative distribution functions for observed and predicted soil moisture at airborne scale [1].

plots, we find that, in general, our mean predictions agree well with the observed SM patterns.

We find discrepancies between the observed and predicted SM on DOY 187, DOY 195 and DOY 196. In particular, we find that the predicted SM under-estimates the SM for wet pixels (Figure 2.15). The same can be seen in Figure 2.16 when comparing the $ecdf$s. To explain this, we analyze the precipitation events occurring on the antecedent days for each day of analysis. We interpolate the daily cumulative rainfall recorded by the rain gauges to the entire domain by inverse distance weighting method. The resulting rainfall patterns are given in Figure 2.17. We find that extreme localized precipitation on DOY 186 and 194 may result in wetter pixels on the succeeding days which are not captured by the surface control driven model. Due to the sparse number of rain

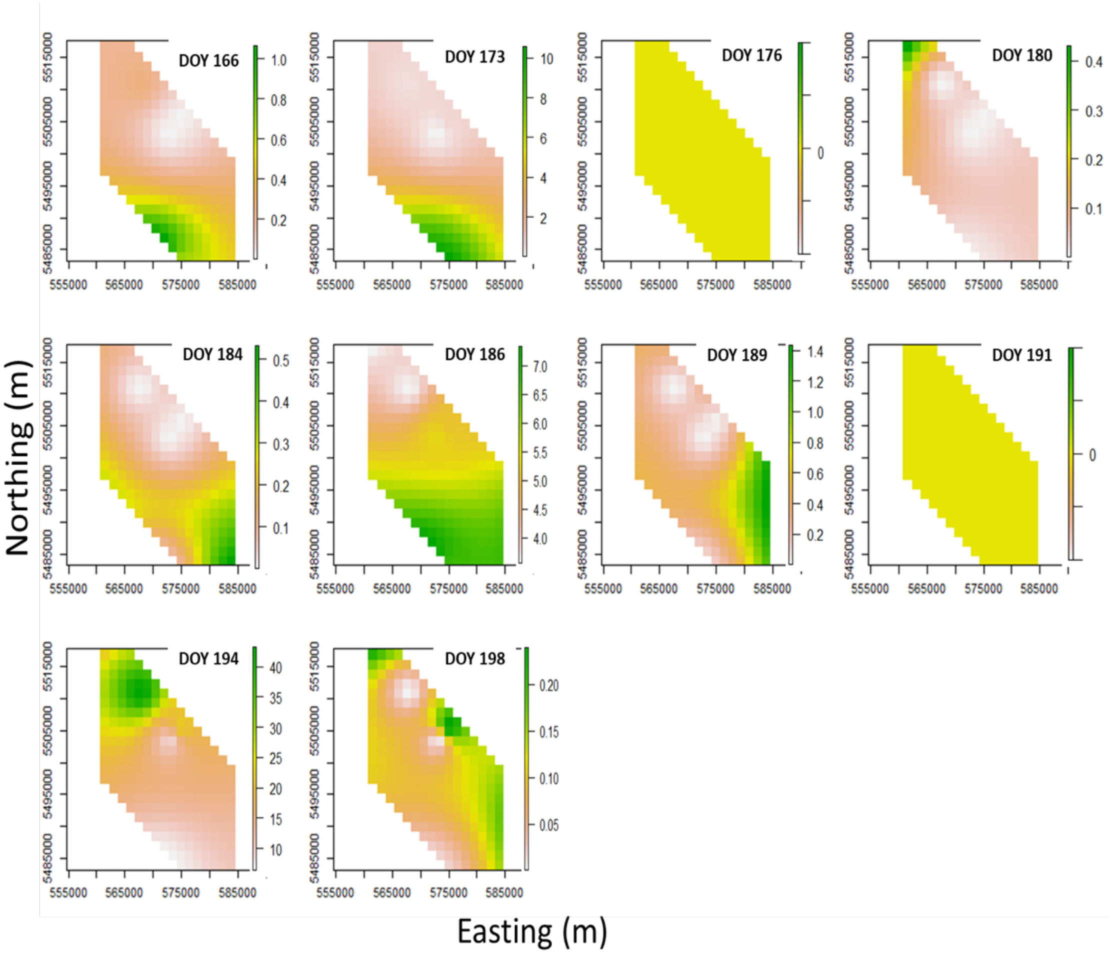

Figure 2.17: Antecedent rainfall (mm) for each day [1].

gauges in the study domain, we were unable to fully account for the effect of rainfall especially if it was patchy/localized on some of the days and we could only make qualitative arguments as the ones above. The absence of precipitation as a covariate in our model might contribute to the high RMSE values on some days in Table 2.5. Also note that, our interpretation of the effects of the surface controls on the SM variance/correlation might be affected if there was patchy rainfall in the domain, especially for DOY 167 and DOY 174 which were preceded by a number of rainy days. In study regions, where the spatial distribution of rainfall can be approximated reasonably accurately, rainfall should also be considered as a control in the vector $\boldsymbol{X}$ in equations (2.1) and (2.7) to improve predictions.

For study regions, where the number of *in situ* SM data points are moderate, fitting the non-stationary model can be challenging because of the high number of parameters to be estimated. In such situations we recommend to either use dimension reduction techniques on the controls (such as Principal Component Analysis) prior to geostatistical modeling or use only the dominant control [45] in the region. We also recommend using one parameter correlation functions such as exponential, spherical, etc. if it gives satisfactory predictions. As an example, in the current analysis, for M = 2, if we took the first Principal Component of LAI, clay and silt for both mean and the covariance, and used exponential correlation functions for the two $C_j$ in equation (2.7), then the number of parameters to be estimated reduces from 14 to 8. Since convergence using the optimization algorithms (GenSA and Genoud) was very slow, algorithms with faster convergence times should also be explored. Lastly, improved formulations for controls driven non-stationary covariance functions should be developed as they have the potential to improve soil moisture predictions across spatial scales.

## 2.7 Conclusions

SM is an important driver for both land and atmospheric processes, and its correct representation in land surface models and remote sensing retrievals is essential for better understanding of the water and energy cycles and reducing biases in model predictions. SM exhibits considerable non-stationary behavior at sub-grid scales, and stationary geostatistical methods fail to capture this behavior. Since this is mainly caused by heterogeneity in surface characteristics, we use a spatial model with a flexible covariance function depending on the local surface characteristics. Using the BIC, we show that the non-stationary model fits significantly better than the stationary counterpart. We examine the effects of individual surface controls on both the variance and correlation structure of SM and find that vegetation and soil texture affects the variance/correlation at different stages of the crop growing cycle. The extent of the dominance of these controls is governed by wetness conditions and also on the stage of the drying cycle succeeding a rainfall event. While soil texture plays a dominant role in the beginning of the crop growing cycle, vegetation drives the SM variance/correlation once the crops begin to gain maturity. We then upscale our SM predictions to airborne support and find that our SM predictions can satisfactorily mimic the spatial patterns at the airborne scale. According to the authors' knowledge, this is the first study that explicitly accounts for the effects of surface controls on the variance/correlation of SM. The proposed framework is statistically optimal and can be used to predict and upscale SM in heterogeneous environments using *in situ* SM and surface controls data.

# 3. MULTISCALE DATA FUSION FOR SOIL MOISTURE ESTIMATION: A SPATIAL HIERARCHICAL APPROACH

## 3.1 Synopsis

[1]Soil moisture (SM) has been identified as a key climate variable governing hydrologic and atmospheric processes across multiple spatial scales at local, regional and global levels. The global burgeoning of SM datasets in the past decade holds a significant potential in improving our understanding of multi-scale SM dynamics. The primary issues that hinder the fusion of SM data from disparate instruments are 1) different spatial resolutions of the data instruments, 2) inherent spatial variability in SM caused due to atmospheric and land surface controls and 3) measurement errors caused due to imperfect retrievals of instruments. We present a data fusion scheme which takes all the above three factors into account using a Bayesian spatial hierarchical model (SHM) combining a geostatistical approach with a hierarchical model. The applicability of the fusion scheme is demonstrated by fusing point, airborne and satellite data for a watershed exhibiting high spatial variability in Manitoba, Canada. We demonstrate that the proposed data fusion scheme is adept at assimilating and predicting SM distribution across all three scales while accounting for potential measurement errors caused due to imperfect retrievals. Further validation of the algorithm is required in different hydroclimates and surface heterogeneity as well as for other data platforms for wider applicability.

## 3.2 Introduction

Soil moisture (SM) acts as a crucial driver in land-atmosphere interactions due to its coupling mechanisms with temperature [10, 7, 8], precipitation [85, 86] and evapotranspiration (ET) [87, 88]. Due to its importance from local to global scales, scientists have spent the last two decades measuring SM at various resolutions ranging from a few centimeters to hundreds of kilometers.

[1] Reprinted with permission from, ' Multiscale data fusion for surface soil moisture estimation: A spatial hierarchical approach' by Kathuria et al. (2019) Water Resour. Res., 55, 10443– 10465, https://doi.org/10.1029/2018WR024581, Copyright 2019 John Wiley and Sons, Inc.

We define an individual SM measurement using the scaling triplet: support, extent and spacing [25, 26]. Support refers to the representative area of the data sample, extent refers to the total coverage area of the data, and spacing refers to the distance between individual samples. *In situ* sensors, such as capacitance sensors and time domain reflectometers, measure SM at a point scale. Proximal sensing techniques, such as active-passive airborne sensors, cosmic-ray neutron probes and Global Positioning Systems (GPS), provide areal-averaged SM data at intermediate scales [27]. Satellite missions, such as Advanced Microwave Scanning Radiometer - Earth Observing System (AMSR-E), Soil Moisture Ocean Salinity (SMOS) and Soil Moisture Active Passive (SMAP), retrieve SM globally at a coarse resolution of 25 $\sim$ 40 km.

The spatial distribution of SM is sensitive to the heterogeneity in land surface controls and atmospheric forcings [12]. As a result, SM distribution usually exhibits non-stationary behavior in space in heterogeneous environments. [54, 89, 90]. From point to remote sensing (RS) scales, SM manifests both short and long range spatial dependence (or correlation) [91] with short range dependence (10-30 km) usually governed by land surface characteristics comprising soil texture, vegetation and topography [18, 17, 45, 15, 20, 14], while long range dependence (larger than 60 km) determined by precipitation [12, 16]. This inherent spatial correlation, however, has not yet been exploited by current state of the art downscaling and data fusion algorithms used to predict SM.

Retrievals from SM instruments are typically subject to systematic (bias) and stochastic (random) errors. The uncertainty in retrievals usually depends on atmospheric and land surface factors. Retrievals from airborne active sensors, for instance, are typically affected by the dominant soil texture and vegetation characteristics of the region [92, 93, 25, 94] while cosmic neutron probes are affected by atmospheric vapor, vegetation, lattice water and organic matter in the soil [27, 95]. Hence, it is essential for downscaling/fusion algorithms to account for this measurement uncertainty to ensure that incorrect inferences are not made.

Currently, there are three broad paradigms to predict SM across different scales [96]. 1) Data fusion of active/passive sensors [84, 97, 98]. 2) Downscaling a coarse SM pixel using ancilliary data such as vegetation, surface temperature, topography and soil texture [99, 56, 100] 3) Downscaling or data assimilation of SM data using a physical model [101, 59, 102]. The majority of the methods in the first two paradigms are empirical and specific to the study region, support of the data and the underlying geophysical heterogeneity [103]. Though the third approach, by contrast, utilizes a physical model to account for spatial dependence of SM across space (and time), predictions using data assimilation "typically bears signatures of the hydrologic forecast model, either as the background field or statistical prior." [104]. Further, data assimilation of RS retrievals using land surface models usually result in biased predictions due to the horizontal and vertical scale mismatch between the model and RS observations [87]. Data assimilation, therefore requires preprocessing of satellite observations to match the spatial resolution of the physical model [59, 105] and a correction for the bias between the satellite data and model simulations [60]. It should be noted, however, that physical models are essential to predict the deep root-zone soil moisture as most of the RS sensors only measure SM.

Though there can be different ways to describe the term "data fusion", in this paper, we adopt the definition of [28]: "Data fusion is the process of combining information from heterogeneous sources into a single composite picture of the relevant process, such that the composite picture is generally more accurate and complete than that derived from any single source alone." Data fusion for a given study domain, combines disparate data sources to improve our understanding of the SM distribution across multiple scales. In a way, data from different sensors can be considered as incomplete fragments of the multi-scale SM composite and data fusion attempts to obtain this holistic picture by fusing these fragments together.

The past few years have seen a proliferation of novel soil moisture sensors from point to RS scales such as Distributed Temperature Sensing (DTS), COsmic-ray Soil Moisture Observing System (COSMOS) and Global Navigation Satellite System-Reflectometry (GNSS-R). On a global scale, satellite missions such as SMAP and SMOS are measuring SM for the past few years with the National Aeronautics and Space Administration (NASA)-Indian Space Research Organization Synthetic Aperture Radar (NISAR) expected to be launched in 2021. The abundance of multiscale SM data presents a never-before opportunity for the scientific community to improve its understanding of multi-scale SM dynamics. There is, thus, a need for a data-driven data fusion scheme to combine this data with traditional point sensors which respects the change of support between the disparate data, accounts for uncertainty in retrievals while also exploiting the spatial dependence/correlation inherent in SM. The data fusion scheme should also be adept in predicting SM at the required support (upscaling/downscaling).

The objective of the present work, therefore, is to fuse SM data from multiple instruments (in an optimal way) while partitioning the inherent spatial dynamics of SM distribution and the measurement errors induced by different data instruments. To achieve our objective, we use a Spatial Hierarchical Model (SHM) combining conventional geostatistics with a hierarchical modeling (HM) paradigm. Traditionally, geostatistics has been used to model soil moisture only at individual scales [18, 46, 16], and in this paper, we use an underlying geostatistical model to fuse multi-scale SM data. While the HM paradigm enables coherent data fusion, the spatial dependence induced by a geostatistical approach allows us to make inferences on unknown locations and scales [106]. A SHM approach is motivated not just by getting a good fit to the data but also enables scientific interpretability. While data mining methods focus on the particular dataset being studied and tease out unusual observations (often not optimally), they are unable to explain what "unusual" means in a physical sense. A SHM approach, by contrast, accounts for these unusual observations (optimally) using an underlying latent spatial process answering "why" some observations behave unusually. The "why" question (rather than just getting a good fit to the data) is generally of more interest to the scientific community.

We present a fusion algorithm combining SM datasets from multiple instruments using ancillary data such as rainfall, soil texture, elevation and vegetation while accounting for the spatial non-stationary nature of SM, the measurement errors caused due to data instruments as well as the change of scale between individual data instruments. As a proof of concept, we apply the data fusion scheme to combine multi-scale SM data in the Red River watershed in Manitoba, Canada. The study area is chosen to test the applicability of the data fusion framework in a region with high heterogeneity in land surface controls and large variability in soil moisture conditions [70]. This paper is organized as follows. We describe the study region and the associated data in Section 3.3. The general SHM and data fusion scheme used in the present study are outlined in Section 3.4. The exact SHM used in the current study to fuse SM data is described in Section 3.5. This is followed by the discussion of results in Section 3.6 before we conclude in Section 3.7.

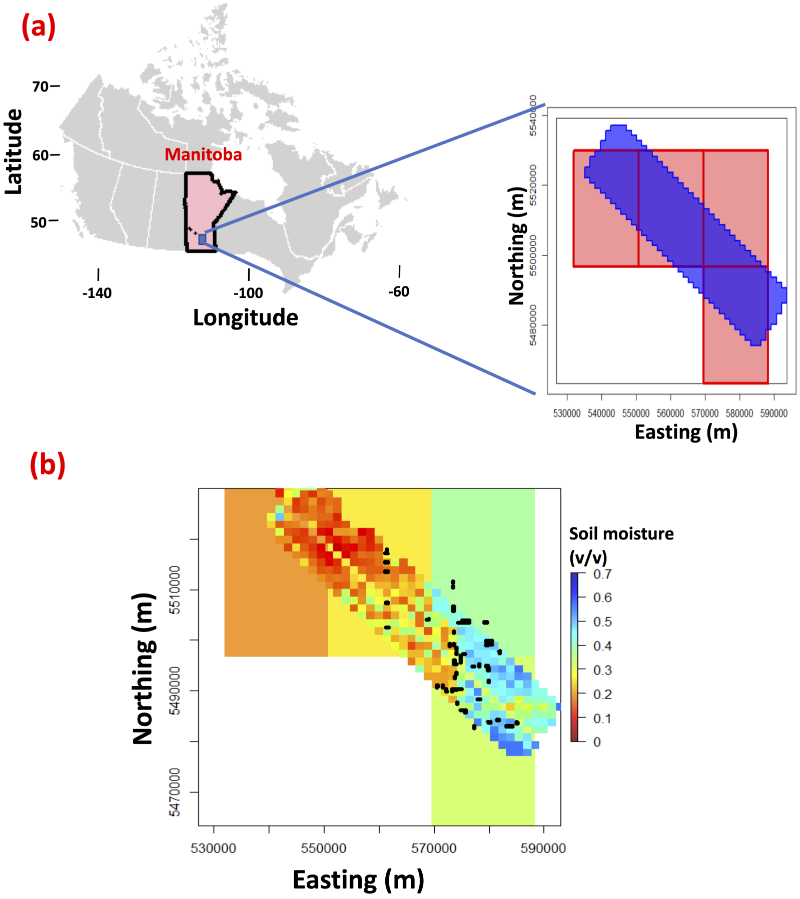

Figure 3.1: (a) Study site located in the Red River watershed in the province of Manitoba, Canada. The blue outline refers to the extent of the airborne soil moisture data retrieved during the Soil Moisture Active Passive Experiment 2012. The 4 red pixels denote the overlapping Soil Moisture Ocean Salinity data with the airborne extent. (b) Spatial variability in soil moisture for DOY 164. For most days, the soil moisture exhibits large variability from west to east. The diagonal strip refers to the airborne data retrieved from passive-active L-band sensor during the campaign, while the 4 coarse pixels represent the satellite data. The black dots refer to the agricultural fields in which *in situ* sampling was carried out at an average of 16 samples (spaced 75 m apart) per field [2].

## 3.3 Study area and data

We apply the data fusion scheme combining point, airborne and satellite SM data retrieved during the Soil Moisture Active Passive Experiment 2012 (SMAPVEX12) in the Red River watershed within Manitoba, Canada (Figure 3.1(a)). SMAPVEX12 was organized from June 6-July 17, 2012 in the agricultural lands south-west of Winnipeg, Manitoba which is classified as having a fully humid climate according to the Köppen-Geiger climate classification [69], with an average annual precipitation of 521 mm [94]. There is a sharp contrast of soil texture across the region, with high clay content in the east to fine loamy sands in the west. The majority of the land is covered with agricultural crops consisting of cereals, soybeans, canola and corn with wetlands and forest cover in the northwest. The campaign was carried out during the crop-growing season with low biomass contents at the start of the campaign ($< 0.1$ kg/m$^2$) while reaching biomass conditions of 1-2 kg/m$^2$ for bean crops and 4 kg/m$^2$ for corn in the final week. The *in situ* SM data was collected using Stevens Water Hydra (frequency-domain reflectometry) sensors and Delta-T Theta (impedance) probes at an average of 16 sampling points spaced 75 m apart in each field. Three replicate readings were taken at each location. On each flight day of the airborne sensor, on-site crews started collecting *in situ* measurements at 6:30 AM local time. For each flight day, one bulk-density core per field was collected to perform site-specific calibration of the sensors. The average root mean square error (RMSE) after developing the site-specific calibration equations was 0.037 v/v. The airborne data was retrieved at a spatial resolution of 1500 m using a passive-active L-band sensor (PALS) mounted on a DHC-6 Twin Otter aircraft. PALS was mounted at the rear of the aircraft at a 40 degree incidence angle. A total of 17 days of SM retrievals were acquired using PALS with the flight commencing approximately at 6:30 AM local time each day. [70] provides a complete description of the study site and the SMAPVEX12 campaign. For the satellite scale, overlapping SMOS (RE04-MIR-CLF31A daily product) pixels ($\sim$ 18.5 km $\times$ 33 km) were used (Figure 3.1(b)). The RE04-MIR-CLF31A is a gridded L3 SMOS product which uses a multi-orbit (MO) retrieval approach for enhanced SM retrievals and is available at a temporal latency of 3.5-7 days. The MO retrieval is motivated by longer autocorrelation length of the vegetation optical depth compared to the corresponding SM autocorrelation enhancing the retrievals at the border of the satellite swath. [107] Here, we use the ascending overpass of SMOS occurring at approximately 6 AM local time as it is nearer to the *in situ* and airborne acquisition times.

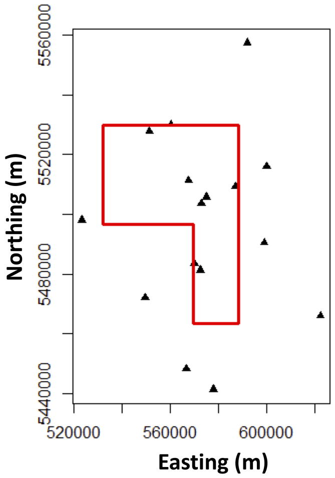

Figure 3.2: Location of sparse rain-gauge stations (denoted by triangles) for the study site. The red outline refers to the satellite extent of the study area. Only 8 of the 16 stations lie within the study region [2].

Leaf Area Index (LAI), used as a proxy for vegetation, was extracted from the four day composite MODIS product (MCD15A3H, version 6) at a 500 meters resolution available at (NASA) Land Processes Distributed Active Archive Center (LPDAAC). Since the LAI data are only available after a period of four days, the LAI values are linearly interpolated for the intervening days. For the airborne extent, the soil texture and elevation data are available at resolution of 1500 m on the campaign website whereas for the rest of area we extracted the data from the Canadian National Soils Database (NSDB). We spatially interpolate rainfall data using Inverse Distance Weighting (IDW) from 16 weather-stations available during the duration of the campaign (Figure 3.2).

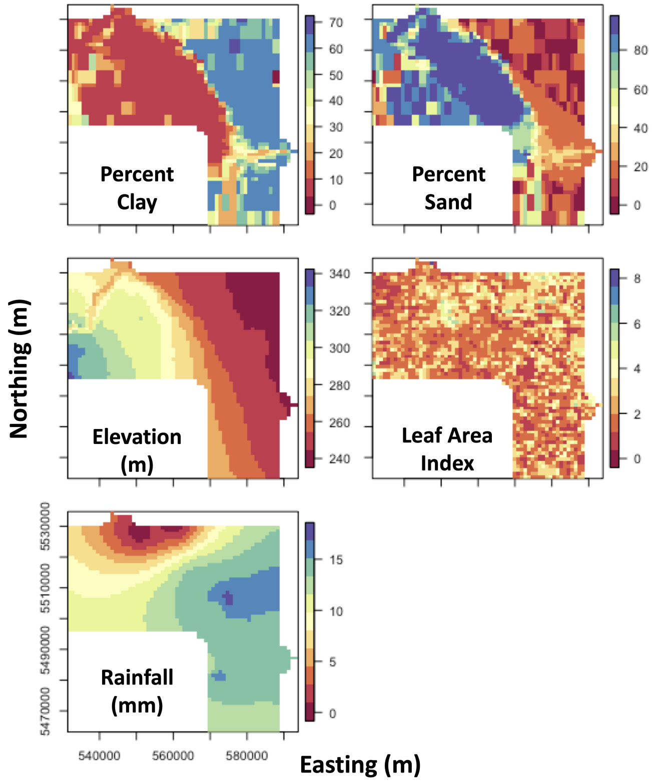

Figure 3.3: Spatial plots of percent clay, percent sand, elevation, leaf area index (DOY 164) and rainfall (DOY 164) [2].

## 3.4 Theory

A SHM finds its roots in one simple fact of probability theory [108]: The joint distribution of a collection of random variables can be defined as a product of a series of conditional distributions. Consider, for instance, the joint distribution of three random variables, $A$, $B$ and $C$ denoted by $[A, B, C]$, where [A] represents the probability distribution of the random variable $A$. We can decompose this joint distribution as $[A|B, C][B|C][C]$, where $[C|A, B]$ denotes the conditional probability distribution of $C$ given $A$ and $B$. Following [109], we specify a spatial hierarchical model (SHM) in three stages:

1. Data model $[\mathbf{z}|y(.), \mathbf{P}_z]$

2. Process model $[y(.)|\mathbf{P}_y]$

3. Parameter model $[\mathbf{P}_z, \mathbf{P}_y] = [\mathbf{P}]$

On top of the hierarchy lies the data model, where we model the observed SM (conditional on the process model, $y(.)$ and parameters in the data model, $\mathbf{P}_z$) from all the data instruments. The bias and errors caused due to imperfect retrievals are modeled here. Next is the process model, where we model the latent underlying SM (conditional on parameters in the process model, $\mathbf{P}_y$). By latent SM, we refer to the underlying (continuous) spatial distribution of SM which is devoid of measurement errors (random or systematic) induced by the sensors. In general, we do not know the exact form (and value) of the measurement errors and therefore make some assumptions regarding the form of the errors (and estimate them from the data). Our definition of the latent process (and subsequent inference and predictions) is subject to those assumptions. Lastly, we define the parameter model consisting of the joint probability distribution of the parameters used in the data and the process model. The parameter model characterizes the parametric uncertainty in the SHM.

### 3.4.1 Data model

A data model is where we model the distribution of the observations conditional on the latent SM process and the parameters. A data model typically (but not always) consists of the bias (systematic) and the error (random) structures of individual data instruments.

The spatial dependence induced in the observed SM at any given scale is due to the underlying latent SM process. Therefore, conditioned on the process model, the observed data can be considered independent of each other making data fusion of multiple instruments relatively straightforward. Note that, if we don't condition the observed data on the process, and instead look at their marginal (unconditional) distributions, then the assumption of independence becomes untenable. This makes a hierarchical approach powerful in data fusion.

For all the observations from a $j^{th}$ instrument on support scale A, we define the data as:

$$\mathbf{z}_j(A) = \boldsymbol{\Delta}_j + \mathbf{y}(A) + \boldsymbol{\epsilon}_j \tag{3.1}$$

where $\mathbf{z}_j(A)$ is the vector of observations from the $j^{th}$ data instrument, $\mathbf{y}(A)$ is the corresponding latent SM at support $A$, $\boldsymbol{\epsilon}_j \sim \mathcal{N}(0, \tau_j^2)$ is the normal measurement error with variance $\tau_j^2$ and $\boldsymbol{\Delta}_j$ is the bias associated with the $j^{th}$ data instrument. $\mathbf{P}_z$ comprises the parameters used to define $\tau_j^2$ and $\boldsymbol{\Delta}_j$. The exact form of error and bias will depend crucially on the data instrument under consideration.

### 3.4.2 Process model

Since we are dealing with SM processes at multiple spatial resolutions, it is beneficial to explicitly define the latent SM process at point scale called $y(.)$ and describe the SM process at other support scales in terms of $y(.)$. Modeling y(.) is the most crucial step in a hierarchical approach because 1) the observed spatial dependence is considered entirely due to y(.) [110] and 2) parameters with physical interpretation can be easily defined at the point scale [111].

*3.4.2.1 Process at point scale*

Since we aim to account for spatial dependence in our model, we define the latent SM as a geostatistical process at the point scale such that $y(.)$ is a Gaussian process with mean function $\mu$ and covariance function $C$ written as:

$$y(.) \sim GP(\mu, C) \tag{3.2}$$

For any given set of $n$ points $\{s_1, ..., s_n\}$ in space, $\mathbf{y} = (y(s_1), ..., y(s_n))' \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{C})$ is a $n$-dimensional normal distribution, where $\boldsymbol{\mu}$ is the $n$-dimensional mean vector and $\mathbf{C}$ is the $n \times n$ covariance matrix. The principal advantage of assuming that y(.) follows a Gaussian process is that the joint distribution for a given set of points can be completely defined by $\boldsymbol{\mu}$ and $\mathbf{C}$ i.e., we can define the spatial distribution of SM in terms of its mean, variance and correlation. The mean structure takes care of the deterministic variability while the covariance function defines the stochastic variations. Using a Gaussian process at the latent scale also leads to computational benefits and is a standard assumption of geostatistical models.

#### *3.4.2.2 Process at aggregate scale*

So far, we have defined the the process y(.) at point support. Since our aim is to combine multiple SM data sources at varying resolutions, we now define the average of $y(.)$ for an areal pixel A at support area $|A|$ as $y(A) = \frac{1}{|A|}{}_A y(s)ds$. *In situ* data can be considered a special case of the above with $|A| = 1$. Since, y(.) is a random variable, the spatial integral is a stochastic integral over all realizations of y(.) and will be difficult to solve for analytically. Under mild regularity conditions, which are satisfied for our model, the mean function $\mu(A)$ and the covariance between two pixels $A_1$ and $A_2$ can be written as:

$$\begin{aligned} \mu(A) &= \frac{1}{|A|}{}_A\mu(s)ds \\ C(A_1, A_2) &= \frac{1}{|A_1||A_2|}{}_{A_1 A_2}C(s_1, s_2)ds_2 ds_1 \end{aligned} \tag{3.3}$$

Note that the pixels $A_1$ and $A_2$ can be at same or different supports and can also consist of point data. Thus, in addition to modeling spatial dependence on a particular support scale, we also take into account dependence between observations across different scales. For the Gaussian process model for point scale, as defined in Section (3.4.2.1), it can be shown that for any given set $\mathcal{S} = \{A_1, ...., A_n\}$, the joint distribution of $\mathbf{y}(\mathcal{S}) = (y(A_1), ..., y(A_n))$ is given as $\mathcal{N}_n(\boldsymbol{\mu}(\mathcal{S}), \mathbf{C}(\mathcal{S}, \mathcal{S}))$, where the elements of $\boldsymbol{\mu}(\mathcal{S})$ and $\mathbf{C}(\mathcal{S}, \mathcal{S})$ are given by equation (3.3).

### 3.4.3 Data fusion

Data fusion of all the observations using the data model and process model in preceding sections would have been straightforward if we could evaluate the integrals in equation (3.3). Since this is not always possible, we proceed using a numerical approximation for the spatial integrals [65, 28, 112]. We assume a fine equidistant grid $\mathcal{G} = \{g_j : j = 1, ..., n_{grid}\}$ of $n_{grid}$ points over the entire study domain $\mathcal{D}$. For a (aggregate) data pixel $A_i$:

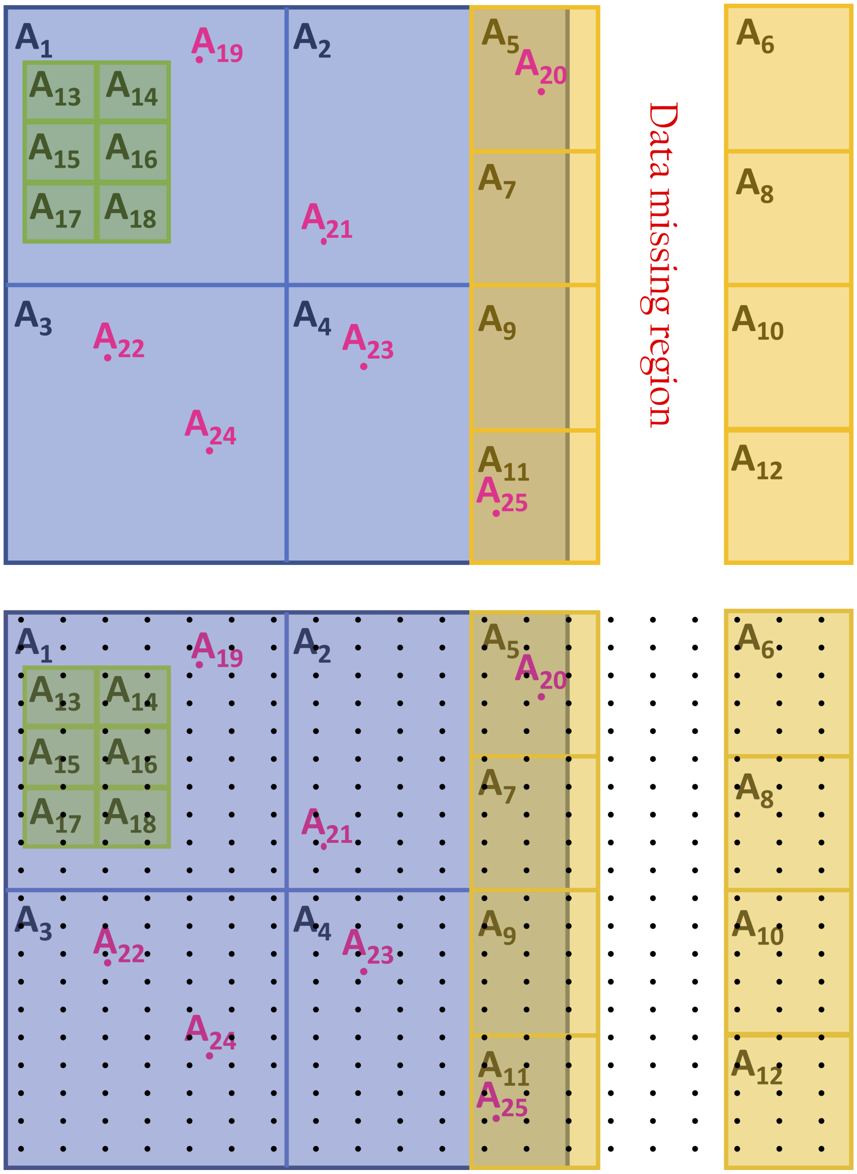

Figure 3.4: (Top) Illustration of a hypothetical study area with data from four instruments (color-coded) at different resolutions and extents. As depicted here, the fusion scheme can be applied even for overlapping and missing data. The pink color refers to point scale data with the total number of point data, $n_{point} = 7$ and areal data, $n_{pixels} = 18$ making the total number of data, $n = 25$. (Bottom) The equidistant point scale grid is denoted by black circles and the total number of black circles equals $n_{grid}$. Since the data fusion considers spatial correlation, we can do predictions in the missing region as well [2].

$$y(A_i) \approx \frac{1}{n_{A_i}} \sum_{g_j \in \mathcal{G} \cap A_i} y(g_j) \tag{3.4}$$

where $n_{A_i} = |\mathcal{G} \cap A_i|$ is the number of grid points inside the pixel $A_i$. We can equivalently write equation (3.4) as $\mathbf{h'_{A_i}} \mathbf{y_{A_i}}$ where $\mathbf{h_{A_i}} = (1/n_{A_i}, ..., 1/n_{A_i})$ and $\mathbf{y_{A_i}}$ is a vector with elements $\{y(g_j) : g_j \in \mathcal{G} \cap A_i\}$.

Following equation (3.4), if we observe $n_{pixels}$ data pixels denoted by the vector $\mathcal{S}_{pixels}$, we can write $\mathbf{y}(\mathcal{S}_{pixels}) \approx \mathbf{H_1} \mathbf{y}(\mathcal{G})$ where $\mathbf{H_1}$ is a $n_{pixels} \times n_{grid}$ matrix such that:

$$(\mathbf{H_1})_{i,j} = \begin{cases} 1/n_{A_i} & \text{if } g_j \in \mathcal{G} \cap A_i \\ 0 & \text{o}therwise. \end{cases} \tag{3.5}$$

In addition to pixel data, we might have point *in situ* data as well, whose locations might not necessarily overlap with the equidistant grid. Therefore, for $n_{point}$ *in situ* data points out of a total $n$ observations $\{A_i : i = 1, ..., n\}$, we define a $n \times (n_{grid} + n_{point})$ matrix $\mathbf{H}$ such that

$$(\mathbf{H})_{i,j} = \begin{cases} 1_{i=j} & \text{if } A_i \text{ is an \textit{in situ} data point} \\ \frac{1}{n_{A_i}} \times 1_{g_j \in \mathcal{G} \cap A_i} & \text{if } A_i \text{ is a data pixel.} \end{cases} \tag{3.6}$$

where $1_R$ is the indicator function defined as $1_R(x) := \begin{cases} 1 & x \in R \\ 0 & x \notin R \end{cases}$. The matrix $\mathbf{H}$ can therefore be considered as a transformation matrix to fuse data at different spatial resolutions. Fusing all the observations, the data model in equation (3.1), becomes:

$$[\mathbf{z}|\mathbf{y}, \mathbf{P}_z] \sim \mathcal{N}_n(\mathbf{\Delta} + \mathbf{H}\mathbf{y}(\mathcal{G}), \mathbf{\Sigma}) \tag{3.7}$$

where $\mathbf{z} = (z_{A_1}, ..., z_{A_n})$ represents the observed SM data from all the instruments, $\mathbf{\Delta} = (\Delta_{A_1}, ..., \Delta_{A_n})$ are the biases associated with each observation and $\mathbf{\Sigma}$ is the $n \times n$ error matrix. The numerical approximation is illustrated using a hypothetical example in Figure 3.4.

### 3.4.4 Parameter model

In the parameter model, we specify the joint distribution of the parameters in the data and process models. In a Bayesian framework, this consists of putting prior probability distribution on the parameters $\mathbf{P} = [\mathbf{P}_z, \mathbf{P}_y]$ characterizing our prior beliefs of the parameters in the data and process models.

## 3.5 Methodology

The exact form for the mean, covariance, bias and error functions for the SHM depends on the sensors, the available ancilliary data and surface heterogeneity of the study region. In the present section, we specify the exact SHM to fuse SM data (*in situ*, airborne and satellite) during SMAPVEX12 campaign. The flowchart of the data fusion scheme is given in Figure 3.5.

### 3.5.1 SHM specification

#### *3.5.1.1 Data model*

For SMAPVEX12, three replicate readings of *in situ* data are available. We average the three readings and regard the resulting value as the "ground truth", i.e. we assume that it has no error/bias. For the active-passive airborne sensors, it has been observed that the backscatter data is affected by second-order effects of clay content and first order effects of vegetation [113]. We therefore model bias as a second order polynomial function of clay content and a linear function of LAI, i.e, $\boldsymbol{\Delta} = \mathbf{X}_{\eta}\boldsymbol{\eta}$ where $\mathbf{X}_{\eta}$ is the vector of (Intercept, LAI, $\%$Clay, $\%$ Clay$^2$) and $\boldsymbol{\eta}$ is the vector of regression coefficients. For the error structure, we assume a normal distribution with zero mean and constant variance $\tau^2$. We assume zero bias and error for the satellite data due to fewer satellite pixels (four) in the study region. This will not induce much error since SM retrievals from passive sensors (used by SMOS) are more robust than active sensors [113]. For a study area with sufficient satellite pixels, an appropriate bias/error structure can be parameterized for the satellite data as well. It should be noted that even though the number of satellite pixels are few in our study region, they are critical to constrain the SM predictions in regions where no other observations (*in*

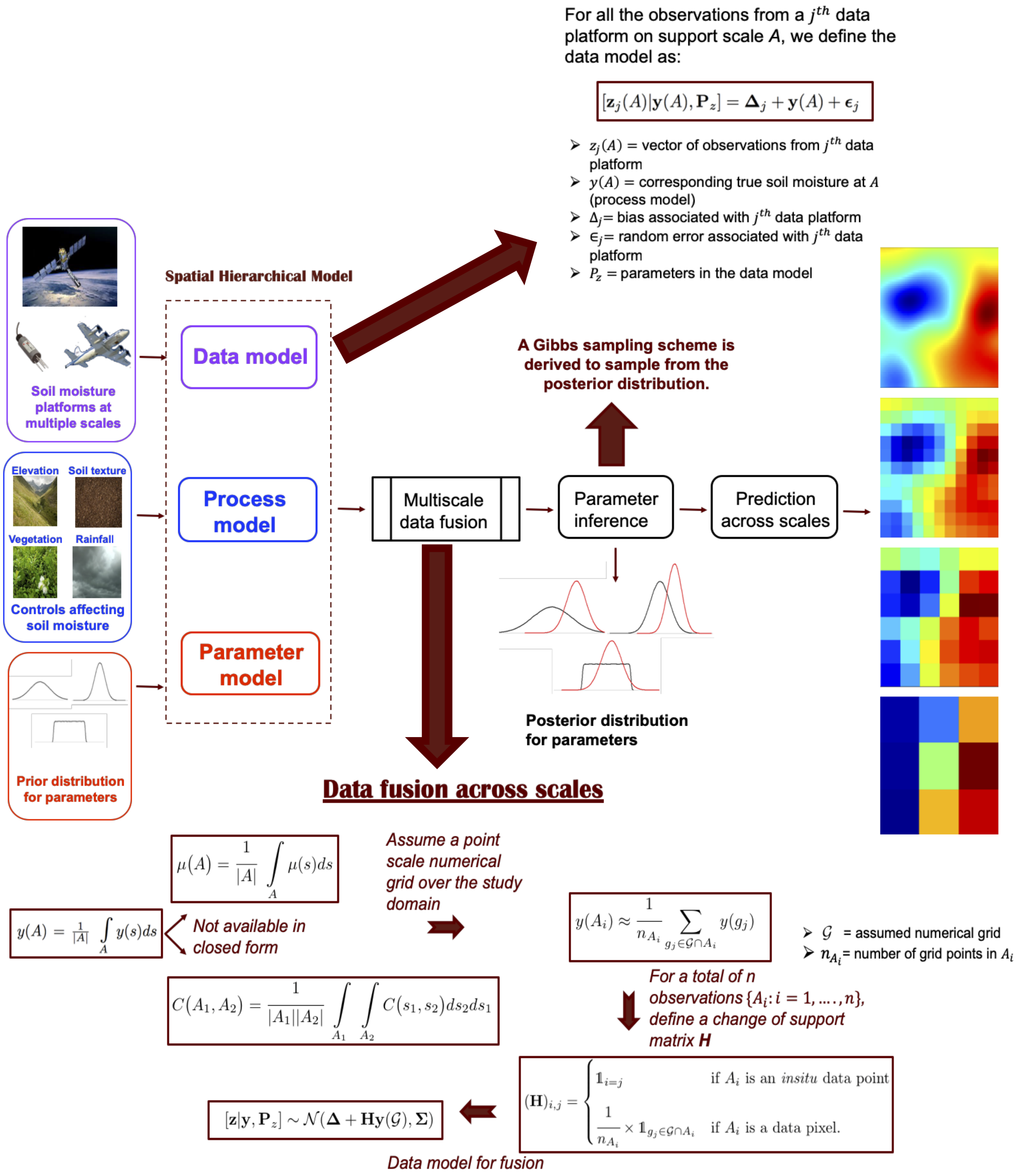
For all the observations from a $j^{th}$ data platform on support scale $A$, we define the data model as:
$[\mathbf{z}_j(A)|\mathbf{y}(A),\mathbf{P}_z] = \boldsymbol{\Delta}_j + \mathbf{y}(A) + \boldsymbol{\epsilon}_j$
➢ $z_j(A)$ = vector of observations from $j^{th}$ data platform
➢ $y(A)$ = corresponding true soil moisture at $A$ (process model)
➢ $\Delta_j$= bias associated with $j^{th}$ data platform
➢ $\epsilon_j$= random error associated with $j^{th}$ data platform
➢ $P_z$ = parameters in the data model
Spatial Hierarchical Model
Data model
Process model
Parameter model
Soil moisture platforms at multiple scales
Elevation
Soil texture
Vegetation
Rainfall
Controls affecting soil moisture
Prior distribution for parameters
A Gibbs sampling scheme is derived to sample from the posterior distribution.
Multiscale data fusion
Parameter inference
Prediction across scales
Posterior distribution for parameters
Data fusion across scales
$\mu(A) = \frac{1}{|A|}\int_A \mu(s)ds$
$y(A) = \frac{1}{|A|}\int_A y(s)ds$
Not available in closed form
$C(A_1,A_2) = \frac{1}{|A_1||A_2|}\int_{A_1}\int_{A_2} C(s_1,s_2)ds_2ds_1$
Assume a point scale numerical grid over the study domain
$y(A_i) \approx \frac{1}{n_{A_i}}\sum_{g_j\in\mathcal{G}\cap A_i} y(g_j)$
➢ $\mathcal{G}$ = assumed numerical grid
➢ $n_{A_i}$= number of grid points in $A_i$
For a total of n observations $\{A_i : i = 1, \dots, n\}$, define a change of support matrix $\mathbf{H}$
$(\mathbf{H})_{i,j} = \begin{cases} \mathbb{1}_{i=j} & \text{if } A_i \text{ is an } insitu \text{ data point} \\ \frac{1}{n_{A_i}} \times \mathbb{1}_{g_j\in\mathcal{G}\cap A_i} & \text{if } A_i \text{ is a data pixel.} \end{cases}$
$[\mathbf{z}|\mathbf{y},\mathbf{P}_z] \sim \mathcal{N}(\boldsymbol{\Delta} + \mathbf{H}\mathbf{y}(\mathcal{G}),\boldsymbol{\Sigma})$
Data model for fusion

Figure 3.5: Flowchart of the data fusion framework using a Bayesian spatial hierarchical model. Soil moisture data along with their errors are modeled in the data model. The spatial dynamics of soil moisture governed by atmospheric and surface controls are parameterized by a non-stationary geostatistical model while the parameter model accounts for the uncertainty in parameters [2].

*situ* or airborne) are available.

### *3.5.1.2 Process model*

The spatial distribution of the latent SM is typically affected by land surface controls such as topography, vegetation and soil texture, as well as atmospheric variables such as rainfall inducing spatial non-stationarity in SM distribution. Naturally, the mean and covariance functions of the geostatistical process model should essentially be functions of these controls.

In absence of strong prior information (or theoretical justification) about the mean structure, it is preferable to assume the mean as a linear function of the covariates, $\mathbf{X}_{\boldsymbol{\beta}}\boldsymbol{\beta}$, where $\mathbf{X}_{\boldsymbol{\beta}}$ is the vector of spatial covariates affecting the mean and $\boldsymbol{\beta}$ are the regression coefficients, and model the residual spatial structure in the covariance function [114]. Defining a covariate-dependent covariance function, is not trivial. Traditionally, geostatistical studies have often relied on isotropic covariance functions whose spatial structure is considered constant for the entire study domain. In areas with high geophysical heterogeneity, the stationarity assumption for SM covariance tends to be violated [90].

To attend to this shortcoming, we adapt the spatial model of [63] and the spatio-temporal covariance function of [64] for our study following [90] such that the covariance function can take into account the non-stationarity arising due to heterogeneity in controls. The covariance function is then a function of the controls (in addition to the location) i.e. the covariance between any two locations $s_1$ and $s_2$ is given by $C(s_1, s_2) = C(s_1, s_2, \mathbf{X}_{\boldsymbol{\alpha}}(s_1), \mathbf{X}_{\boldsymbol{\alpha}}(s_2))$ where $\mathbf{X}_{\boldsymbol{\alpha}}(s)$ is the vector of spatial covariates affecting the covariance at point $s$. We define the covariance function as a weighted sum of isotropic parametric covariance functions where the weights explicitly depends on the controls $\mathbf{X}_{\boldsymbol{\alpha}}$. Therefore,

$$C(s_1, s_2, \mathbf{X}_{\boldsymbol{\alpha}}(s_1), \mathbf{X}_{\boldsymbol{\alpha}}(s_2)) = \sum_{j=1}^{M} w_j(\mathbf{X}_{\boldsymbol{\alpha}}(\mathbf{s_1}))w_j(\mathbf{X}_{\boldsymbol{\alpha}}(\mathbf{s_2}))C_j(|\mathbf{s_1} - \mathbf{s_2}|) \tag{3.8}$$

where $w_j$ is the weight for the $j^{th}$ covariance function $C_j$ and M is the total number of isotropic covariance functions. Though the weights $w_j$'s can be parameterized in any way, we assume them to be multinomial logistic functions of the controls (ensuring that the sum of the weighting functions equals one) i.e.

$$w_j(\mathbf{X}_{\boldsymbol{\alpha}}(\boldsymbol{s})) = \frac{exp(\mathbf{X}_{\boldsymbol{\alpha}}(\boldsymbol{s})^T \boldsymbol{\alpha}_j)}{\sum_{l=1}^{M} exp(\mathbf{X}_{\boldsymbol{\alpha}}(\boldsymbol{s})^T \boldsymbol{\alpha}_l)} \tag{3.9}$$

where $\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_M$ are vectors of regression coefficients that describe the effect of geophysical controls on the covariance function. If the number of controls affecting the covariance equals $q$, then $\boldsymbol{\alpha}_j = [\alpha_{j1}, ..., \alpha_{j(q+1)}]$, where $\alpha_{j1}$ is the coefficient for the intercept term. Equation (3.9) can also be used to model isotropic covariance function (in such a case, one weighting function out of all $w_j$s equal to one, rest equal to zero) which might be present in relatively homogeneous areas. Further details regarding the non-stationary covariance function can be found in [90].

*3.5.1.3 Parameter model*

After specifying the SHM, the main interest is in parameter inference (posterior parameter distribution $[\mathbf{P}|\mathbf{z}]$) from the observed data $\mathbf{z}$ and then subsequently doing predictions of the latent process (posterior predictive distribution $[\mathbf{y}^{Pr}|\mathbf{z}]$) at unobserved locations. Let $\boldsymbol{\gamma}$ be the vector denoting the parameters in the covariance function and $\tau^2$. Then, the posterior parameter distribution $[\mathbf{P}|\mathbf{z}] = [\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}|\mathbf{z}]$. Since it is difficult to sample from the joint distribution $[\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}|\mathbf{z}]$, we instead, carry out a Gibbs sampler to sample from the full-conditional distributions (FCDs) of $\boldsymbol{\beta}$, $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ respectively. Specifically, given an initial vector of parameters $\mathbf{P}^{[0]} = \{\boldsymbol{\beta}^{[0]}, \boldsymbol{\eta}^{[0]}, \boldsymbol{\gamma}^{[0]}\}$, we generate $\mathbf{P}^{[l]}$ from $\mathbf{P}^{[l-1]}$ as follows:

$$\begin{aligned} &\text{sample } \boldsymbol{\beta}^{[l]} \sim [\boldsymbol{\beta}|\boldsymbol{\eta}^{[l-1]}, \boldsymbol{\gamma}^{[l-1]}, \mathbf{z}] \\ &\text{sample } \boldsymbol{\eta}^{[l]} \sim [\boldsymbol{\eta}|\boldsymbol{\beta}^{[l]}, \boldsymbol{\gamma}^{[l-1]}, \mathbf{z}] \\ &\text{sample } \boldsymbol{\gamma}^{[l]} \sim [\boldsymbol{\gamma}|\boldsymbol{\beta}^{[l]}, \boldsymbol{\eta}^{[l]}, \mathbf{z}] \end{aligned} \tag{3.10}$$

After a burn-in period, for large $L$, the sampling distribution of $\mathbf{P}^{[1]}, .., \mathbf{P}^{[L]}$ approaches the posterior parameter distribution $[\mathbf{P}|\mathbf{z}]$ [115], where $\mathbf{P}^{[l]} = \{\boldsymbol{\beta}^{[l]}, \boldsymbol{\eta}^{[l]}, \boldsymbol{\gamma}^{[l]}\}$.

The FCDs on the right hand side in equation 3.10 are derived in Appendix A. To get good starting values for our parameters, we use a global optimization algorithm, Generalized Simulated Annealing (GenSA) [80], to find the point parameter estimates that maximize the likelihood of the SHM. After a burn-in period, parameter convergence is monitored using trace-plots, and samples (from equation 3.10) are used to approximate the posterior parameter distribution. Note that, in A, the FCDS for $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are available in closed form only for normal prior distributions. If a normal prior distribution is untenable, sampling can still be carried out using Metropolis-Hastings [115] though the convergence will be slower.

The effects of the controls on the mean, covariance and bias are analyzed by examining the 95$\%$ highest posterior density (HPD) interval for the parameters. The 95$\%$ highest posterior density (HPD) interval denotes the subset of the parameter space such that the posterior probability of the parameter lying in that subset equals 95$\%$. For the mean (bias), if the 95$\%$ HPD interval includes zero or is very close to zero, we assume that there is little evidence in the data that the control affects the mean (bias) and if the 95$\%$ HPD interval excludes zero (or is not close to zero), we assume that there is evidence in the data that the control affects the mean (bias). Note that, we cannot say with certainty that the effect is present because there might be a small probability that the parameter equals zero. For the covariance, the HPD intervals are calculated for different lags. Since the covariance between any two points (governed by equations (3.8) and (3.9)) is a function of their lag-distance in space, as well as the vector of controls affecting the covariance ($\boldsymbol{X_\alpha}$), the effect of an individual control on the covariance is not trivial. The effect of an individual control on the covariance is quantified by comparing the covariance when the control is at the mean value (of the study region) compared to when the control is at extreme value ($10^{th}$ and $90^{th}$ percentile) while keeping the other controls at their mean values [64]. Since all the controls are standardized (mean = 0, variance = 1), for a $k^{th}$ control, a simple way to do this is to compute the 95$\%$ HPD region for $\delta_{x_k}^{low}|s|$ and $\delta_{x_k}^{high}|s|$ for any lag $|s|$ given as:

$$\begin{aligned} \delta_{x_k}^{low}|s| &= \frac{C_{x_k}^{low}(|s|)}{C_{x_k}^{mean}(|s|)} = \frac{\sum_{j=1}^{M} \frac{exp(\alpha_{j1}+(x_k^{0.1})\alpha_{jk})}{\sum_{l=1}^{M} exp(\alpha_{l1}+(x_k^{0.1})\alpha_{lk})} C_j(|s|)}{\sum_{j=1}^{M} \frac{exp(\alpha_{j1})}{\sum_{l=1}^{M} exp(\alpha_{l1})} C_j(|s|)} \\ \delta_{x_k}^{high}|s| &= \frac{C_{x_k}^{high}(|s|)}{C_{x_k}^{mean}(|s|)} = \frac{\sum_{j=1}^{M} \frac{exp(\alpha_{j1}+(x_k^{0.9})\alpha_{jk})}{\sum_{l=1}^{M} exp(\alpha_{l1}+(x_k^{0.9})\alpha_{lk})} C_j(|s|)}{\sum_{j=1}^{M} \frac{exp(\alpha_{j1})}{\sum_{l=1}^{M} exp(\alpha_{l1})} C_j(|s|)} \end{aligned} \tag{3.11}$$

where for the $k^{th}$ control $x_k$, $x_k^{0.1}$ and $x_k^{0.9}$ are the $10^{th}$ and $90^{th}$ percentiles. Since $\alpha_{j1}$ is the intercept term, it is always included. Again, if the HPD region includes one (or is very close to one), then we assume that there is little evidence in the data that the control affects the covariance and if the 95% HPD interval excludes one (or is not close to one), we assume that there is evidence in the data that the control affects the covariance [64]. A similar argument can be made for the variance by analyzing the HPD region of $\delta_{x_k}^{low}|0|$ ($\delta_{x_k}^{high}|0|$) and for correlation at any lag $|s|$ by analyzing the HPD region of $\delta_{x_k}^{low}|s|/\delta_{x_k}^{low}|0|$ ($\delta_{x_k}^{high}|s|/\delta_{x_k}^{high}|0|$).

### 3.5.2 Spatial Prediction

Once we have the samples from the posterior parameter distribution, spatial prediction is straightforward. Given observations $\mathbf{z} = \mathbf{z}(\mathcal{S})$ at a set of locations $\mathcal{S}$ and the posterior samples $\mathbf{P}^{[1]}, .., \mathbf{P}^{[L]}$ from $[\mathbf{P}|\mathbf{z}]$, the posterior predictive distribution $[\mathbf{y}^{Pr}|\mathbf{z}] = [\mathbf{y}(\mathcal{S}^{Pr})|\mathbf{z}]$ at the unobserved locations $\mathcal{S}^{Pr} = \{\boldsymbol{s}_1^{Pr}, ..., \boldsymbol{s}_{n_{Pr}}^{Pr}\}$ can be approximated by $[\mathbf{y}^{Pr}|\mathbf{z}] = (1/L)\sum_{l=1}^{L}[\mathbf{y}^{Pr}|\mathbf{z}, \mathbf{P}^{[l]}]$. For the SHM proposed in the study, it can be derived that for a posterior sample $\mathbf{P}^{[l]}$, $[\mathbf{y}^{Pr}|\mathbf{z}, \mathbf{P}^{[l]}] \sim \mathcal{N}_{n_{Pr}}(\mu_{\mathbf{y}^{Pr}|\mathbf{z}}^{[l]}, \Sigma_{\mathbf{y}^{Pr}|\mathbf{z}}^{[l]})$ where:

$$\mu_{\mathbf{y}^{Pr}|\mathbf{z}}^{[l]} = \mu^{[l]}(\mathcal{S}^{Pr}) + C^{[l]}(\mathcal{S}^{Pr}, \mathcal{G})\mathbf{H}'(\Sigma_{\mathbf{H}}^{-1})^{[l]}(\mathbf{z} - \mu^{[l]}(\mathcal{S}) - \Delta^{[l]}(\mathcal{S})) \tag{3.12}$$

$$\Sigma_{\mathbf{y}^{Pr}|\mathbf{z}}^{[l]} = C^{[l]}(\mathcal{S}^{Pr}, \mathcal{S}^{Pr}) - C^{[l]}(\mathcal{S}^{Pr}, \mathcal{G})\mathbf{H}'(\Sigma_{\mathbf{H}}^{-1})^{[l]}\mathbf{H}C^{[l]}(\mathcal{G}, \mathcal{S}^{Pr}) \tag{3.13}$$

where $\Sigma_{\mathbf{H}} = \mathbf{H}C(\mathcal{G}, \mathcal{G})\mathbf{H}' + \mathbf{\Sigma}$, and $\mu_{\mathbf{y}^{Pr}|\mathbf{z}}^{[l]}$ and $\Sigma_{\mathbf{y}^{Pr}|\mathbf{z}}^{[l]}$ are the predicted mean and covariance respectively for the posterior sample $\boldsymbol{P}^{[l]}$. To obtain predictions at any spatial support $A$, we predict at the point level following equations (3.12) and (3.13) and then upscale to any spatial support. The predicted mean and the covariance at support scale $A$ would then be given as:

$$\begin{aligned} \mu^{[l]}_{\mathbf{y}_A^{Pr}|\mathbf{z}} &= \mathbf{H}_A \mu^{[l]}_{\mathbf{y}^{Pr}|\mathbf{z}} \\ \Sigma^{[l]}_{\mathbf{y}_A^{Pr}|\mathbf{z}} &= \mathbf{H}_A \Sigma^{[l]}_{\mathbf{y}^{Pr}|\mathbf{z}} \mathbf{H}'_A \end{aligned} \tag{3.14}$$

where, $\mathbf{H}_A$, for a support $A$ is given by equation (3.6). Posterior summaries of predicted mean and covariance (such as average and 95% HPD intervals) can then be calculated using the sampling distribution resulting from applying equation 3.14 to $\{\mathbf{P}^{[1]}, .., \mathbf{P}^{[L]}\}$.

Table 3.1: Summary statistics for soil moisture (v/v) for point, airborne and satellite scales [2].

| | Point | | | | Airborne | | | | Satellite | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DOY | Min | Mean | Max | SD | Min | Mean | Max | SD | Min | Mean | Max | SD |
| 164 | 0.075 | 0.334 | 0.668 | 0.114 | 0.077 | 0.296 | 0.551 | 0.133 | 0.188 | 0.286 | 0.371 | 0.079 |
| 167 | 0.084 | 0.280 | 0.629 | 0.113 | 0.070 | 0.221 | 0.545 | 0.089 | 0.193 | 0.198 | 0.205 | 0.005 |
| 169 | 0.118 | 0.344 | 0.665 | 0.125 | 0.062 | 0.280 | 0.551 | 0.135 | 0.171 | 0.253 | 0.336 | 0.067 |
| 174 | 0.063 | 0.307 | 0.640 | 0.130 | 0.084 | 0.270 | 0.551 | 0.111 | 0.167 | 0.189 | 0.220 | 0.025 |
| 177 | 0.019 | 0.245 | 0.642 | 0.116 | 0.030 | 0.162 | 0.480 | 0.071 | 0.096 | 0.100 | 0.105 | 0.004 |
| 179 | 0.025 | 0.224 | 0.688 | 0.100 | 0.030 | 0.126 | 0.480 | 0.063 | 0.077 | 0.093 | 0.101 | 0.011 |
| 192 | 0.019 | 0.170 | 0.517 | 0.079 | 0.030 | 0.143 | 0.418 | 0.072 | 0.072 | 0.082 | 0.096 | 0.011 |
| 196 | 0.025 | 0.174 | 0.545 | 0.087 | 0.030 | 0.144 | 0.443 | 0.071 | 0.056 | 0.115 | 0.147 | 0.041 |
| 199 | 0.054 | 0.252 | 0.547 | 0.095 | 0.057 | 0.210 | 0.507 | 0.092 | 0.168 | 0.179 | 0.210 | 0.020 |
| 201 | 0.056 | 0.189 | 0.420 | 0.084 | 0.036 | 0.162 | 0.421 | 0.071 | 0.103 | 0.131 | 0.153 | 0.022 |

### 3.6 Results and Discussion

We apply the proposed fusion scheme separately for each day for which *in situ*, airborne and satellite data are available resulting in a total of 10 days, i.e. the data fusion scheme is spatial-only. Due to the less number of days of the dataset, it will be difficult to specify and estimate a valid controls-driven spatio-temporal mean/covariance function using the data though analyzing the temporal dynamics of SM is equally important. The summary statistics for SM at each scale are given in Table 3.1. For *in situ* SM, we average the three replicate measurements at each location. We take soil texture (percent sand and percent clay), elevation and vegetation (LAI) to

Table 3.2: Summary statistics for land surface controls and antecedent rain [2]. LAI = Leaf Area Index, ant. = antecedent

| | LAI | | | | 1-day ant. rain (mm) | | | | 2-day ant. rain (mm) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DOY | Min | Mean | Max | SD | Min | Mean | Max | SD | Min | Mean | Max | SD |
| 164 | 0.00 | 2.39 | 6.00 | 1.09 | 0.01 | 6.71 | 11.42 | 2.17 | 0.00 | 4.67 | 9.11 | 1.84 |
| 167 | 0.00 | 1.80 | 7.00 | 1.03 | 0.00 | 0.26 | 1.65 | 0.16 | 0.00 | 1.20 | 3.03 | 0.47 |
| 169 | 0.00 | 1.05 | 7.00 | 1.41 | 0.03 | 15.88 | 29.91 | 5.21 | 0.00 | 0.00 | 0.00 | 0.00 |
| 174 | 1.00 | 3.11 | 6.00 | 1.28 | 0.00 | 2.50 | 12.66 | 2.53 | 0.00 | 0.40 | 2.27 | 0.21 |
| 177 | 1.00 | 3.35 | 7.00 | 1.35 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 179 | 1.00 | 3.32 | 7.00 | 1.27 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 192 | 1.00 | 3.37 | 7.00 | 1.21 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 | 0.25 | 0.04 |
| 196 | 1.00 | 3.03 | 7.00 | 1.19 | 0.00 | 0.04 | 0.25 | 0.04 | 2.44 | 17.94 | 57.53 | 8.57 |
| 199 | 1.00 | 3.09 | 6.00 | 1.20 | 0.00 | 0.16 | 1.01 | 0.12 | 5.79 | 8.37 | 12.26 | 0.87 |
| 201 | 1.00 | 3.28 | 7.00 | 1.33 | 0.00 | 0.01 | 0.08 | 0.01 | 0.00 | 0.21 | 1.69 | 0.23 |
| | % Clay | | | | %Sand | | | | Elevation (m) | | | |
| | 0.00 | 31.36 | 69.00 | 24.57 | 0.00 | 47.85 | 93.00 | 35.29 | 240.00 | 269.26 | 335.37 | 23.96 |

be the dominant land surface controls and the two-day cumulative antecedent precipitation to be the dominant atmospheric-control affecting the SM distribution [16, 46, 45, 14, 103] for both the watersheds. The summary statistics for the controls are given in Table 3.2. Prior to our analysis, we standardize all the controls by their respective means and standard deviations. This is done to ensure that they are at the same scale and range, so that their effects on the mean, covariance and bias be compared. For the process model, we assume that the mean structure is affected by both precipitation and land surface controls while the covariance is affected only by the land surface controls. We do this because we spatially interpolate precipitation using IDW which might not represent point-scale rainfall dynamics accurately. For the study region, we find that percent sand, percent clay and elevation have a high correlation with each other ($\sim$ 0.7). Therefore, to avoid multicollinearity, we perform principal component analysis on all the three variables and take the first PC (explaining more than 85% of the variance) in our analysis. For each day, the first PC equals $0.60 \times (\%Clay) - 0.60 \times (\%Sand) - 0.54 \times (Elevation)$. Therefore for the mean function, we use antecedent 2-day cumulative rainfall, LAI and first PC of percent sand, percent clay and elevation as our controls, and for the covariance we exclude rainfall.

As the assumed grid $\mathcal{G}$ (Section 3.4.3) becomes finer, the numerical approximation approaches the spatial integrals in equation 3.3 but at the expense of increased computational cost. Therefore, we assume a grid spacing of 750 m (half of the airborne support) for the entire study domain to strike a balance between accuracy and computational expense. For the covariance given by equation 3.8, we use M=2 and use a mixture of exponential covariance functions for $C_js$. We restrict M=2, as increasing M did not improve predictions.

Assuming we have no prior knowledge of SM distribution, we use *apriori* independent diffuse/vague prior distributions for all the parameters. For the $\boldsymbol{\beta}$, $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$ regression coefficients, we assume normal distributions $N(0, 10^2)$. Since the parameters in the covariance functions $C_js$ in equation (3.8) and error variance $\tau^2$ are always non-negative, we assume a uniform distribution $U(0, 20)$. We divide the Northing-Easting coordinates (in meters) by 100000 so that abovementioned uniform prior is appropriate for the range parameters of the covariance functions.

### 3.6.1 Validation

To explore the validity of the fusion algorithm at multiple scales, we hold out observations at both point and airborne support. We don't hold out observations at the satellite support because of the limited number of pixels. As the difference in support between the three data instruments is large, if we are able to predict at all the three supports reasonably well, it will be a good validation for the algorithm. Since we are trying to combine observations across the three spatial scales, if we get satisfactory results at all the scales, we hypothesize that that the predictions for the intermediate scales will be satisfactorily interpolated as well. For validation, we cannot randomly hold-out observations (as is usually done for independent and identically distributed models) since the data fusion model accounts for spatial correlation. Therefore, we hold-out a section in the south-eastern region of the study area which accounts for roughly 20$\%$ of the airborne pixels and the *in situ* data lying inside the section. The area of the hold-out region varies from 142 to 193 km$^2$ (Table 3.3). The hold-out area for each day is chosen such that it has a high density of *in situ* measurements (37$\%$ to 56$\%$ of the total *in situ* measurements (Table 3.3)) and shows high variability in soil moisture values. The hold-out area for DOY 164 is given in Figure 3.6.

Table 3.3: Number of hold-out data at point and airborne scales for individual days [2].

| DOY | Hold-out area (km$^2$) | total point data | hold-out point data | total airborne pixels | hold-out airborne data |
|---|---|---|---|---|---|
| 164 | 171.00 | 879 | 347 (39.48%) | 373 | 76 (20.38%) |
| 167 | 186.75 | 735 | 288 (39.18%) | 412 | 83 (20.15%) |
| 169 | 173.25 | 828 | 294 (35.51%) | 380 | 77 (20.26%) |
| 174 | 180.00 | 827 | 318 (38.45%) | 393 | 80 (20.36%) |
| 177 | 189.00 | 864 | 349 (40.39%) | 413 | 84 (20.34%) |
| 179 | 141.75 | 806 | 322 (39.95%) | 312 | 63 (20.19%) |
| 192 | 191.25 | 805 | 341 (42.36%) | 422 | 85 (20.14%) |
| 196 | 191.25 | 807 | 299 (37.05%) | 418 | 85 (20.33%) |
| 199 | 193.50 | 789 | 292 (37.01%) | 427 | 86 (20.14%) |
| 201 | 193.50 | 313 | 176 (56.23%) | 425 | 86 (20.24%) |

We find the mean of the posterior predictive distribution at fine and airborne support for the hold-out region using equations 3.12 and 3.14 respectively. To validate against the hold-out observations we average the predicted mean resulting from all the posterior samples (Section 3.5.2). Since we assumed bias in PALS data, we add the bias back to the predictions at the airborne scale, so we can validate against the observed airborne data. The one-to-one plots between the observations and the average of the mean posterior predictive distribution at *in situ* and airborne scale along with their correlation (R) are given in Figure 3.7. The correlation values range from 0.40 to 0.88 for *in situ* data and from 0.70 to 0.93 for the airborne data. Though, the general trend of the observations is well predicted for all days at both the scales, the airborne predictions perform better than *in situ*. This can be attributed to the fact that the *in situ* data in each field is spaced 75 m away from each other [70], while the spatial support of the land-surface controls used to drive the data fusion model vary from 500 m (LAI) to 1500 m (soil and elevation). Also, our estimate for the spatial distribution of rainfall using IDW on sparse rainfall data (Figure 3.2) is approximate at best, and does not account for the spatial heterogeneity induced in SM by complex rainfall patterns. The correlation coefficient for both the supports is the lowest for DOY 196 which receives the heaviest antecendent rainfall as recorded by the raingauges (Table 3.2).

We also do predictions for the 4 SMOS pixels analyzed in the study. The RMSE values for the *in situ*, airborne and the 4 SMOS pixels are given in Table 3.4. For *in situ* data, the RMSE values

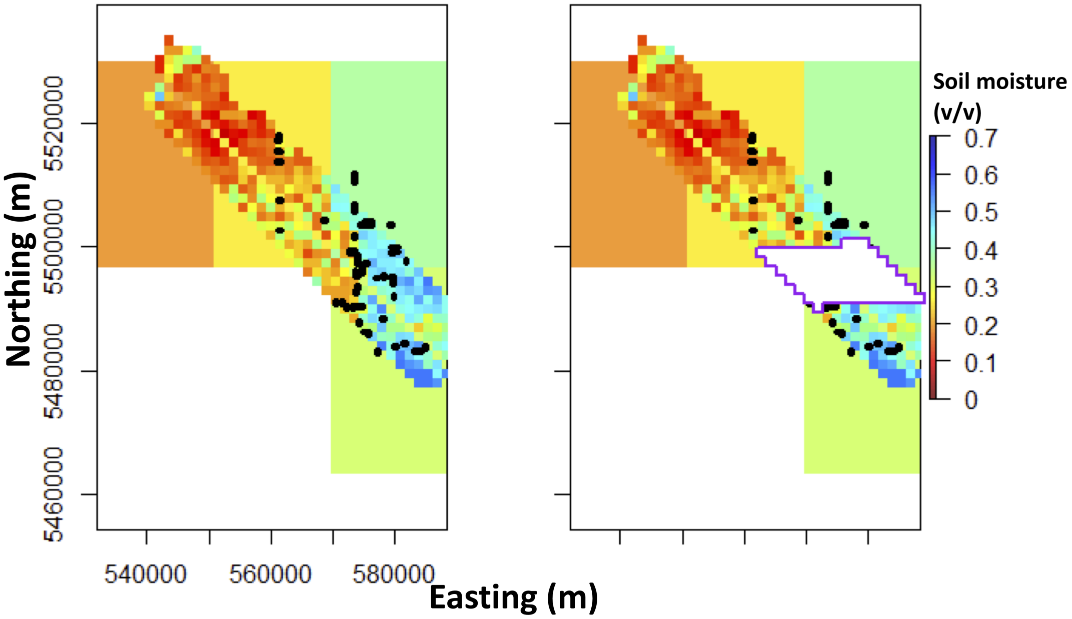

Figure 3.6: (Left) Soil moisture data at point (denoted with black), airborne (diagonal strip) and satellite (four pixels in the background) support for DOY 164. (Right) The sub-area, denoted by purple border, represents the hold-out area for validation of the algorithm at point and airborne scales. The area is chosen such that it has a high density of *in situ* data and also exhibits good soil moisture variability [2].

lie between 0.032 to 0.099 v/v while for the airborne scale the RMSE lies between 0.029 v/v to 0.074 v/v. Again, airborne predictions do better than *in situ* in terms of RMSE. Since the first four days have the highest mean wetness conditions (Table 3.1) and large SM variability (in terms of the standard deviation), they also register high RMSE values. For the SMOS pixels, the RMSE values lie below 0.04 v/v. In general, the SHM provides a good fit to the data across all the three scales and is therefore able to satisfactorily capture the multi-scale SM dynamics in the region.

### 3.6.2 Effect of controls on soil moisture

The 95% HPD intervals of the posterior distribution of the mean parameters ($\boldsymbol{\beta}$) are given in Figure 3.8. As mentioned in Section 3.5.1.3, if the 95% HPD interval for the mean parameters ($\boldsymbol{\beta}$) includes zero or is very close to zero, we assume that there is little evidence in the data that

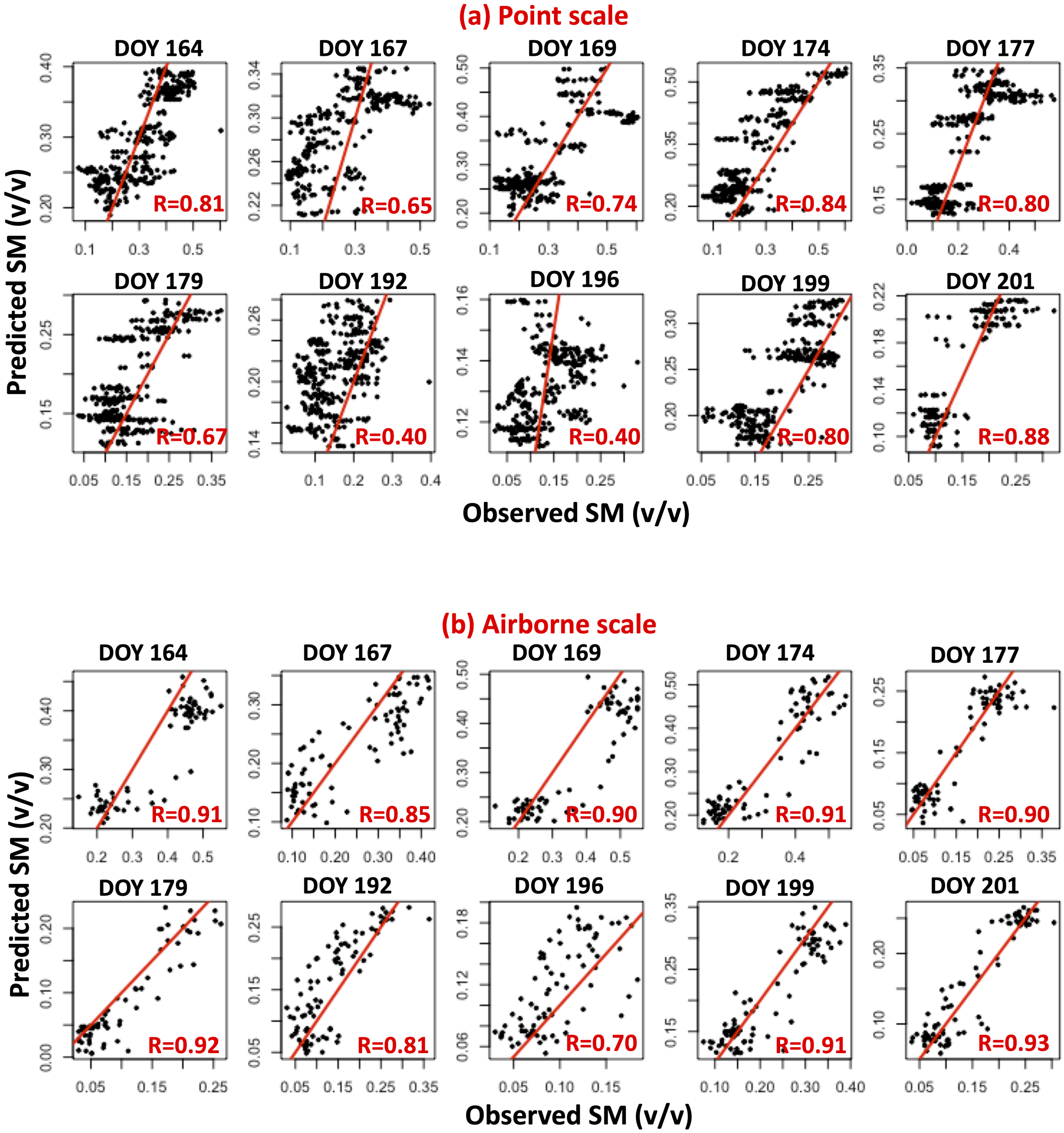

Figure 3.7: One to one plots of the hold-out observations and the average of the mean predictive posterior distribution for point and airborne support scales. The red line denotes the 1:1 line [2].

Table 3.4: Root mean squared error (v/v) for Soil moisture predictions for point (hold-out region), airborne (hold-out region) and satellite (four SMOS pixels) support scales [2].

| DOY | Point | Airborne | Satellite |
|---|---|---|---|
| 164 | 0.071 | 0.068 | 0.004 |
| 167 | 0.096 | 0.064 | 0.011 |
| 169 | 0.099 | 0.074 | 0.010 |
| 174 | 0.096 | 0.056 | 0.018 |
| 177 | 0.076 | 0.037 | 0.020 |
| 179 | 0.059 | 0.033 | 0.020 |
| 192 | 0.081 | 0.050 | 0.011 |
| 196 | 0.054 | 0.038 | 0.012 |
| 199 | 0.055 | 0.040 | 0.012 |
| 201 | 0.032 | 0.029 | 0.005 |

the control affects the mean and if the 95% HPD interval excludes zero (or is not close to zero), we assume that, subject to some uncertainty, there is evidence in the data that the control affects the mean. There is little evidence in the data that LAI (as a proxy for vegetation) affects the mean structure of SM for SMAPVEX12 on any day whereas there is evidence in the data that soil and elevation, represented by the first principal component of percent clay, percent sand and elevation, affect the mean on the majority of the days. There is little evidence in the data that precipitation affects the mean of SM on any day. This can be attributed to the fact that the IDW interpolation of sparse rain-gauge does not reflect the true rainfall patterns. Since, precipitation is a major driver of land surface processes, better spatial representation of rainfall dynamics (such as by using a stochastic rainfall model) will improve the applicability of the data fusion algorithm.

Figure 3.9 depicts the 95% HPD intervals for $\delta_{LAI}^{low}$ and $\delta_{LAI}^{high}$ for the variance and correlation (for certain lag-distances). There is little evidence that LAI affects the variance of SM but there is evidence that it has an effect on correlation during the later stages of the crop-growth season. There is evidence that soil and elevation affect the variance/correlation of SM, though in the later stages, when the crops begin to gain maturity, the 95% HPD is very close to one (Figure 3.10). Therefore, it can be hypothesized (with some uncertainty) that the land surface controls induce non-stationarity in the covariance structure of SM. This can be attributed to high spatial heterogeneity

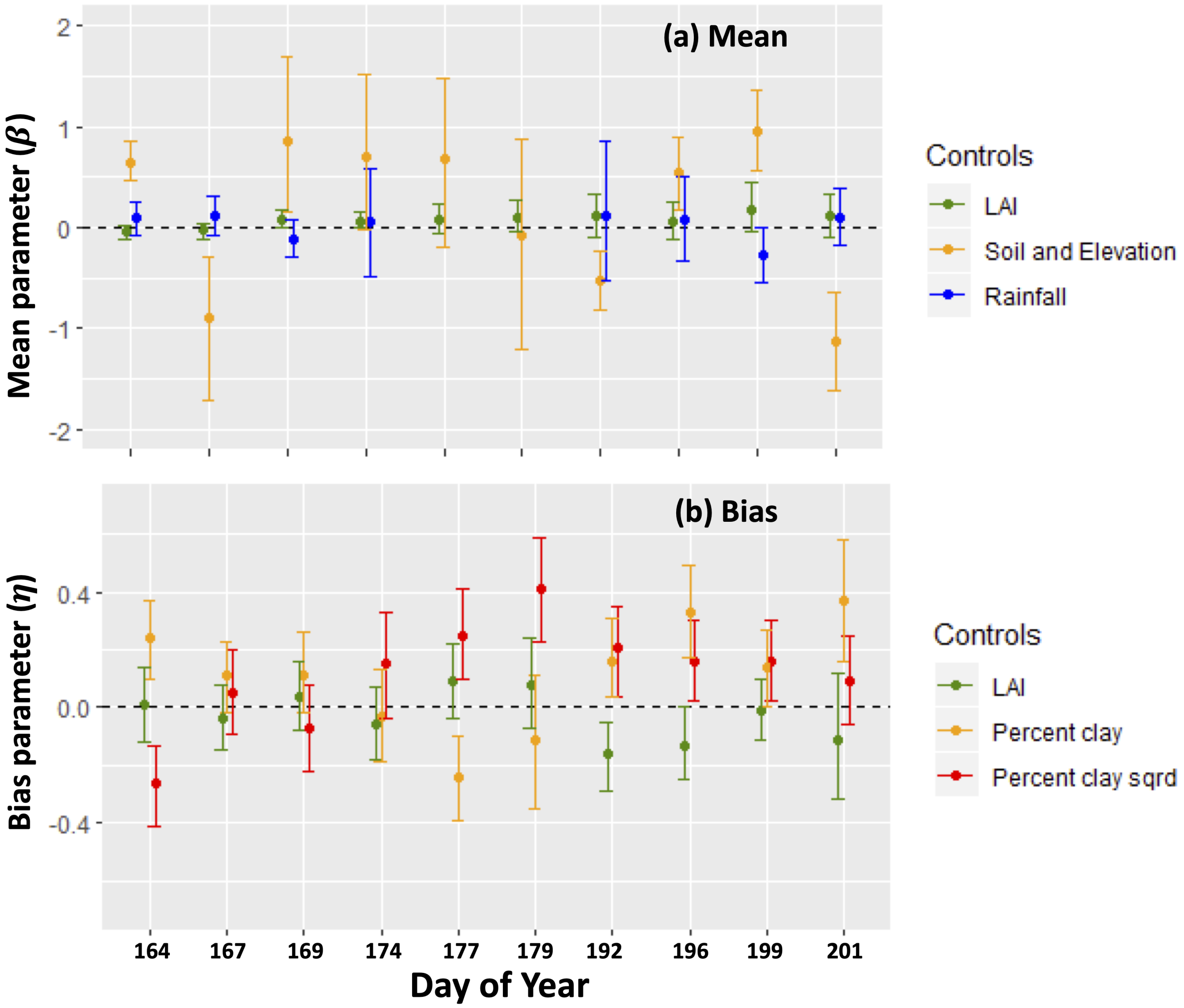

Figure 3.8: The dots represent the average of the posterior distribution while the error-bars denote the 95% highest posterior density intervals for the mean (top) and bias (bottom) for the physical controls. If the HPD region includes zero (or is very close to zero), we assume that there is little evidence in the data that the control affects the mean (bias) and if the 95% HPD interval excludes zero (or is not close to zero), we assume (with some uncertainty) that there is evidence in the data that the control affects the mean (bias) [2].

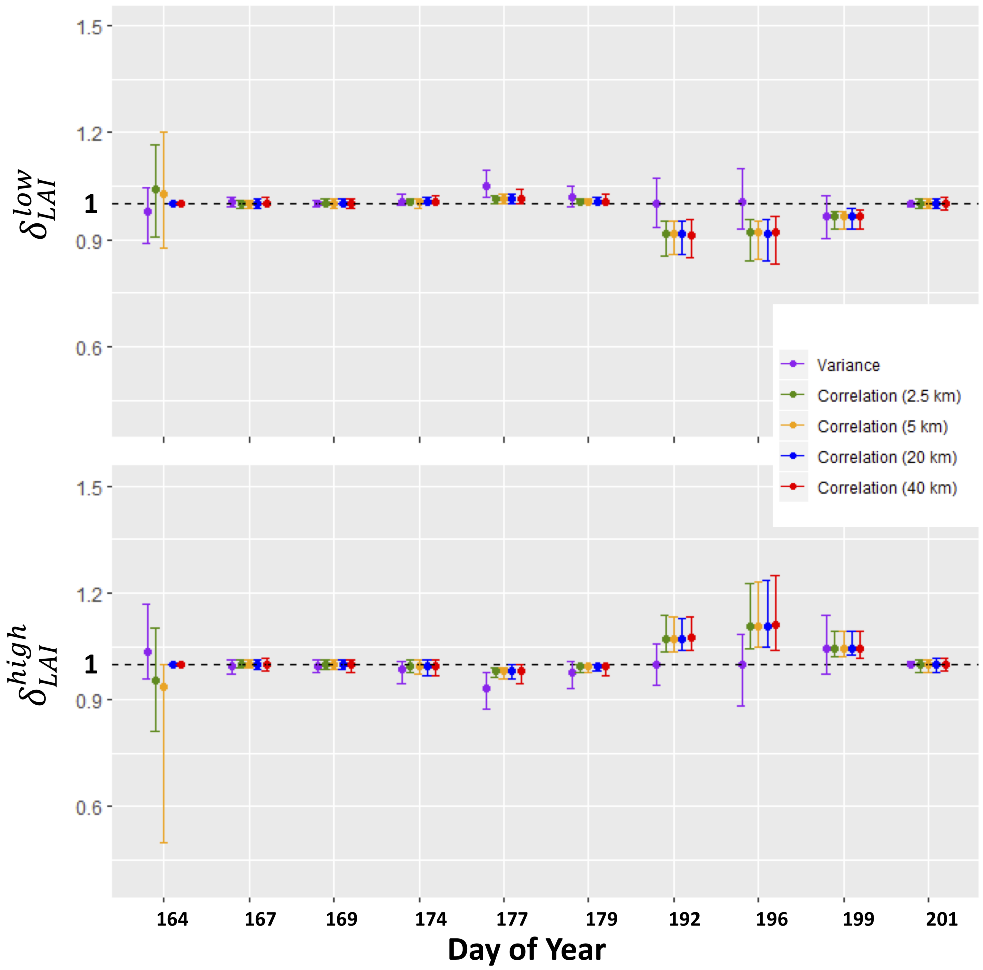

Figure 3.9: The dots represent the average of the posterior distribution while the error-bars denote the 95% highest posterior density (HPD) intervals for $\delta_{LAI}^{low}$ and $\delta_{LAI}^{high}$. If the HPD region includes one (or is very close to one), we assume that there is little evidence in the data that LAI affects the covariance and if the 95% HPD interval excludes one (or is not close to one), we assume that (with some uncertainty) that there is evidence in the data that LAI affects the covariance [2].

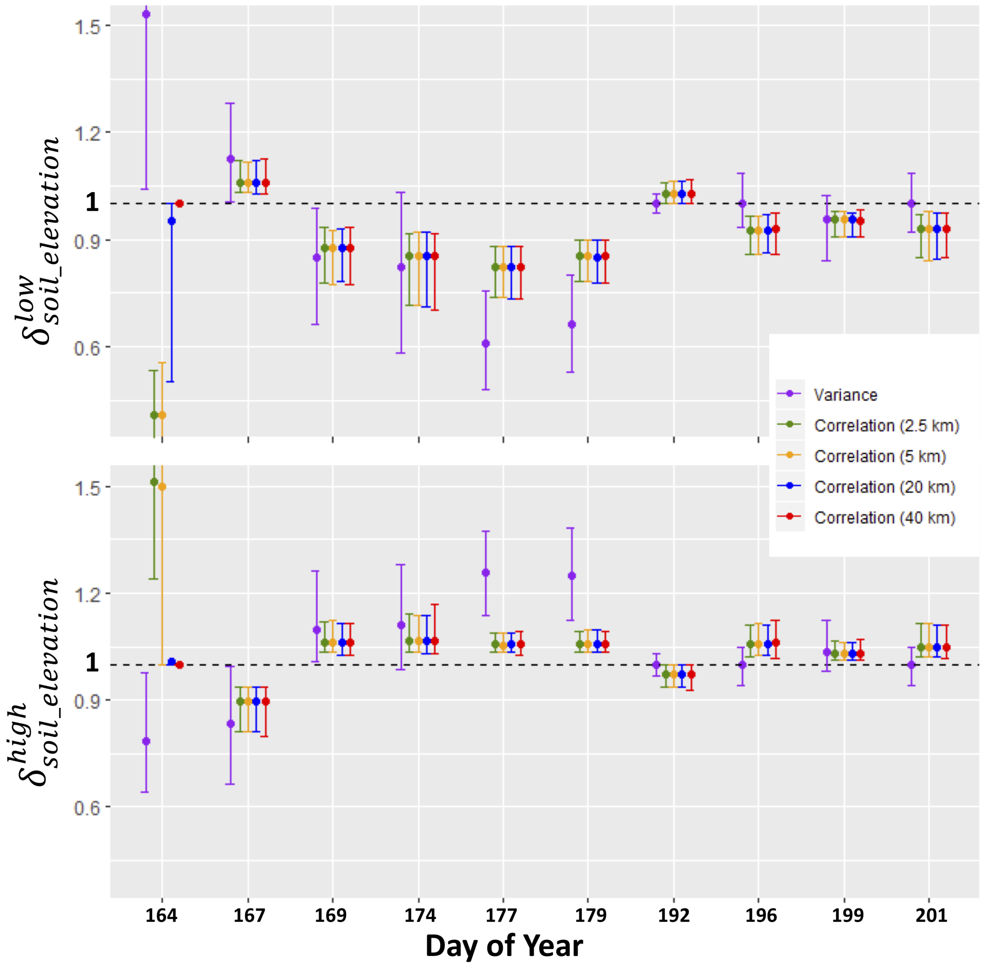

Figure 3.10: The dots represent the average of the posterior distribution of the variance/correlation while the error-bars denote the 95% highest posterior density intervals for $\delta^{low}_{soil\_elevation}$ and $\delta^{high}_{soil\_elevation}$. If the HPD region includes one (or is very close to one), we assume that there is little evidence in the data that soil and elevation affects the covariance and if the 95% HPD interval excludes one (or is not close to one), we assume (with some uncertainty) that there is evidence in the data that soil and elevation affect the covariance [2].

both in terms of soil texture and vegetation in SMAPVEX12 leading to significant variability in SM over short distances [70].

### 3.6.3 Effect of controls on PALS soil moisture bias

The 95% HPD intervals of the posterior distribution of the bias parameters ($\boldsymbol{\eta}$) are given in Figure 3.8. In general, there is little evidence that LAI affects the bias. This is because low-frequency L-band sensors are in general less affected by vegetation than high-frequency sensors such as those operating at C- and X-band [93]. Subject to some uncertainty, there is evidence in the data that the PALS bias is affected by both first and second order of % clay content on the majority of the days, though the 95% HPD region of second-order effects gets close to zero as the growing season progresses.

The average of the posterior distribution of bias ($\boldsymbol{\Delta}$) for all the observed airborne pixels (excluding the hold-out data) is given in Figure 3.11. Except DOY 174, it can be seen that the majority of the pixels (depicted by the peak of the histogram) have a negative bias, i.e the PALS sensor under-estimates the SM for most of the airborne pixels.

### 3.6.4 Predictions at intermediate scales

Since the data fusion scheme predicts well at each of the three scales, we can use the SHM to predict the latent SM at the intermediate scales following equation 3.14. As an illustration, Figure 3.12 gives the average of the mean posterior predictive distribution for 1.5 km, 3 km, 9 km and SMOS support respectively. Such predictions are crucial as multi-scale SM distribution is typically required from local to regional levels for varied applications. Predictions at the edges are prone to high uncertainty as they are very far away from *in situ* and airborne data. Since we are assuming a continuous (Gaussian Process) SM distribution, on dry days (for instance, DOY 177 - DOY 196), this can cause predictions in dry regions become even less than zero. Thus, in our study we lower-bound the predictions at 0. This illustrates a limitation of using a Gaussian process model for a bounded variable such as SM. Though SM has been found to be generally normally distributed [116, 117, 46], this might not hold true in extreme wet and dry conditions [23]. An alternative in

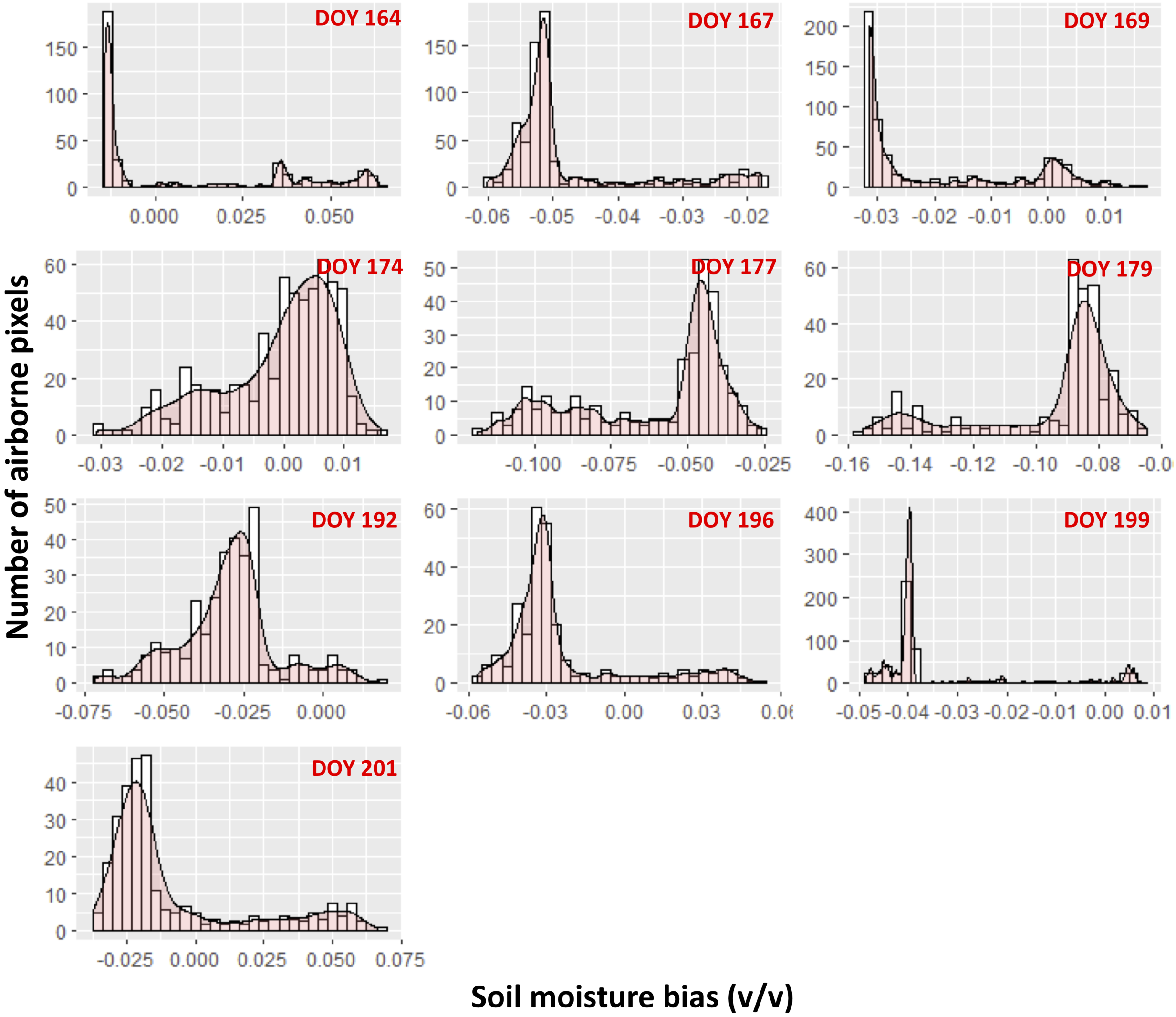

Figure 3.11: Histogram of the average of the posterior distribution of soil moisture bias for all airborne pixels. For most days, the passive-active L-band sensor underestimates the soil moisture for the majority pixels [2].

such cases is to do prior transformations on SM data such as a logit transformation to ensure that the SM distribution lies between 0 and 1.

Note that the Bayesian sampling scheme in Appendix A holds true only when the measurement error is additive to the Gaussian process model. In situations where such assumptions are untenable, the data fusion scheme can still be implemented and posterior distributions can be found using numerical procedures such as the Laplace approximation of the integral but at an added computational expense. Since, the data fusion framework is controls-driven, its applicability will be influenced by the accuracy and resolution of the governing surface and atmospheric controls. For instance, for this study, the predictions can definitely be improved with better estimates of ancillary data such as high-resolution rainfall and soil texture. A possible extension to the data fusion framework will be to also account for uncertainties/errors in these controls. Another possible area of future research will be to modify existing scaling algorithms such as those motivated by the vegetation-temperature triangular feature space [118, 99] for the mean function of the geostatistical process and model the covariance using a stationary/non-stationary covariance function.

## 3.7 Conclusions

SM data are going to become increasingly available at multiple scales and instruments in the $21^{st}$ century, and techniques to combine such data to provide a holistic picture of SM dynamics will become increasingly important. In addition to the inherent SM dynamics that vary with atmospheric and land surface conditions, each data instrument will be accompanied by its own set of errors. Data fusion using spatial hierarchical modeling is a promising methodology to account for these uncertainties and more importantly, make a distinction between the inherent SM variability and measurement errors of data instruments, to ensure that incorrect scientific inferences are avoided. We demonstrate that the proposed fusion scheme can be used to fuse SM data across multiple instruments while accounting for the above-mentioned uncertainties. Such a technique can also be used to characterize measurement errors occurring in SM retrievals from other data instruments (using appropriate data models) such as cosmic-ray neutron probes and GPS sensors though the algorithm needs to be validated with existing SM datasets from such sensors in different

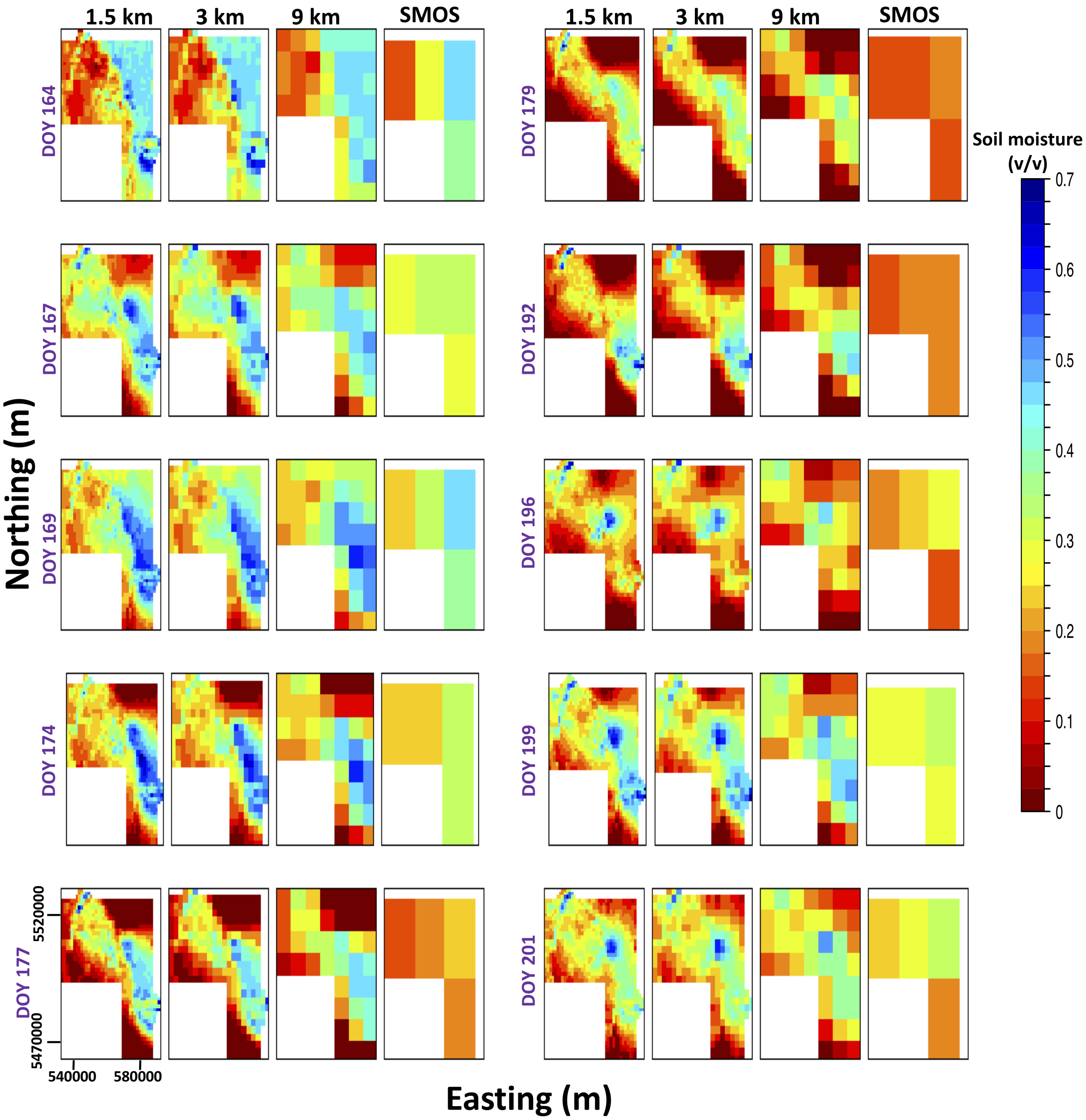

Figure 3.12: Prediction of soil moisture at different spatial supports [2].

hydroclimatic conditions to vet its wider applicability.

Data fusion using SHM is not a black-box approach and depends crucially on the choice of the mean, covariance and measurement error functions as well as the prior distributions of the parameters. Therefore, while fusing data, care must be taken to select the model parameterization which best approximates reality. In regions with sparse or no *in situ* data, aggregate SM data can still be fused using the framework presented in this paper. In such cases, though the model is expected to perform well at the support of aggregate data, the point scale behavior of the process will be influenced more by the choice of the prior distributions. Although we applied the fusion scheme on a watershed scale, the real potential of the data fusion framework lies in its applicability to fuse SM data in an operational environment at large scales such as fusing data from intermediate sensors such as GNSS-R and SMAP/Sentinel-1, and coarse radiometer products derived from SMOS and SMAP using Big Data geostatistical techniques [61, 62]. In such a case, the SHM can be further extended to a spatio-temporal setting by defining spatio-temporal mean, bias and covariance functions. Since we are working in a Bayesian framework, the posterior parameter distributions found in data-rich regions can be used as strong priors in regions with similar land surface heterogeneities and hydroclimates to aid in parameter estimation. The SM predictions resulting from fusion of all available data can also be assimilated in a physical model to potentially improve estimates of root-zone soil moisture and other hydrologic processes (for an example, refer [60]). To quote [106]: "Hierarchical statistical modeling is where data, Science and uncertainty join forces" and can therefore serve as a valuable tool to optimally combine disparate SM (and other environmental) data across multiple scales.

# 4. A MULTISCALE SPATIO-TEMPORAL BIG DATA FUSION ALGORITHM FROM POINT TO SATELLITE FOOTPRINT SCALES

## 4.1 Synopsis

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.

## 4.2 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 [75, 119]. 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” [120] in global weather prediction. Data assimilation has also found success in oceanography [121, 122] and land-surface hydrology [123]. Physical models are vital for predicting variables poorly observed by RS platforms such as ocean mixed layer [124] and root-zone soil moisture (SM) [125]. However, the rigid parameterization of physical models can be a hindrance when knowledge of the underlying spatio-temporal process is incomplete [126]. The resulting predictions can suffer from signatures of strong (and sometimes incorrect) assumptions [104]. Moreover, RS observations usually need to be pre-processed for correcting bias and scale-mismatch before assimilation in the numerical model [87].

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 [127, 128, 129, 130, 131]. Though classical ML models are inept at accounting for spatio-temporal dependence, recent research in Deep Learning seems promising [132, 133, 131]. Accuracy without interpretability, however, is insufficient [134]; 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 [134].

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 [135, 136], geostatistics has been widely used in modeling the spatio-temporal distribution of environmental variables including precipitation [137], temperature [138], soil properties [139, 140, 141, 142], carbon dioxide [143], ground-water quality [144] and SM [46, 1, 18]. Recent work on covariate-driven non-stationary models have also enabled the seamless integration of covariates into geostatistical models [64, 145] 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 [146] and inverse-covariance matrices [147]. Among these, the Vecchia approximation [148] 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 [149]. Moreover, recent work [150, 151] 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 [65, 2]. 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 [152], streamflow [153] and evapotranspiration (ET) [104]. 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. [2] 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 [2]. The rest of the chapter is organized as follows. We describe the SM datasets used in the case study in Section 4.3. The data fusion algorithm along with its big data extension is detailed in Section 4.4. This is followed by the discussion of results in Section 4.5 before we conclude in Section 4.6. Note that in the following sections, all vectors are assumed to be column vectors.

## 4.3 Study Area and Data

### 4.3.1 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) [154] and Soil Climate Analysis Network (SCAN) [155] with satellite data from Soil Moisture Ocean Salinity (SMOS) [156] and Soil Moisture Active Passive (SMAP) [157] 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 4.1).

Both SMOS and SMAP use L-band radiometers to measure surface brightness temperature ($T_b$) at an average revisit time of three days [158, 159]. 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/m2 [160]. 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 [161]. Soil and elevation data were extracted from Soil Survey Geographic Database (1 km) [162] and Leaf Area Index (LAI) (as a proxy for vegetation) were extracted from Moderate Resolution Imaging Spectroradiometer (MCD15A3H, 500m) [163].

## 4.4 Methodology

### 4.4.1 Multiscale data fusion

Let the environmental variable varying across space and time (such as SM, ET, temperature, etc.) be denoted by $y$. We assume that $y(.)$ is a Gaussian Process (GP) (a standard geostatistical assumption) at the point scale in a domain or extent D in $d$ dimensions ($d = 1, 2, 3 \dots$). For instance, if $y$ represents daily land-surface temperature (LST) varying spatially (latitude and longitude) and temporally (days), then d equals 3. The variable $y$ is defined at the point scale using a mean function $\mu$ and a covariance function $C$

$$y(.) \sim GP(\mu, C) \tag{4.1}$$

For any environmental variable $y$, in addition to point data, we might observe data at aggregate resolutions from RS platforms or large-scale numerical models. For instance, surface SM

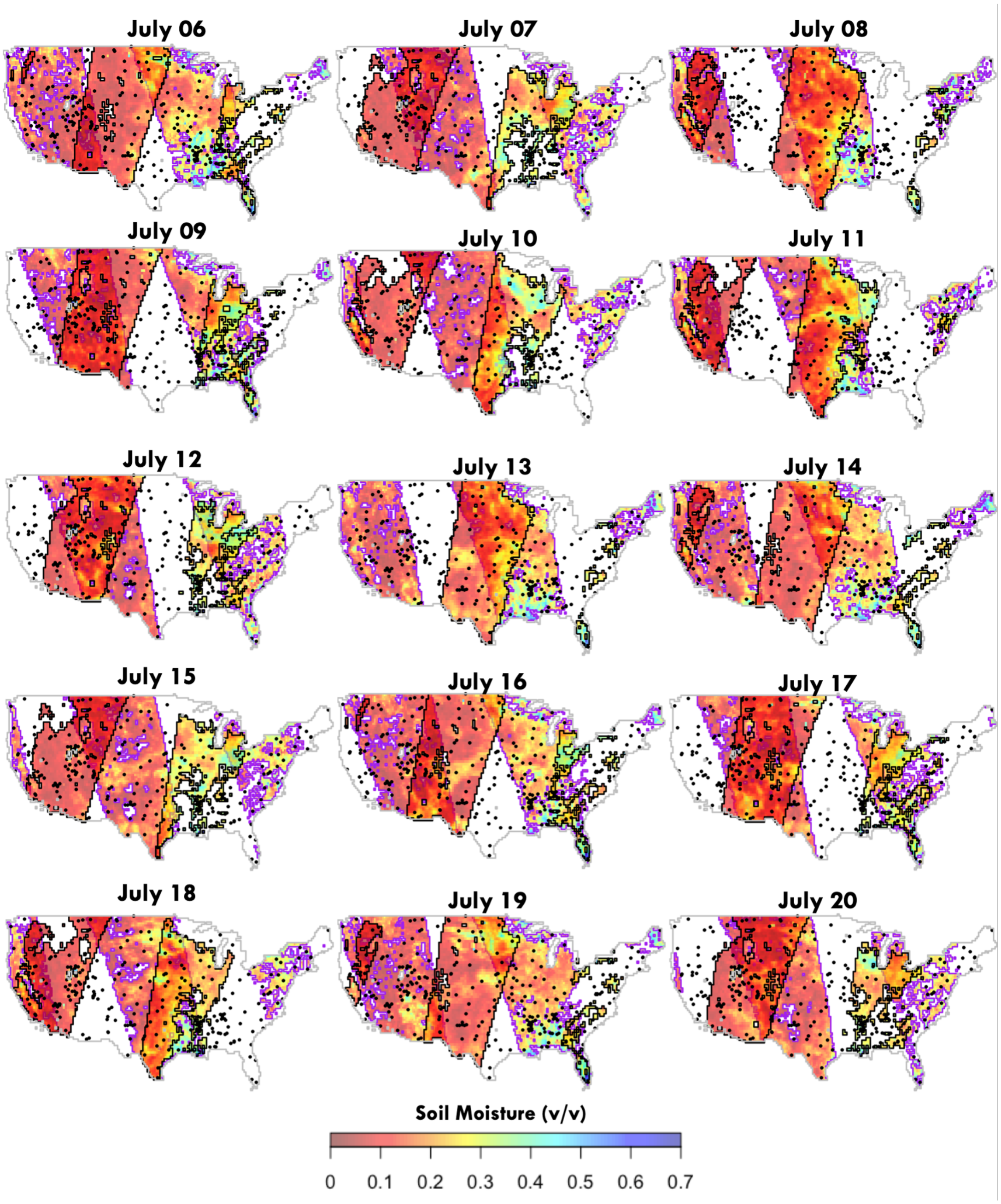

Figure 4.1: Fifteen day soil moisture data from USCRN and SCAN (black cross), SMOS (swath - black outline) and SMAP (swath - purple outline) for July 06-20, 2017. For individual days, both SMOS and SMAP observe different regions of Contiguous US (CONUS) and there is a significant overlap between the data. The size of the SM data and the extent of study domain (CONUS) are both massive making data fusion computationally demanding.

is observed at aggregate resolutions from SMAP ($\sim$36 km × 36 km, daily) and SMOS ($\sim$25 km × 25 km, daily) while ET is observed using ECOSTRESS ($\sim$70m × 70m, daily) and MODIS ($\sim$500m × 500m, 8-day). Since $y$ is defined at point 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 mean and covariance as:

$$\begin{aligned} \mu(A_i) &= \frac{1}{|A_i|}\int_{A_i} \mu(s)ds \\ C(A_i, A_j) &= \frac{1}{|A_i||A_j|}\int_{A_i}\int_{A_j} C(s_1, s_2)ds_2 ds_1 \end{aligned} \tag{4.2}$$

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

Let the total number of observed pixels be $n$ and be denoted by $A = \{A_1, \ldots, A_n\}$ with $A_i \subset D$. The joint distribution of $y(A) = (y(A_1), \ldots, y(A_n))$ can be shown to be multivariate normal [65]:

$$y(\mathcal{A}) = \mathcal{N}_n(\mu(\mathcal{A}), C(\mathcal{A}, \mathcal{A})), \tag{4.3}$$

where $\mu(A)$ is a vector of length $n$ and $C(A, A)$ is a matrix of size $n \times n$. The individual elements of $(\mu(\mathcal{A}))_i$ and $(C(\mathcal{A}, \mathcal{A}))_{ij}$ are given by equation 4.2. Since we cannot always analytically solve the above integrals, we use a numerical approximation [65] by assuming an equidistant numerical grid $\mathcal{G}$ over the extent $\mathcal{D}$ with $n_\mathcal{G}$ number of grid points such that $\mathcal{G} = \{g_1, \ldots, g_{n_\mathcal{G}}\}$ or equivalently $\mathcal{G} = \{g_k : k = 1, \ldots, n_\mathcal{G}\}$. Here $g_k$ denotes the location of the $k^{th}$ grid point in $\mathcal{G}$. We can then approximate $y(A_i)$ as:

$$y(A_i) \approx \frac{1}{n_{A_i}} \sum_{g_k \in \mathcal{G}_{A_i}} y(g_k) \tag{4.4}$$

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:

$$\begin{aligned} \mu(A_i) &\approx \frac{1}{n_{A_i}} \sum_{g_k \in \mathcal{G}_{A_i}} \mu(g_k) \\ C(A_i, A_j) &\approx \frac{1}{n_{A_i}} \frac{1}{n_{A_j}} \sum_{g_k \in \mathcal{G}_{A_i}} \sum_{g_l \in \mathcal{G}_{A_j}} C(g_k, g_l) \end{aligned} \tag{4.5}$$

We illustrate the numerical approximation using a hypothetical example in Figure 4.2. Figure 4.2 (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 4.2 (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 4.2 (c) are given by equation 4.5. 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 4.2 (d) is simply given as $\mu(A_3)$ while $C(A_1, A_3)$ is given by $\frac{1}{n_{A_1}} \sum_{g_k \in \mathcal{G}_{A_1}} C(g_k, A_3)$.

We can write $\frac{1}{n_{A_i}} \sum_{g_k \in \mathcal{G}_{A_i}} y(g_k)$ (equation 4.4) 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 $h_{A_i} = (1/n_{A_i}, ....., 1/n_{A_i})$, 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 4.5, $\mu(A_i)$ can be written as $h_{A_i}^T \mu_{A_i}$ (with $\mu_{A_i}$ having elements $\{\mu(g_k) : g_k \in G_{A_i}\}$). We also write $C(A_i, A_j)$ in equation 4.5 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 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 [164, 165] for SM, [166, 167] for LST, [168] for water storage, [169, 170] 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

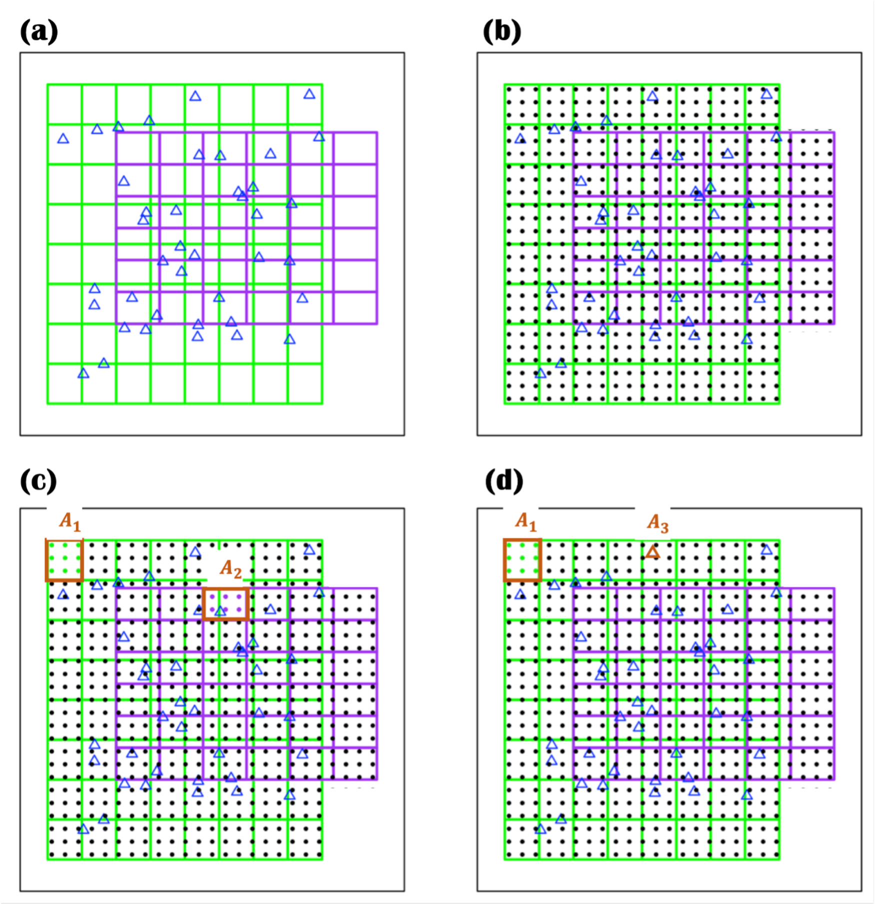

Figure 4.2: (a) Example depicting two areal (green and purple) and one point (blue triangles) data platforms (b) Equidistant point grid assumed throughout the study domain (c) The mean and covariance of pixels $A_1$ and $A_2$ approximated using the numerical grid (d) The mean and covariance between a pixel $A_1$ and point observation $A_3$.

parameterized errors. For a given observation $z(A_i)$ (from a data platform) for pixel $A_i$, we thus write:

$$z(A_i) = y(A_i) + \delta(A_i) + \kappa(A_i)y(A_i) + \epsilon(A_i) \tag{4.6}$$

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 $\epsilon(A_i) \sim \mathcal{N}(0, \tau^2_{A_i})$ with variance $\tau^2_{A_i}$. We then write:

$$\begin{aligned} z(A_i) &\approx h^T_{A_i} y_{A_i} + \delta(A_i) + \kappa(A_i) h^T_{A_i} y_{A_i} + \epsilon(A_i) \\ &= (1 + \kappa(A_i)) h^T_{A_i} y_{A_i} + \delta(A_i) + \epsilon(A_i) \\ &= (h^{\kappa}_{A_i})^T y_{A_i} + \delta(A_i) + \epsilon(A_i), \end{aligned} \tag{4.7}$$

where $h^{\kappa}_{A_i} = (1 + \kappa(A_i)) h^T_{A_i}$. The mean ($\mu$) and covariance function ($C$) in equation 4.1 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 ($\tau^2_{A_i}$) for a pixel $A_i$ in equation 4.7 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 $\theta$. Elements of $\theta$ 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 $\theta$ is estimated by maximizing the likelihood $f(z(A)|\theta)$ 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:

$$-2log(f(z(\mathcal{A})|\theta) = log(det(\Sigma_z)) + (z(\mathcal{A}) - \mu_z)^T \Sigma_z^{-1} (z(\mathcal{A}) - \mu_z) + nlog(2\pi), \tag{4.8}$$

where the $i^{th}$ element of the vector $\mu_z$ (size $n$) and the $(i, j)^{th}$ element of the matrix $\Sigma_z$ (size $n \times n$) in equation 4.8 are given as:

$$\begin{aligned}\mu_{z,i} &\approx (h^{\kappa}_{A_i})^T \mu_{A_i} + \delta(A_i) \\ \Sigma_{z,i,j} &\approx (h^{\kappa}_{A_i})^T C(\mathcal{G}_{A_i}, \mathcal{G}_{A_j}) h^{\kappa}_{A_j} + \tau^2_{A_{i,j}},\end{aligned} \tag{4.9}$$

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.

### 4.4.2 *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 $\Sigma_z$ and finding its inverse $\Sigma_z^{-1}$ in equation 4.8 requires $\mathcal{O}(n_G^2) + \mathcal{O}(n^3)$ 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 $f(z(A)|\theta)$ as a product of univariate conditional distributions as

$$f(z(\mathcal{A})|\theta) = f(z(A_1)|\theta) \times \prod_{i=2}^{n} f(z(A_i)|\boldsymbol{z(A_{1:i-1})}, \theta), \tag{4.10}$$

where $\boldsymbol{A_{1:i-1}}$ denotes $\{A_1, \ldots, A_{i-1}\}$ and thus $\boldsymbol{z(A_{1:i-1})}$ denotes $\{z(A_1), \ldots, z(A_{i-1})\}$. Following [148] we approximate the likelihood $f(z(\mathcal{A})|\theta)$ as:

$$\hat{f}(z(\mathcal{A})|\theta) = f(z(A_1)|\theta) \times \prod_{i=2}^{n} f(z(A_i)|\boldsymbol{z(A_{m_i})}, \theta), \tag{4.11}$$

where $\boldsymbol{A_{m_i}}$ is a subvector of $\boldsymbol{A_{1:i-1}}$ 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 4.10. The elements of subvector $\boldsymbol{A_{m_i}}$ consist of $m_i$ elements from $\boldsymbol{A_{1:i-1}}$ which are closest to $A_i$ in space. The subvector $\boldsymbol{z(A_{m_i})}$ is the observed data vector corresponding to $\boldsymbol{A_{m_i}}$ . To illustrate the approximation, we again use the hypothetical example in Figure 4.2 (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 4.3 using a random permutation of the pixels A and choosing $m = 20$. Column (a) presents the conditional distributions in equation 4.10 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 4.10 and 4.11 (where $i = 2, ..., 140$ increases from top to bottom in the columns) is color-filled in red while the pixels (or points) of the conditioning vector $\boldsymbol{z(A_{1:i-1})}$ (equation 4.10) or $\boldsymbol{z(A_{m_i})}$ (equation 4.11) are color-filled in green ($R_1$), purple ($R_2$) and blue ($P_1$). It can be seen in Figure 4.3 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 $\Lambda(\Lambda^T\Lambda = \Sigma_z^{-1})$. This leads to fast evaluation of $\Sigma_z^{-1}$ (and consequently the likelihood) in equation 4.8 used for estimating the parameter vector $\theta$ as well as doing subsequent predictions. The detailed algorithm for parameter estimation and subsequent predictions is given in Appendix B. We call this approximation *Vecchia-multiscale*.

#### *4.4.2.1 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 $m << n$. Further, equation 4.11 results in a product of independent univariate distributions which is readily parallelized for faster computations.

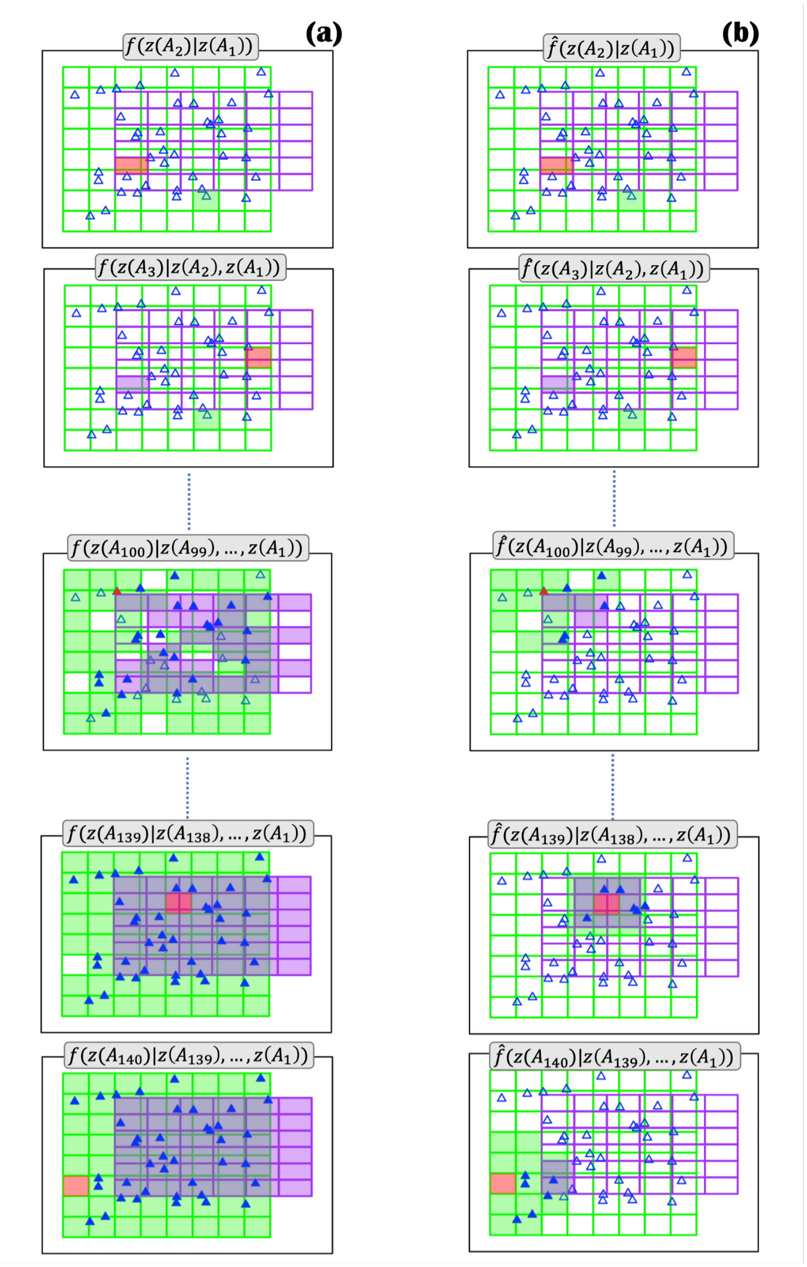
(a)
$f(z(A_2)|z(A_1))$
$f(z(A_3)|z(A_2),z(A_1))$
$f(z(A_{100})|z(A_{99}),\dots,z(A_1))$
$f(z(A_{139})|z(A_{138}),\dots,z(A_1))$
$f(z(A_{140})|z(A_{139}),\dots,z(A_1))$
(b)
$\hat{f}(z(A_2)|z(A_1))$
$\hat{f}(z(A_3)|z(A_2),z(A_1))$
$\hat{f}(z(A_{100})|z(A_{99}),\dots,z(A_1))$
$\hat{f}(z(A_{139})|z(A_{138}),\dots,z(A_1))$
$\hat{f}(z(A_{140})|z(A_{139}),\dots,z(A_1))$

Figure 4.3: Illustration of the Vecchia-multiscale to the hypothetical data in Figure 4.2 (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 $i = 2, ..., 140$ 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 ($P_1$).

Regarding accuracy for a fixed value of m, as the right side of equation 4.11 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 4.11, for a pixel $A_i$ ($i \leq 2$), we select the subset $A_{m_i}$ (of length $m_i$) from elements of $A_{1:i-1}$ which are closest in space to $A_i$. This leads to different values for $\boldsymbol{z(A_{m_i})}$ in equation 4.11 based on how we permute $\{A_1, ..., A_n\}$. Thus, the approximation accuracy will depend upon what permutation of $\{A_1, ..., A_n\}$ we choose for the pixels (and points) for computing $\hat{f}(z(\mathcal{A})|\theta)$ in equation 4.11. When the size of the multiscale data is massive, it is infeasible to explore all such permutations. For point data, [149] 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 [149]: 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 Appendix C and [149] 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 4.4. We assume the centroid of an areal pixel as its location for applying the permutations.

Using the hypothetical example in Figure 4.2 (a), we illustrate the effect of these eight chosen permutations on how the pixels and points are ordered in $\mathcal{A}$ and how it affects the evaluation of $\hat{f}(z(A)|\theta)$. 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 Appendix C.

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 [171, 172] and it should be used with caution when using *Vecchia-multiscale*. The subvector $\boldsymbol{A}_{m_i}$ (equation 4.11) consists of a good mix of both far and near pixels as well as nearby point data for *Separate-Maxmin* and *Separate-Random* (Appendix C). 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 4.1), 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 $\{A_1, ..., A_n\}$. 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.

## 4.5 Results and Discussion

### 4.5.1 Case Study: Soil Moisture Data

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 (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 $-$ to  and also make the distribution less skewed (Appendix C). Overlapping data from SMOS and SMAP during the

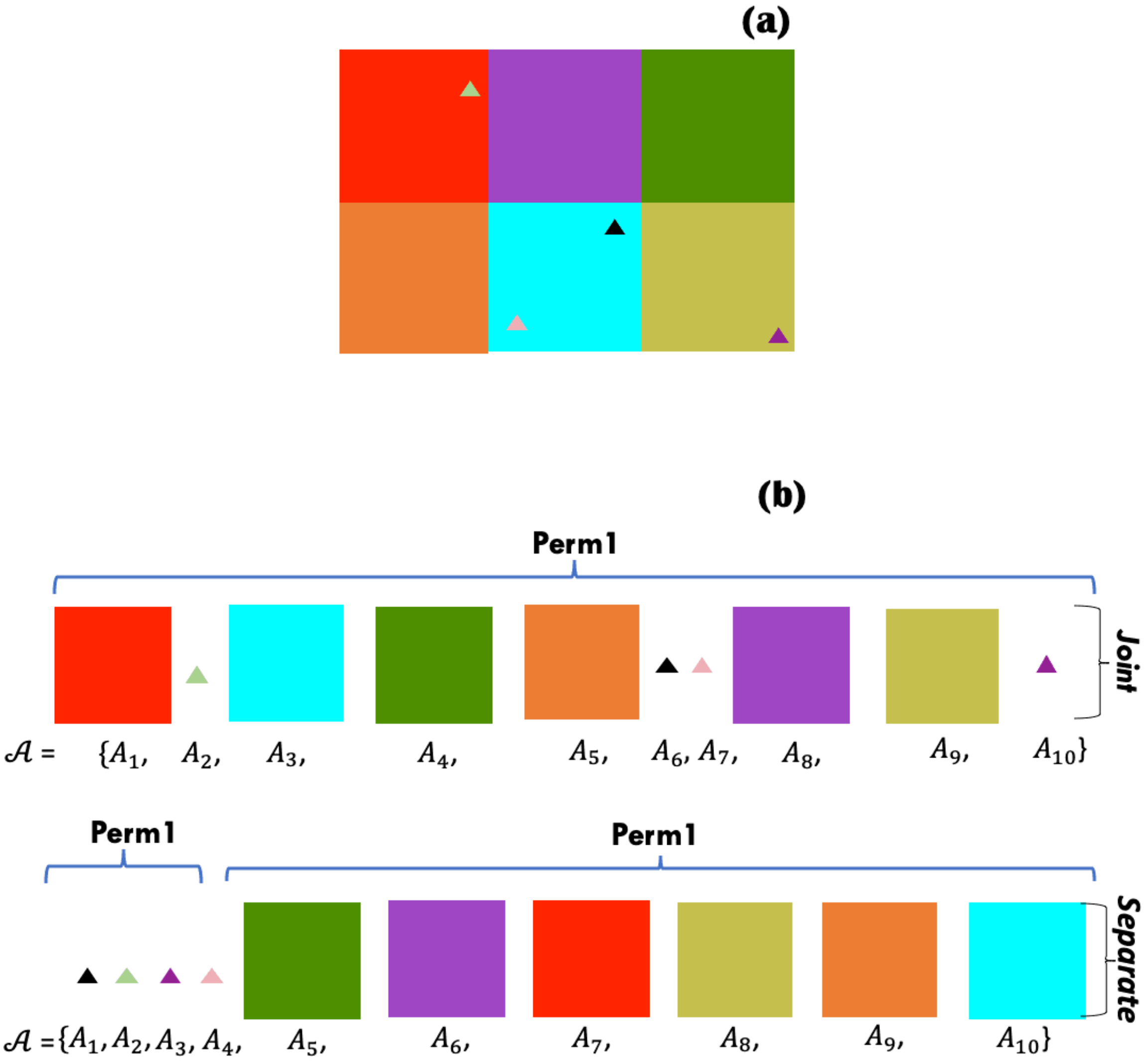
(a)
(b)
Perm1
Joint
𝒜 = {A1, A2, A3, A4, A5, A6, A7, A8, A9, A10}
Perm1
Perm1
Separate
𝒜 = {A1, A2, A3, A4, A5, A6, A7, A8, A9, A10}

Figure 4.4: 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 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.

analyzed period also exhibit slightly better correlation on the transformed scale (Appendix C).

### 4.5.2 Mean, covariance and bias

Numerous studies [13, 47, 12, 45, 14, 15, 46, 90, 16, 48, 21] 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 $\leq 5kg/m^2$, 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 (Appendix C) 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 $\tau^2_{SMAP}$ and $\tau^2_{SMOS}$ 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 exploratory analysis. For the exploratory analysis, the covariates are linearly averaged to the SMAP resolution. For rainfall, we assume 3-day antecedent mean rainfall as a covariate. On the original scale (Figure 4.5 (a)), the relationship between SM and the physical controls is non-linear. But after some non-linear transformations of the covariates (and logit transform of SM), an approximate linear relationship between SM and the covariates can be assumed (Figure 4.5 (b)). The mean trend of SM can be therefore written as:

$$\mu(log(\frac{SM}{1-SM})) = \mu(SM') = \beta_0+\beta_1 log(LAI)+\beta_2 exp(-\frac{rain}{p^{\beta}_{rain}})+\beta_3 exp(-\frac{elevation}{p^{\beta}_{elevation}}) \quad (4.12)$$

We fix $p^{\beta}_{rain}$ and $p^{\beta}_{elevation}$ as 3.3 mm and 342.6 m based on exploratory analysis. These two parameters represent the range of the exponential functions in equation 4.12 for which an approxi-

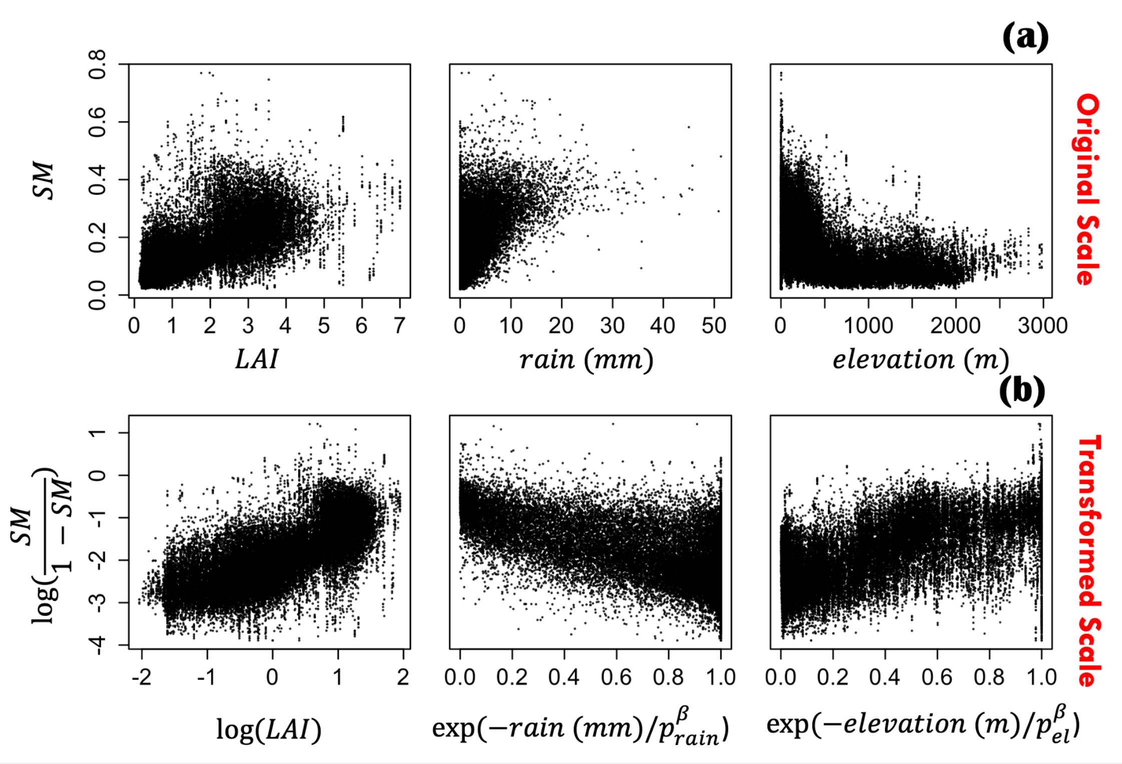

Figure 4.5: Exploratory analysis of soil moisture with physical covariates. (a) The relationship of soil moisture with the physical covariates is non-linear on the original scale. (b) Appropriate covariate transformation results in an approximate linear relationship of SM (on the logit scale) with the physical covariates. The values of $p_{rain}$ and $p_{elevation}$ are fixed as 3.3 mm and 342.6 m in the plots.

mate linear relationship holds between SM′ and the transformed covariates in Figure 4.5 (b). Note that the covariates are resampled only for exploratory analysis and no resampling of (SM and covariate) data is required for implementing the actual algorithm in Section 4.4. Since we use an equidistant grid to approximate multiscale SM data, the grid points are assigned values according to the covariate pixels in which they lie. Though this results in grid points lying in a covariate pixel getting the same values, this allows us to work with covariate data at different resolutions and avoid errors introduced due to resampling of covariate data.

The covariance between any two points $(x_1, y_1, t_1)$ and $(x_2, y_2, t_2)$, where $x$, $y$, $t$ represent the latitude, longitude and time respectively, will also vary based on the underlying covariate heterogeneity and therefore the assumption of a stationary covariance function is too simplistic. Thus, for the covariance function $C$ (equation 4.1), we use a non-stationary covariance function [90, 64] such that:

$$C(SM'(x_1, y_1, t_1), SM'(x_2, y_2, t_2)) = C(s_1, s_2) = \sum_{j=1}^{M} w_j(X_{cov}(s_1))w_j(X_{cov}(s_2))C_j(|s_1 - s_2|) \tag{4.13}$$

The covariance function in equation 4.13 is a weighted sum of $M$ isotropic covariance functions $\{C_j; j = 1, 2, \ldots.., M\}$ where the weights $\{w_j; j = 1, 2, \ldots.., M\}$ are a function of the underlying physical covariates $X_{cov}(s)$ affecting the covariance. The weighting functions $w_j$s are modeled using a multinomial logistic function of the underlying covariates: $w_j(s) = \frac{exp(X_{cov}(s)^T \alpha_j)}{\sum_{l=1}^{M} exp(X_{cov}(s)^T \alpha_l)}$. For our analysis, we choose exponential covariance functions (Matern with smoothness =0.5) for individual $C_j$s (equation 4.13) with different range parameters for space ($r^j_{xy}$) and time ($r^j_t$) [149]:

$$C_j(s_1, s_2) = \sigma_j^2 exp(-\sqrt{\frac{||(x_1, y_1) - (x_2, y_2)||^2}{(r_1^j)^2} + \frac{|t_1 - t_2|^2}{(r_2^j)^2}}) \tag{4.14}$$

We chose the exponential covariance functions for individual $C_j$s as changing the smoothness parameter for Matern resulted in insignificant change in the estimated maxmimum likelihood, and exponential functions are computationally faster to evaluate than Matern due to the added cost of evaluating the Bessel functions for the Matern function. We fix $M = 3$ to keep the number of parameters to be estimated relatively low. We include LAI, three-day mean antecedent rain, clay and elevation in $X_{cov}$. As mentioned in Section 4.4, both the mean and covariance functions are defined at point scale with computations at areal supports done as outlined in Section 4.4.1. In this work, since point data are sparse, the parameter estimates of the mean and covariance functions are expected to be mainly driven by SMAP and SMOS data. Note that we do not include latitude, longitude or time as covariates in either the mean or covariance function to make the fusion scheme more general and transferable.

### *4.5.2.1 Parameter estimation and inference*

We assume a numerical grid G (Section 4.4.1) spaced approximately $0.09$ degrees apart across the CONUS for each of the fifteen days resulting in close to $100,000$ grid points per day ($n_G \approx 15 \times 100,000 = 1,500,000$). The total number of observations $n$ from all platforms (SMAP, SMOS, and USCRN/SCAN) for fifteen days equal $100,386$. Parameter estimation and subsequent predictions by computing exact likelihood is computationally intractable for such a big dataset and thus requires an approximation. We use the approximation detailed in Section 4.4 using the *Separate-Maxmin* orderings. Since SMAP and SMOS observe SM at an interval of $3 - 7$ days, we compute the *Separate-Maxmin* ordering only considering the spatial coordinates (latitude and longitude) of the data so that the temporal information of SM is also adequately represented in the conditioning vector $z(A_{m_i})$ in equation 4.11. We fix the number of neighbors as $m = 60$; the choice of $m$ was taken to balance the predictive accuracy and computational speed. We carry out parameter estimation using a global optimization algorithm called Generalized Simulated Annealing [80], a generalized and improved form of simulated annealing, to find the parameter estimates that maximize the likelihood.

On the logit scale, the estimated mean parameters (equation 4.12) are $\beta = \{\beta_0, \beta_1, \beta_2, \beta_3\} = \{-1.71, 0.08, -0.35, 0.17\}$ thus showing a good correlation of mean SM with the controls especially antecedent rainfall. The additive and multiplicative bias for SMOS are $\delta = -0.003$ and $\kappa = 0.15$ respectively, while the measurement error variance for SMAP and SMOS are

$\tau^2_{SMAP} = 0.026$ and $\tau^2_{SMOS} = 0.023$. To quantify the effect of covariates on the spatio-temporal covariance of SM, we first transform the covariance to the original scale. For a specified covariance between two points (from the covariance function in equation 4.13) on the logit scale, we use the well-known Cholesky-Decomposition method to simulate $(50,000)$ pairs of values for these two points. We then back-transform these values to the original scale and use the empirical covariance of the pairs as an approximation of the covariance at the original scale. For the non-stationary covariance function, since the covariance between any two points depends on the lag-distance in space and time as well as the covariates($X_{cov}$), the effect of an individual covariate on the covariance is nontrivial. We thus quantify the effect of a covariate by comparing the covariance for different lags (in space and time) when the control is at the mean value (of the study domain) to when the control is at extreme value ($5^{th}$ and $95^{th}$ percentile) while keeping the other controls at their mean values [1, 64]. The resulting correlation plots are given in Figure 4.6. We find that all four covariates affect the correlation in space with higher values of rainfall, LAI, percent clay and lower values of elevation associated with increase in spatial correlation. For the temporal correlation, we found only a slight effect of the covariates on the correlation. Note, however, inclusion of other physical covariates as well as analysis of a longer time-period might show the effect of certain covariates on the temporal SM correlation.

Of course, individual plots in Figure 4.6 represent only three combinations of the physical covariates. In reality, all the covariates exhibit considerable heterogeneity across CONUS (Appendix C) and act together to give vastly different correlation patterns. To illustrate this effect, we choose 5 points (A-E, Figure 4.7) across CONUS under contrasting covariate heterogeneity and look at the spatial correlation of these points with surrounding points ($\sim$3 km apart) within an approximately 60 km × 60 km region for July 06, 2017. We see that the correlation pattern differs significantly based on the how the quartet of rainfall, LAI, clay and elevation vary in the surrounding region of the respective points.

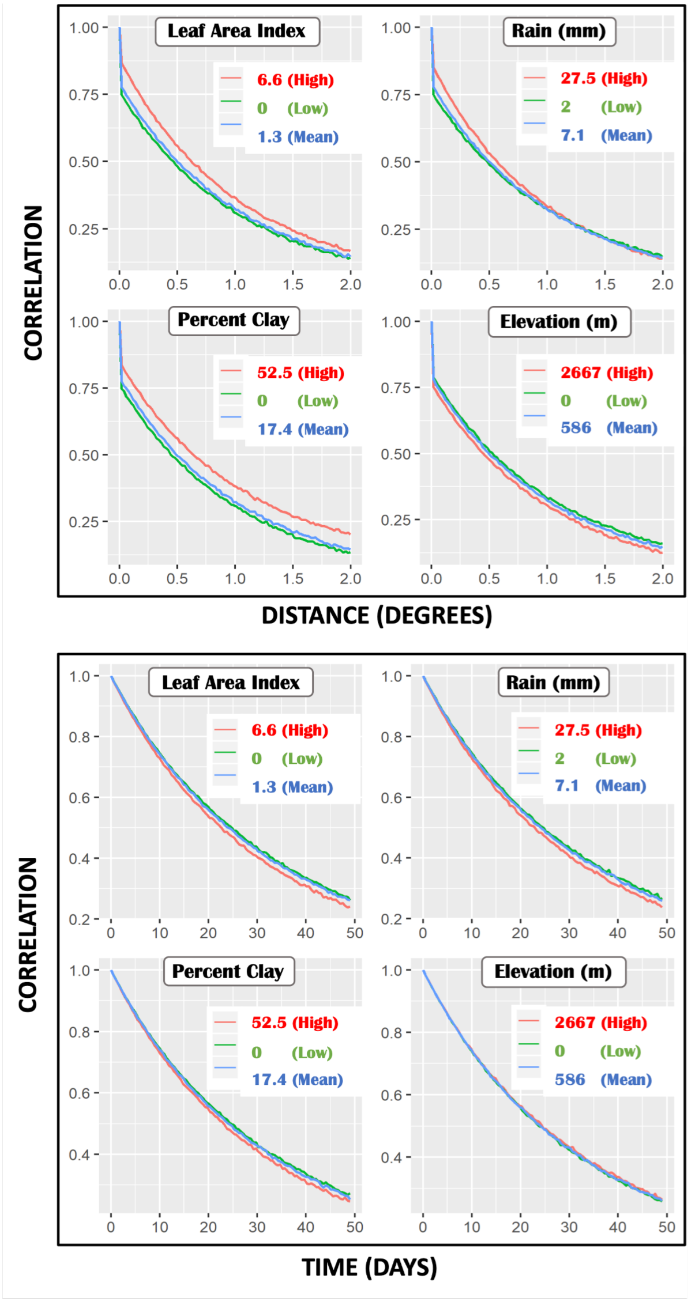

Figure 4.6: Spatial (top) and temporal (bottom) correlation plots when one of the physical covariates (Leaf Area Index, rainfall, clay and elevation) is changed from the mean value (of the study domain) to high ($95^{th}$ percentile) and low values ($5^{th}$ percentile). For each of the plots, the blue 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.

### 4.5.3 Predictions at different Scales

Once the parameters have been estimated, we compute multiscale SM predictions (Appendix B) across CONUS. As a final step, we back-transform the predictions i.e., $SM = exp(SM')/(1 + exp(SM'))$ to the original scale. We compare our SM predictions at four support scales: point (USCRN and SCAN), 3 km (SMAP/Sentinel-1), 25 km (SMOS) and 36 km (SMAP). We compute five-day SM forecasts from July 21-25, 2017 on all four support scales.

*4.5.3.1 USCRN and SCAN Scale*

As mentioned before, we randomly held out 27 USCRN and SCAN stations across CONUS (Figure 4.8) as test data. Figure 4.9 depicts the SM for the "observed" (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.

#### *4.5.3.2 SMAP/Sentinel-1 Scale*

The SMAP/Sentinel-1 L2 SM [173] product uses concurrent 36 km SMAP $T_b$ measurements and 3 km backscatter measurements from Sentinel-1 radars to give 3 km SM in the overlapping regions of the two platforms. The Sentinel-1 radars have a much narrower swath ($\sim$250 km) however, compared with the relatively wide swath (1,000 km) of SMAP which significantly reduces the spatial coverage of the SMAP/Sentinel-1 product. The average temporal revisit time of Sentinel-1 radars is 6 days and due to different revisit times of SMAP and Sentinel-1 radars, the temporal resolution of the SMAP/Sentinel-1 SM product varies from 6-12 days. Therefore, for any given day, the coverage of the SMAP/Sentinel-1 product across CONUS is quite limited.

We compute SM predictions at 3 km (assuming the equidistant grid points G to be 1 km apart) for the observed SMAP/Sentinel-1 pixels during the 20-day period and compare with the observed SMAP/Sentinel-1 product (Figure 4.10). We also compare the SMAP/Sentinel-1 observations with the SMAP product from which it is derived. We see that for the majority of the days the SM predictions agree well with the SMAP/Sentinel-1 product outperforming the original SMAP product even for the forecast period. This shows that fusing SMAP SM with SMOS (and USCRN-SCAN data) and accounting for the effects of physical covariates on SM distribution results in better predictive accuracy at 3 km support scale than just using the SMAP SM. Since the spatio-temporal coverage of SMAP/Sentinel-1 is extremely limited, predictions using the data fusion scheme are useful as they help predict SM across the entire CONUS at a daily scale.

#### *4.5.3.3 SMAP and SMOS Scale*

Since we use all of SMAP and SMOS data for the "observed" period (July 06-20, 2017) for estimating our parameters, we compare SM predictions with observed SMOS and SMAP data for the forecast period (Figure 4.11 (a)). We make predictions assuming an equidistant numerical grid spaced approximately 9 km apart and remove pixels which have less than 7 grid points lying inside the pixels. We find that the predictions satisfactorily agree with the observed SM with RMSE ranging from 0.039 v/v to 0.055 v/v for SMAP, and 0.049 v/v to 0.067 v/v for SMOS while R

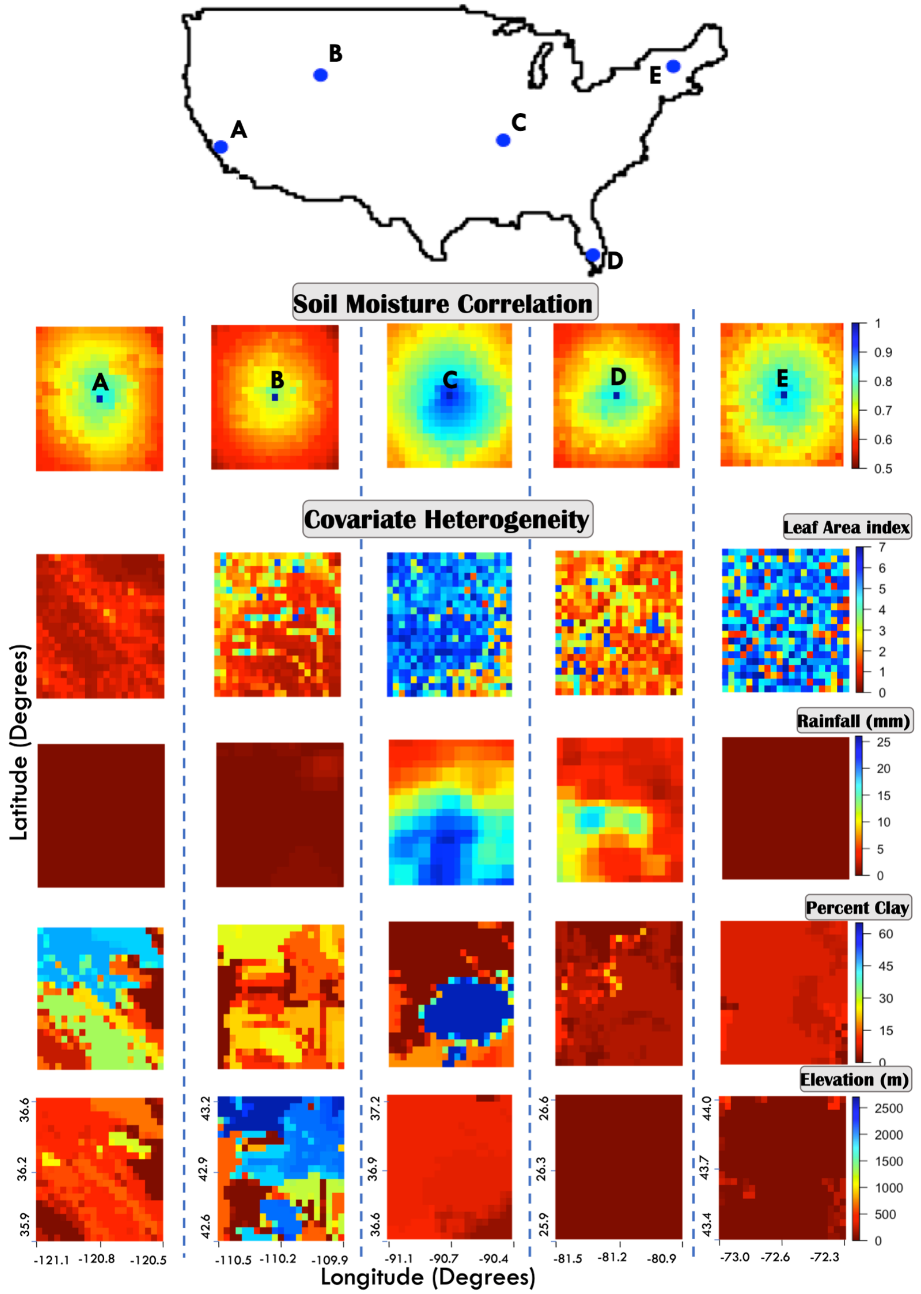

Figure 4.7: Spatial Correlation pattern of Soil moisture for five points (A-E) across Contiguous US for July 06, 2017. The correlation of the five points with their surrounding region varies considerably due to the covariate heterogeneity of the regions.

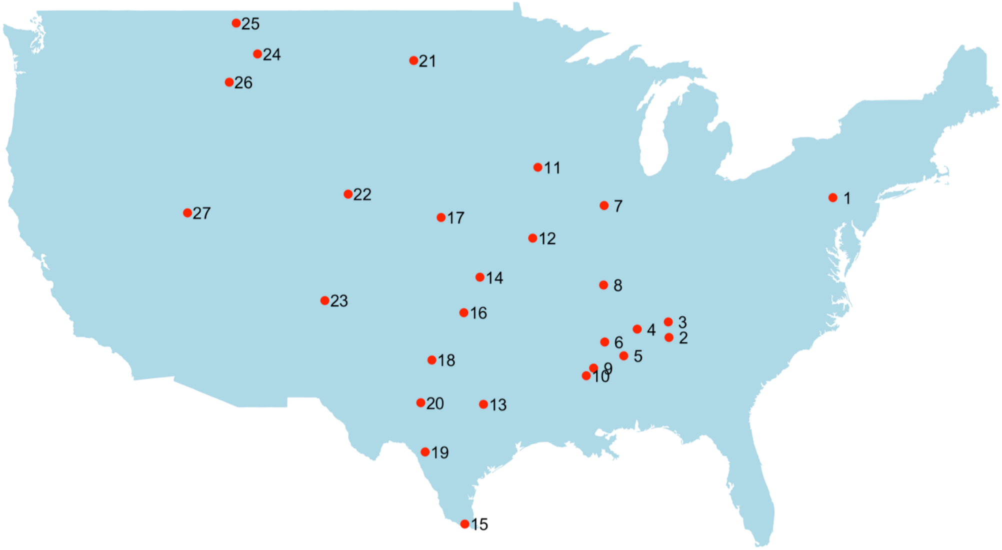

Figure 4.8: Location of the validation USCRN/SCAN stations across Contiguous US. We randomly hold out the 27 USCRN/SCAN stations to compare soil moisture predictions at the point scale across Contiguous US. The locations span different hydroclimates and surface heterogeneities.

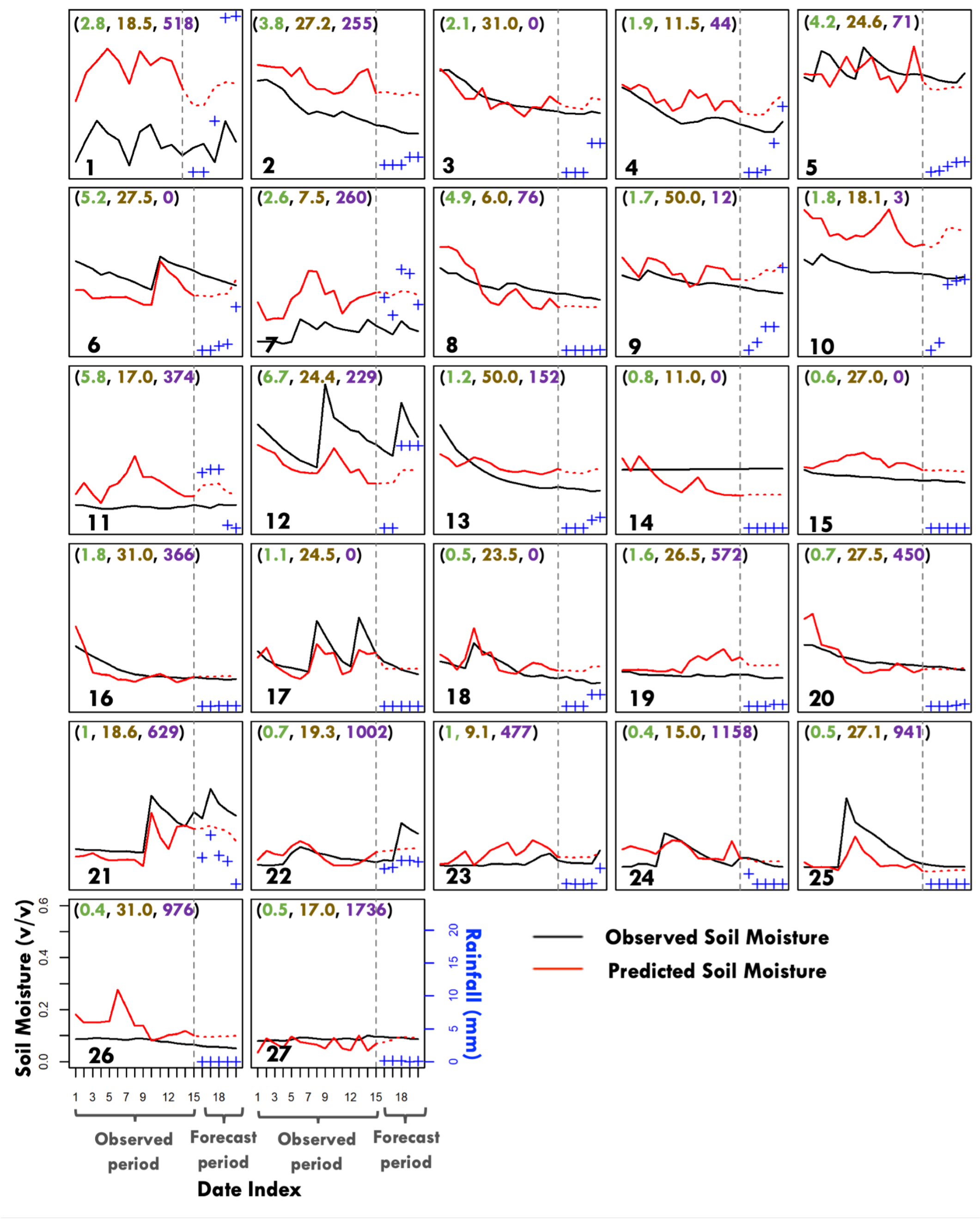

Figure 4.9: Comparison of soil moisture predictions with the observed SCAN/USRN data for the "observed" (July 06-20, 2017) and "forecast" period (July 21-25, 2017). The covariate values of LAI (averaged during the forecast period), percent clay and elevation (m) are denoted by green, brown and purple colors respectively. The three-day mean antecedent rainfall is also given in blue during the forecast period to demonstrate its effect on SM forecasts

ranging from 0.84 to 0.90 for SMAP, and 0.76 to 0.87 for SMOS. As an illustration, the mean SM predictions as well as the prediction variance for July 21, 2017 are given in Figure 4.11 (b). It should be noted that since the multiscale predictions are derived from both SMOS and SMAP, their accuracy is affected by how well the two platforms agree with each other. To get a rough estimate of this, we bilinearly interpolated the SMOS pixels which overlap with the SMAP pixels for July 21-25, 2017 and found an RMSE of 0.051 v/v to 0.076 v/v while R varied from 0.74 to 0.86.

The proposed data fusion scheme thus shows good potential for improving SM predictions across scales. Future research efforts should focus on applying the algorithm for bigger time periods and across different seasons using high performance computing systems. Improved formulations of the mean, bias and covariance functions as well as the inclusion of other physical covariates should be explored. The accuracy of the data fusion scheme at multiple scales can be improved by fusing SM estimates from other platforms such as the Cyclone Global Navigation Satellite System (CYGNSS) and the highly anticipated NASA–ISRO Synthetic Aperture Radar (NISAR) mission. The data fusion allows seamless integration of any number of platforms at varied scales; appropriate parametrization of the bias and error for individual platforms, however, is necessary. As mentioned earlier, the proposed algorithm is general and can be potentially used to fuse other spatio-temporally correlated environmental variables which have measurements available from multiple platforms.

## 4.6 Conclusions

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

As the volume of environmental data are expected to dramatically increase in the future, further research into finding better orderings becomes critical. We chose our orderings based only on space and time; future work will focus on proposing physically-based orderings where, in addition to the

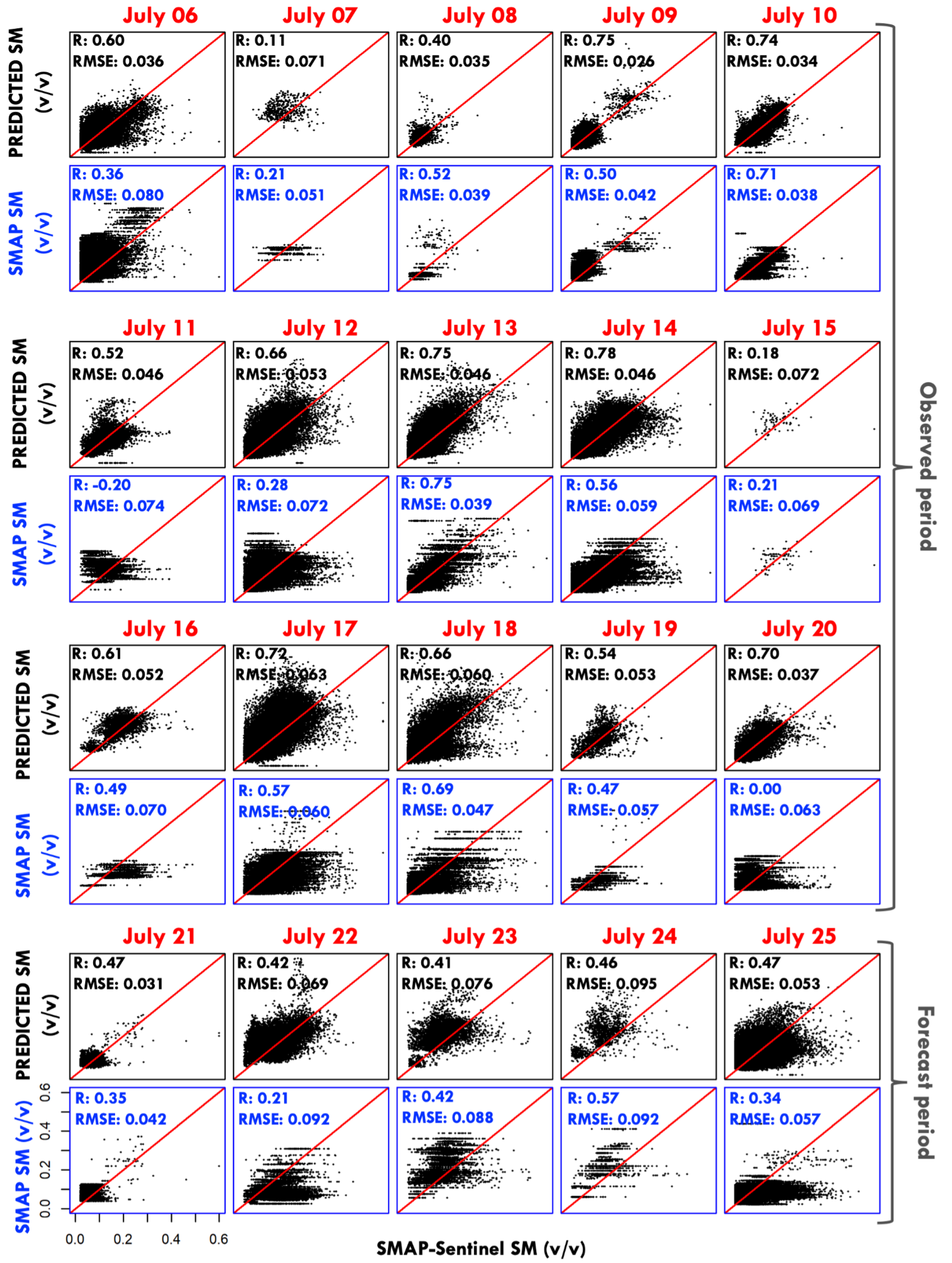

Figure 4.10: Comparison of soil moisture predictions and SMAP soil moisture with the observed SMAP/Sentinel-1 soil moisture at 3 km scale. For the majority of the days, the predicted soil moisture using the fusion approach outperforms the original base SMAP product (even for the forecast period). The red line denotes the 1:1 line.

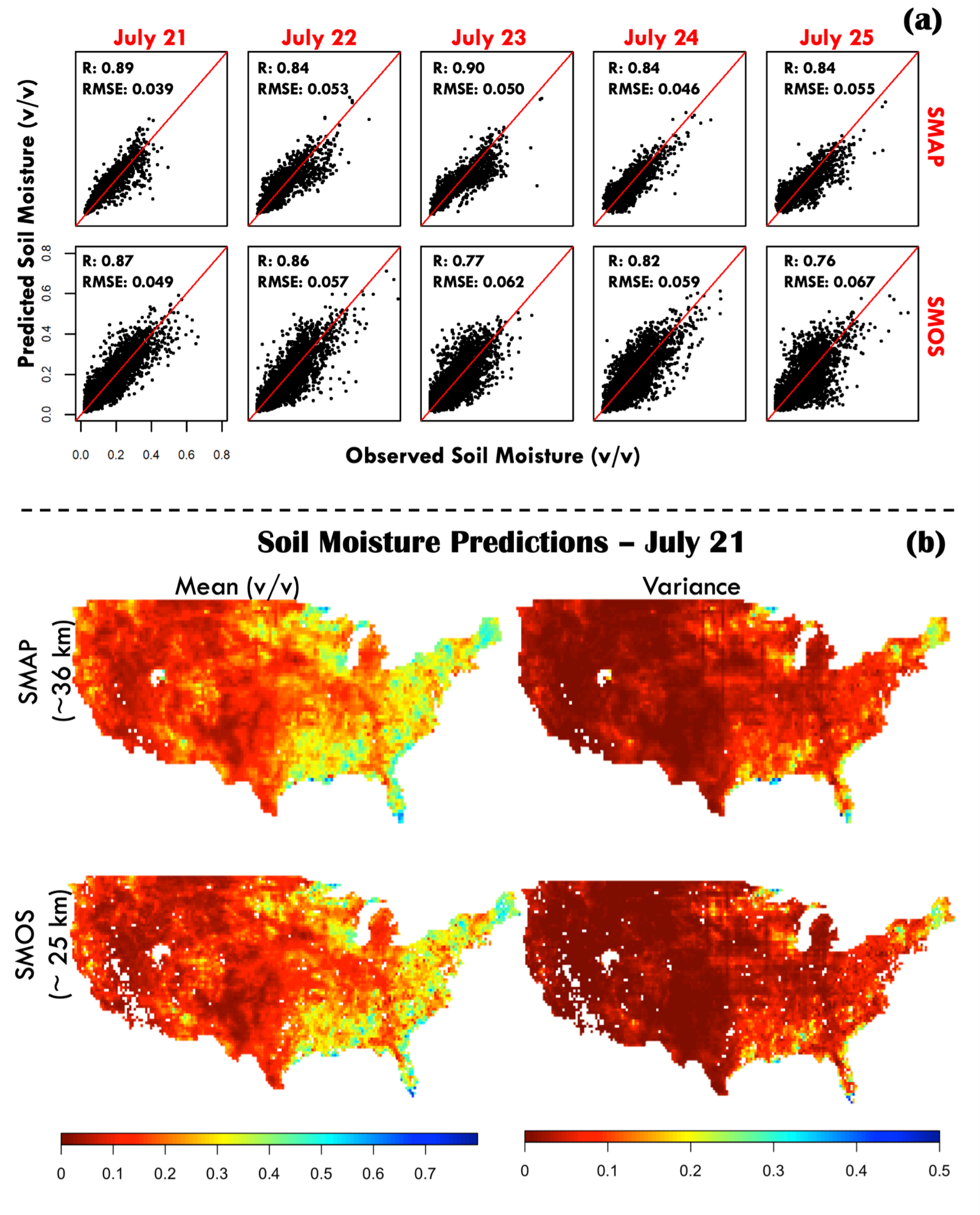

Figure 4.11: . (a) Comparison of soil moisture predictions and SMAP and SMOS observed soil moisture for July 21-25, 2017. The red line denotes the 1:1 line. (b) Soil moisture predictions across Contiguous US along with the prediction variance. Predictions are unavailable for certain regions due to absence of covariate data.

mean and covariance, the ordering will also be covariate-driven. We applied Vecchia-multiscale to simulated data and real SM observations; further application to diverse (spatio-temporally correlated) environmental variables will vet the widespread utility of the algorithm. An advantage of the proposed approach is that it is not a "black-box" and its components can be readily modified based on the underlying physical variable and expert-knowledge. Note that this algorithm can only be applied under a Gaussian Process assumption. In cases where such an assumption is untenable, recent research indicates that the approximation can be further extended using Generalized Gaussian Processes [174].

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.

# 5. A DYNAMIC ENHANCED SPATIAL AND TEMPORAL ADAPTIVE REFLECTANCE FUSION MODEL FOR EVAPOTRANSPIRATION

## 5.1 Synopsis

There has been an increased interest in combining data from satellites which provide observations at a fine spatial resolution but have extended revisit times with observations from satellites which provide data at a coarse spatial resolution albeit at a high temporal frequency. Currently the Spatial and Temporal Adaptive Reflectance Fusion Model (STARFM) and its subsequent extension for heterogenous landscapes, the Enhanced STARFM (ESTARFM), are the most widely used techniques to fuse such satellite data. These algorithms however, do not take the measurement errors of the satellites into account while combining data. They are also unsupervised techniques and thus their application consists of ad-hoc choices of important parameters that govern the overall accuracy of the algorithm. Finally, these algorithms do not have a framework to include ancillary data or theory-driven understanding of the underlying environmental processes. To account for the above limitations, we propose a Dynamic ESTARFM model which utilizes the original ESTARFM algorithm in a stochastic state-space modeling framework. The proposed model is applied to combine daily Evapotranspiration data from satellites ECOSTRESS and MODIS for agricultural sites in the lower Brazos Basin, Texas and we find that the proposed approach outperforms the original ESTARFM when compared to daily Eddy-Covariance data. The proposed scheme is general and can be utilized to combine other environmental variables.

## 5.2 Introduction

Six decades into the launch of satellites for observing Earth-Science processes, few global remote sensing platforms can retrieve data at a fine spatial resolution while providing frequent data coverage temporally. Multiple satellites have been launched however, which either observe the Earth frequently at a coarse spatial resolution or have long revisit time periods retrieving data at fine spatial resolution. Some examples of the former are Moderate Resolution Imaging Spectroradiometer (MODIS) [175], Soil Moisture Active Passive (SMAP) [157] and Soil Moisture and Ocean Salinity (SMOS) [156] while satellites comprising the latter include LANDSAT [176], ECOsystem Spaceborne Thermal Radiometer Experiment on Space Station (ECOSTRESS) [177] and SENTINEL-1 [178]. Naturally, interest in combining these disparate satellite platforms to provide a seamless coverage of remote sensing retrievals across space and time continues. The process of combining these disparate platforms such that the fused product provides better spatio-temporal information than provided by any individual platform is called data fusion [28].

There are three broad approaches [179] for fusing remote sensing data: 1) *weighted-function* approach, 2) *unmixing* approach, and 3) *dictionary-pair learning* approach. The *weighted-function* based methods [180, 181, 182] predict a fine resolution pixel using a weighted sum of predictions computed from the surrounding "similar" pixels. The *unmixing* based approaches [183, 184, 185] classify the fine resolution images into various classes or "end-members" (obtained from various classification techniques) and computes the fractions of each of these end-members in the corresponding coarse resolution image at an observed time $t_p$. Using a moving window algorithm, the coarse resolution image on the prediction date $t_q$ is then unmixed (using the classification at $t_p$) to provide predictions at the fine scale. The *dictionary-pair learning* methods [186, 187] use sparse representation to establish non-linear similarities between the coarse and fine resolution images to capture land-cover and phenological changes.

All the above data fusion algorithms combine a coarse spatial resolution (fine temporal resolution) platform—hereafter called platform $C$—with a fine spatial resolution (and coarse temporal resolution) platform—hereafter called platform $F$. The data fusion algorithms require concurrent images from the $F$ and $C$ platforms for one or more training days. They further require an image from the $C$ platform on the prediction day to generate a synthetic fine resolution image corresponding to the $F$ platform. In this work we focus on the *weighted-function* approach to data fusion.

The Spatial and Temporal Adaptive Reflectance Fusion Model (STARFM) [180] represents the first landmark data fusion study based on the *weighted-function* approach and is arguably the most widely used among all data fusion approaches. The STARFM is an unsupervised algorithm which requires a concurrent pair of images from an observed day (say day $t_p$) from $C$ and $F$ platforms as well as an image from $C$ platform on the prediction day (say day $t_q$) to generate the fine resolution image at day $t_q$. To predict a particular fine resolution pixel inside the study region at day $t_q$, the STARFM proceeds by selecting a subset of "similar" pixels within a moving window centered on the particular pixel. Then the STARFM computes individual predictions for each similar pixel, by adding the corresponding fine resolution values (from platform $F$) at day $t_p$ to the temporal difference of the overlapping coarse values (from platform $C$) between days $t_p$ and $t_q$. Each of these similar pixels is then assigned a weight based on: 1) the homogeneity of the overlapping $C$ pixel at date $t_p$, 2) the temporal change in platform $C$ between time $t_p$ and $t_q$, and 3) the spatial distance of pixels from the central pixel under consideration. The final prediction of the central pixel is then the weighted sum of of all these individual predictions.

The dependence of the STARFM algorithm on the existence of homogenous or "pure" coarse scale pixels limits its utility in heterogenous landscapes such as small agriculture fields. To improve upon this limitation, the Enhanced STARFM (ESTARFM) was proposed by [182] by using the spectral unmixing theory together with the original STARFM approach. The ESTARFM approach, however, requires two days ($t_o$ and $t_p$) for which concurrent images from $C$ and $F$ platforms are available (as opposed to just one pair for STARFM) and one coarse resolution image from platform $C$ at the prediction day $t_q$ to compute fine scale predictions at day $t_q$. The details of the ESTARFM algorithm are given in Section 5.4.1.

The STARFM and ESTARFM algorithms were developed to fuse low-level (Level-1) products such as surface reflectance which generally exhibit less spatio-temporal variability than high level environmental products (Level-2 or Level-3) such as evapotranspiration (ET) [188]. Moreover, high-level products suffer from additional errors due to the uncertainty induced by the retrieval algorithm applied to convert the low-level surface reflectance data (Level-1) into high-level environmental products. Often the retrieval algorithms of two different platforms retrieving the same environmental variable differs due to the physical differences in the sensors. For ET retrieval for instance, ECOSTRESS uses the PT-JPL algorithm [177] to generate daily ET while MODIS employs the MOD16 retrieval algorithm [189]. Another shortcoming of the STARFM/ESTARFM fusion approach is that since they are unsupervised algorithms, an end-user has to make ad-hoc choices when choosing important parameters such as the number of clusters used in clustering the data and the size of the moving window. This not only affects the accuracy of the algorithm if sub-optimal parameter values are chosen but also prevents the automatic processing and implementation of these methods for different study regions thus limiting their widespread applicability. The same has been acknowledged by [182].

ET is a critical variable in energy and water cycle from sub-field to global scales. It influences a variety of processes and activities such as agricultural water management [29, 30], drought estimation [31, 32], water and energy balance closures [33, 34], water rights studies [35], and atmospheric processes [36]. Remote sensing has emerged as an accurate and relatively inexpensive way to estimate ET over vast spatio-temporal domains and fusion of existing ET remote sensing platforms thus is critical. Recent research [190, 188] have proposed ET fusion methodologies combining the ALEXI/DisALEXI algorithm [191, 192] with the STARFM approach. In this approach, instead of directly fusing multi-platform ET obtained from different retrieval algorithms, Land Surface Temperature from Landsat (or ECOSTRESS) and MODIS (along with other meteorological and remote sensing data) are used in the ALEXI/DisALEXI algorithm to generate consistent ET maps from both platforms at the 10 km scale. The consistency helps to reduce uncertainty between the two platforms allowing the use of the STARFM algorithm to provide fused ET at a daily scale at high spatial resolution. Both ALEXI and DisALEXI models however, require a variety of remote sensing, regional modeled meteorological data ,and/or local radiosonde data and are thus dependent on the accuracy and availability of these products. The use of ALEXI/DisALEXI algorithm also prevents the use of the actual retrieval algorithms of the two platforms which might be more accurate (for certain study regions) for the specific sensor employed by the satellite. Using the ALEXI/DisALEXI algorithm along with the STARFM approach makes the processing computationally intensive. Also, this limits the application of the STARFM/ESTARFM approaches to other high-level products where an equivalent ALEXI/DisALEXI algorithm is not available.

Therefore in this work, our objective is to propose an extension to the STARFM/ESTARFM approach which allows the inclusion of parameterization of errors in the platforms as well as additional physical/empirical knowledge of the underlying variable, thus making it more applicable to fuse high-level products. The proposed algorithm also allows for the optimal selection of the parameters of the STARFM/ESTARFM algorithm such as the number of clusters, the size of the search window, etc. making the implementation of the algorithm automatic and increasing its applicability to a variety of study regions. We demonstrate the utility of the algorithm by fusing ET estimates from ECOSTRESS and MODIS for three agricultural sites in the Lower Brazos Basin in Texas, USA for the year 2020.

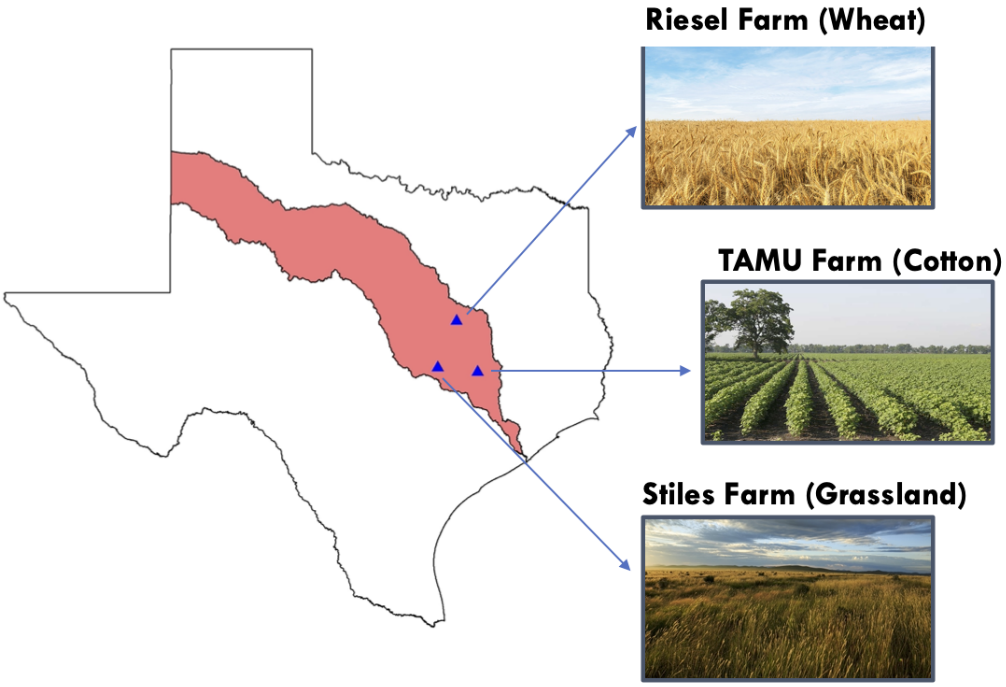

Figure 5.1: The proposed algorithm is validated for the year 2020 using daily Eddy-Covariance data on three sites across the Brazos Basin (in red), Texas having distinct land use and landcover.

## 5.3 Study Area and Data

We validate the proposed algorithm at three agricultural sites (Figure 5.1) lying in the lower Brazos Basin, Texas using data from Eddy-Covariance stations installed as part of the Texas Water Observatory (TWO). The first site (TfPr) is a cotton farm near Texas AM University, College Station, the second site (RfAa) is a wheat aspirational agricultural site located in Riesel and the third site (SfPr) is a grassland in Stiles. All three sites are classified as having Humid Subtropical climate (Cfa) according to the Köppen-Geiger climate classification [69]. The three sites have different agricultural crops and soil types and therefore serve as good testbeds for the validation of the algorithm. We apply the proposed algorithm for the year 2020 for a $2.5\,km \times 2.5\,km$ area centered on the Eddy-covariance location for each day.

Fine resolution daily ET at 70m is available from ECOSTRESS using the PT-JPL algorithm [177] and is provided as a Level-3 (L3) latent energy flux (LE) product (ECO3ETPTJPL v001). THE ECOSTRESS comprises a thermal radiometer aboard the International Space Station (ISS) and has an irregular orbit (unlike polar or geostationary satellites) with an overpass frequency of 1-5 days, with higher latitudes getting sampled more frequently compared to lower latitudes. ECOSTRESS derives the daytime average ET from the instantaneous ET retrieval by generating a sinusoidal curve mimicking the diurnal radiation intensity from sunrise to sunset. The exact sinusoidal curve depends on the date and latitude with additional refinements governed by different atmospheric and surface characteristics. Since the daily ET product resulting from the PT-JPL algorithm only gives daytime ET, it does not utilize night-time ECOSTRESS retrievals. Moreover, cloud cover can further increase the time interval between successful daily ET retrievals from ECOSTRESS.

Eight-day averaged ET is also available from retrievals from MODIS sensor aboard the Aqua and Terra platforms using the MOD16 algorithm [189]. The MOD16 algorithm is motivated by the Penman-Monteith equation using daily meteorological reanalysis data and MODIS vegetation data to derive 8-day average ET at 500 m resolution for the entire globe. The MODIS sensor is aboard the Terra and Aqua platforms having a revisit time of 1-2 days at approximately 10:30 AM local time. The MODIS 8-day average ET has few missing time periods for our analyzed sites. We choose ECOSTRESS and MODIS as the two ET platforms to be fused as they do not require any local information for estimating ET and provide ET estimates for the entire globe thus potentially increasing the future global utility of the proposed data fusion algorithm.

## 5.4 Methodology

### 5.4.1 Summary of ESTARFM algorithm

Predictions using the STARFM approach [180] are prone to errors in highly heterogeneous regions due to the absence of homogenous or pure pixels. To solve for this limitation, [182] proposed an extension to the STARFM approach called the ESTARFM such that the data fusion scheme could be applied to highly heterogenous regions. Since the spatial distribution of many environmental variables is heterogenous under complex surface and atmospheric variabilities, the ESTARFM algorithm was an important extension to the STARFM. In this Section, we briefly summarize data fusion between a fine scale platform $F$ and coarse scale platform $C$ using the ESTARFM algorithm.

To fuse data platforms and compute predictions at fine resolution for a time $t_q$, the ESTARFM requires 2 pairs of fine resolution and coarse resolution data (from platforms $F$ and $C$ respectively) at time periods near to $t_q$ (say $t_o$ and $t_p$) and one coarse resolution image at time $t_q$. For a given study region, the ESTARFM approach is iteratively applied for each pixel lying inside the study region. For a pixel $s_i$ lying inside the study region, a search window of size $w \times w$ centered on the pixel $s_i$ is utilized to compute predictions for pixels “similar” to the pixel $s_i$ within the moving window. The fine resolution prediction for pixel $s_i$ at time $t_q$ based on the observed pairs at time $t_p$ (or $t_o$) is then given as:

$$F_{t_q}(s_i) = F_{t_p}(s_i) + \sum_{j=1}^{N_i} W_j(s_i) \times V_j(s_i) \times (C_{t_q}(s_i) - C_{t_p}(s_i)) \tag{5.1}$$

where $N_i$ refers to the number of similar pixels within the moving window which are considered similar to the central pixel $s_i$. The weight $W_j(s_i)$ refers to the weight assigned to each similar pixel while $V_j(s_i)$ refers to the conversion factor accounting for the bias between the the platforms $C$ and $F$ for each similar pixel.

#### *5.4.1.1 Calculating the number of similar pixels*

To calculate the number of similar pixels for a pixel $s_i$, the $j^{th}$ neighboring pixel is chosen as a similar pixel if it satisfies:

$$|F_{t_p}(s_i) - F_{t_p}(s_j)| < \sigma(F) \times 2/n_c \tag{5.2}$$

where $\sigma(F)$ is the standard deviation of the variable and $n_c$ is the assumed number of classes used for the clustering in equation 5.2. The standard deviation $\sigma(F)$ is approximated using the sample standard deviation of the fine scale observations at time $t_p$. From equation 5.2, higher the value of $n_c$ less the number of similar pixels chosen for a particular pixel $s_i$. In the ESTARFM approach we use both the dates $t_o$ and $t_p$ to determine the similar pixels for each date and the eventual set of similar pixels in equation 5.1 represent the common similar pixels for the two time periods. If there are no similar pixels for a particular central pixel $s_i$ then the weight for $s_i$ is set to 1 (and all others equal to zero).

#### *5.4.1.2 Weight calculation for ESTARFM*

The weight $W_j(s_i)$ for the $j^{th}$ similar pixel is calculated as:

$$W_j(s_i) = \frac{(1/D_j)}{\sum_{j=1}^{N_i}(1/D_j)} \tag{5.3}$$

The range of values for $W_j(s_i)$ varies between 0 and 1, with the total weight of $N_i$ similar pixels equal to 1. Here $D_j = (1 - R_j) \times d_j$. The factor $R_j$ is a measure of the correlation between the values of fine and coarse scale pixels. For a signal with multiple spectral bands, $R_j$ is taken as the correlation coefficient between the fine and coarse resolution pixel values at time $t_o$ and $t_p$. For univariate variables, the value of $R_j$ is given as:

$$R_j = 1 - 0.5(|\frac{F_{t_o}(s_j) - C_{t_o}(s_j)}{F_{t_o}(s_j) + C_{t_o}(s_j)}| + |\frac{F_{t_p}(s_j) - C_{t_p}(s_j)}{F_{t_p}(s_j) + C_{t_p}(s_j)}|) \tag{5.4}$$

The factor $d_j$ determines the geographic distance between the $j^{th}$ similar pixel and the $i^{th}$ (central) pixel. The ESTARFM algorithm gives higher weights to pixels that are closer in space to $s_i$ and is given as:

$$d_j = 1 + \frac{\sqrt{||s_i - s_j||^2}}{(w/2)}, \tag{5.5}$$

where $w$ is the size of the moving window.

#### *5.4.1.3 Calculation of the conversion coefficient*

The conversion coefficient $V_j(s_i)$, which accounts for the bias existing between coarse and fine resolution pixels, is calculated for the $j^{th}$ similar pixel (separately for each coarse pixel) by linearly regressing the fine resolution values with the coarse resolution values for all the similar pixels. The conversion coefficient $V_j(s_i)$ is then the slope parameter of this regression. If the regression model cannot be built with statistical significance ($p < 0.05$), then the value of $V_j(s_i)$ is assumed to be 1 (or no bias is assumed between fine and coarse resolution pixels).

The ESTARFM therefore combines data from fine and coarse resolutions and provide fine resolution predictions at frequent time intervals. However, the ESTARFM approach does not account for uncertainty in observations which might be especially present in high level products (Level-2 and Level-3) due to imperfect retrieval algorithms. Also, the ESTARFM approach is unsupervised and therefore has no mechanism to determine the correct number of clusters $n_c$ and the size of the moving window $w$. The number of clusters $n_c$ is especially important as it governs the number of similar pixels for a central pixel within a moving window and incorrectly determining the number of clusters leading to erroneous predictions. Also, as mentioned earlier ESTARFM cannot incorporate ancillary information about the underlying process such as additional data and domain knowledge. In the next section, we propose a dynamic state-space approach to the ESTARFM which accounts for all the above-mentioned shortcomings while preserving the salient features of the ESTARFM algorithm.

### 5.4.2 Dynamic ESTARFM

To formulate the dynamic state-space model [193, 119] for ESTARFM, we begin by looking at the ESTARFM model in equation 5.1 as an evolution model which evolves the state of the variable from $t_k$ to $t_{k+1}$, where $t_k$ and $t_{k+1}$ are not necessarily consecutive time periods. Since our objective is to account for errors in the remote sensing retrievals, for a pixel $s_i$, we differentiate between the observed value $F_{t_k}(s_i)$ and the underlying latent value of the variable free from parameterized errors, say $x_{t_k}(s_i)$. Substituting $F_{t_k}(s_i)$ by $x_{t_k}(s_i)$ in equation 5.1, we write:

$$x_{t_{k+1}}(s_i) = x_{t_k}(s_i) + E_{t_k}(s_i), \tag{5.6}$$

where $E_{t_k}(s_i) = \sum_{j=1}^{N_i} W_j(s_i) \times V_j(s_i) \times (C_{t_{k+1}}(s_i) - C_{t_k}(s_i))$. Equation 5.6, however, might not fully represent the how the variable evolves in time and we thus write equation 5.6 as:

$$\begin{aligned} x_{t_{k+1}}(s_i) &= x_{t_k}(s_i) + \alpha_k(s_i) \times E_{t_k}(s_i) + \beta_k(s_i) + \eta(s_i) \\ &= x_{t_k}(s_i) + E^*_{t_k}(s_i) + \eta(s_i), \end{aligned} \tag{5.7}$$

where $\alpha_k(s_i)$ and $\beta_k(s_i)$ are corrections to the evolution equation and can be either fixed through expert knowledge, physical understanding of the underlying process or be estimated from the data. For spatially correlated variables, the residual spatial structure not captured by $E^*_{t_k}(s_i)$ can be modeled through $\eta(s_i)$.

We relate the actual observed value $F_{t_k}(s_i)$ with $x_{t_k}(s_i)$ by assuming systematic and random errors in the platform $F$. Specifically we write:

$$F_{t_k}(s_i) = x_{t_k}(s_i) + \Delta_k(s_i) + \tau_k(s_i) \tag{5.8}$$

Suppose we have a study region, consisting of $n_f$ fine scale pixels $s = \{s_1, s_2, ..., s_{n_f}\}$ which are fully or partly observed by the fine scale platform for a day $t_k$ where $k = 1, 2, 3, \ldots, T$. Note that $t_k$ and $t_{k+1}$ need not be consecutive time intervals. For example, $t_1$ can be equal to Day 1, $t_2$ equal to Day 10, $t_3$ equal to Day 32 and so on. For any day $t_k$, we can write the evolution equation as:

$$\boldsymbol{x}_{\boldsymbol{t_{k+1}}}(\boldsymbol{s}) = \boldsymbol{x}_{\boldsymbol{t_k}}(\boldsymbol{s}) + \boldsymbol{E}^*_{\boldsymbol{t_k}}(\boldsymbol{s}) + \eta(\boldsymbol{s}) \tag{5.9}$$

where $\eta(s) \sim \mathcal{N}_{n_f}(0, Q)$, and $Q$ is a $n_f \times n_f$ covariance matrix denoting the residual spatial covariance structure of the process. For any day $t_k$, there might be some missing pixels in the study region due to errors in retrievals caused due to surface or atmospheric factors such as cloud cover. If the number of pixels observed by the platform on day $t_k$ be $m_k$ ($m_k \leq n_f$), then we write the observation equation as:

$$\boldsymbol{F}_{\boldsymbol{t_k}} = \boldsymbol{A}_k \boldsymbol{x}_{t_k}(\boldsymbol{s}) + \boldsymbol{\Delta_k}(\boldsymbol{s}) + \tau_k(\boldsymbol{s}) \tag{5.10}$$

where $\tau_k \sim \mathcal{N}_{m_k}(0, R_k)$ and $R_k$ represents an $m_k \times m_k$ error covariance matrix. For any day $t_k$, $\boldsymbol{A}_k$ is an $m_k \times n$ dimension matrix. The matrix $\boldsymbol{A}_k$ is an identity matrix for $m_k = n$. For $m_k < n$, the $m_k$ elements of $\boldsymbol{A}_k$ are equal to 1 corresponding to the observed pixels on day $t_k$ and are zero otherwise. The evolution model denoted by equation $5.9$ and the observation model in equation $5.10$ comprise the dynamic state-space ESTARFM model.

#### *5.4.2.1 Filtering and Parameter Estimation*

For the state-space model defined by equations $5.9$ and $5.10$, parameter inference and predictions are done using filtering. Specifically, the state-space model is a linear Gaussian model and we employ the Kalman Filter [194] to estimate parameters and compute predictions. The Kalman filter consists of two steps — 1) forecast step and 2) update step.

Let $[V|W]$ denote the conditional probability distribution of $V$ given $W$. Assuming that the initial conditional distribution of $[x_{t_0}(\boldsymbol{s})|F_{t_0}]$ is $\mathcal{N}_{n_f}(\hat{\mu}_0, \hat{P}_0)$, for $t_k = t_1, t_2, ...., t_T$ the Kalman filter iteratively applies the forecast and update step as follows:

$$\begin{aligned}
Forecast\,Step :& x_{t_k}(\boldsymbol{s})|F_{1:t_k} \sim \mathcal{N}_{n_f}(\hat{\mu}_t, \hat{P}_t) \\
& \hat{\mu_k} = \tilde{\mu_k} + KG_k(z_{t_k} - A_k\tilde{\mu}_k - \Delta_k), \\
& \hat{P_k} = \tilde{P_k} - KG_kA_k\tilde{P}_k \\
Update\,Step :& x_{t_k}(\boldsymbol{s})|F_{1:t_{k-1}} \sim \mathcal{N}_{n_f}(\tilde{\mu}_k, \tilde{P_k}) \\
& \tilde{\mu}_k = \hat{\mu}_{k-1} + \boldsymbol{E^*_{t_k}}(\boldsymbol{s}), \\
& \tilde{P_k} = \hat{P}_{k-1} + Q,
\end{aligned} \tag{5.11}$$

where $KG_k = \tilde{P}_kA_k'(A_k\tilde{\Sigma}_kA_k' + R_k)^{-1}$ is the Kalman gain matrix of size $n \times m_k$.

For $k = 1, 2..., T$, let the parameters $n_c$, $w$, $Q$, $R_k$, $\Delta_k$, $\alpha_k$ and $\beta_k$ be denoted by $\Theta$. The parameter vector $\Theta$ is estimated by maximizing the likelihood $L(\Theta)$ or equivalently minimizing the negative log likelihood $-ln(L(\Theta))$, computed using innovations $\epsilon_1, \epsilon_2, ..., \epsilon_T$ [193] where $\epsilon_k = z_{t_k} - A_k\tilde{\mu}_k - \Delta_k$ such that:

$$-ln(L(\Theta)) = \frac{1}{2}\sum_{k=1}^{T} ln|\Sigma_k(\Theta)| + \frac{1}{2}\sum_{k=1}^{T} \epsilon_k(\Theta)'\Sigma_k(\Theta)^{-1}\epsilon_k(\Theta), \tag{5.12}$$

where $\Sigma_k = A_k\tilde{P}_kA_k' + R_k$. The function in equation 5.12 is highly non-linear and complex. To estimate the parameters in $\Theta$, we use a global optimization algorithm to ensure that we arrive at the parameter vector $\Theta_{est}$ which globally minimizes equation 5.12. We use a a generalized and improved form of simulated annealing called GenSA [80] to find the parameter estimates that globally maximizes the likelihood in equation 5.12.

## 5.5 Results and Discussion

### 5.5.1 Implementation of Dynamic ESTARFM to Evapotranspiration

#### *5.5.1.1 Data Processing*

The Eddy-Covariance towers installed at the three TWO sites (Figure 5.1) give ET at half-hour intervals. To convert the half-hour ET to daily average daytime ET, we take the mean of the half-hourly ET values during the daytime. Eddy-covariance data are prone to errors during time periods close to sunrise and sunset and therefore we only use values of Eddy-Covariance ET from 10 AM to 5 PM. We apply the energy-balance closure correction following the data-processing pipeline developed by the FLUXNET community [195]. Since ECOSTRESS only uses daytime instantaneous retrievals for computing daily ET and due to frequent cloud cover during the year 2020, daily ET data for the three sites is extremely sparse for the entire year. For each day, we also filter out the ECOSTRESS pixels which have a Band-5 emissivity (sensitive to clouds) less than 0.95 to

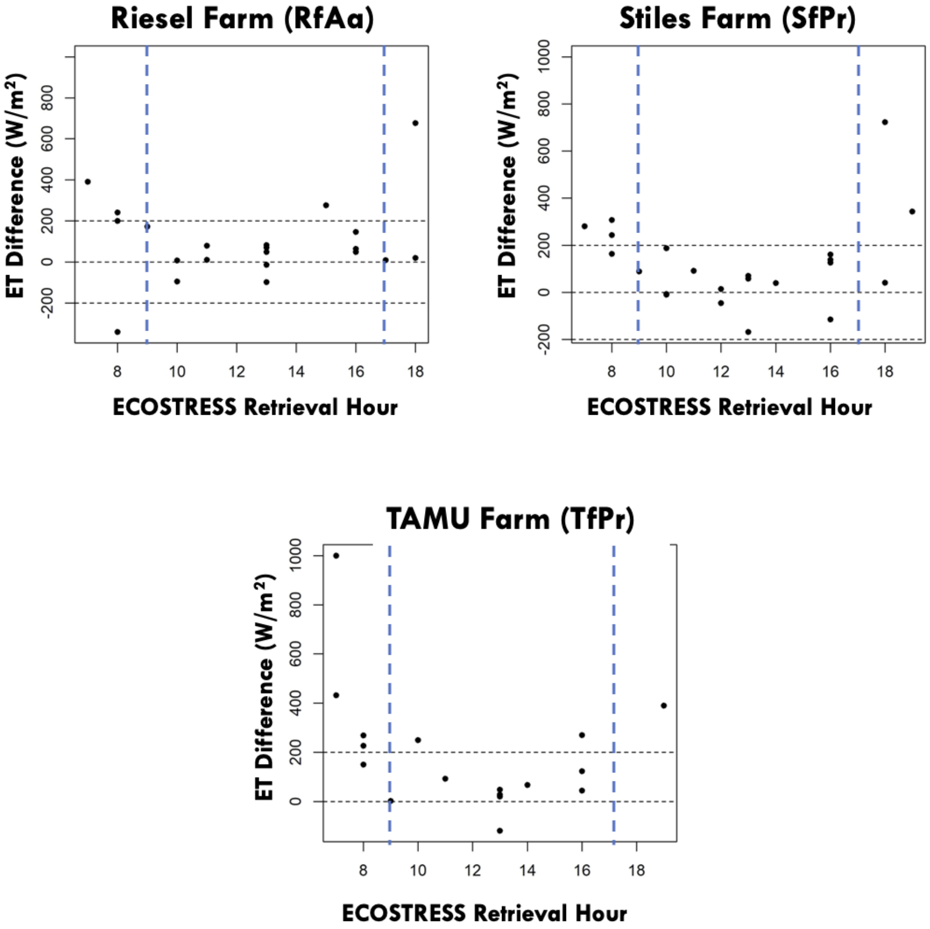

Figure 5.2: Analysis of the ECOSTRESS bias when compared with the observed daily Eddy Covariance ET as a function of the ECOSTRESS retrieval hour for the year 2019 reveals that ECOSTRESS exhibits the lowest bias between hours 9 AM to 5 PM (blue vertical lines).

ensure minimal cloud contamination. [190] found that even for ECOSTRESS daytime retrievals, limiting the retrieval times between 9 AM to 5 PM yielded good quality ECOSTRESS retrievals. For the three TWO sites, we analyzed the days in 2019, for which daily ECOSTRESS data was available, and compared them with daily Eddy-Covariance ET. Figure 5.2 gives the difference between the ECOSTRESS and Eddy-Covariance ET for 2019 summarized by the retrieval hours of ECOSTRESS. We also find that the times between 9 AM to 5 PM result in lowest bias between the two platforms and therefore for the year 2020, select only days where the ECOSTRESS retrievals occurred between 9 AM to 5 PM. This resulted in 10 days for RfAa (Figure 5.3), 11 days for SfPr (Figure 5.4) and 10 days for TfPr (Figure 5.5) for the year 2020.

The ESTARFM approach requires continuous daily MODIS data as an input. Therefore, we smooth the MODIS 8-day average ET product using locally estimated scatterplot smoothing (LOESS) to smooth any missing values and remove any effect of outliers. As an illustration, the smoothed ET plots for the pixel containing the Eddy-Covariance stations are given in Figure 5.6. The temporally smoothened plots are computed separately for each MODIS pixel.

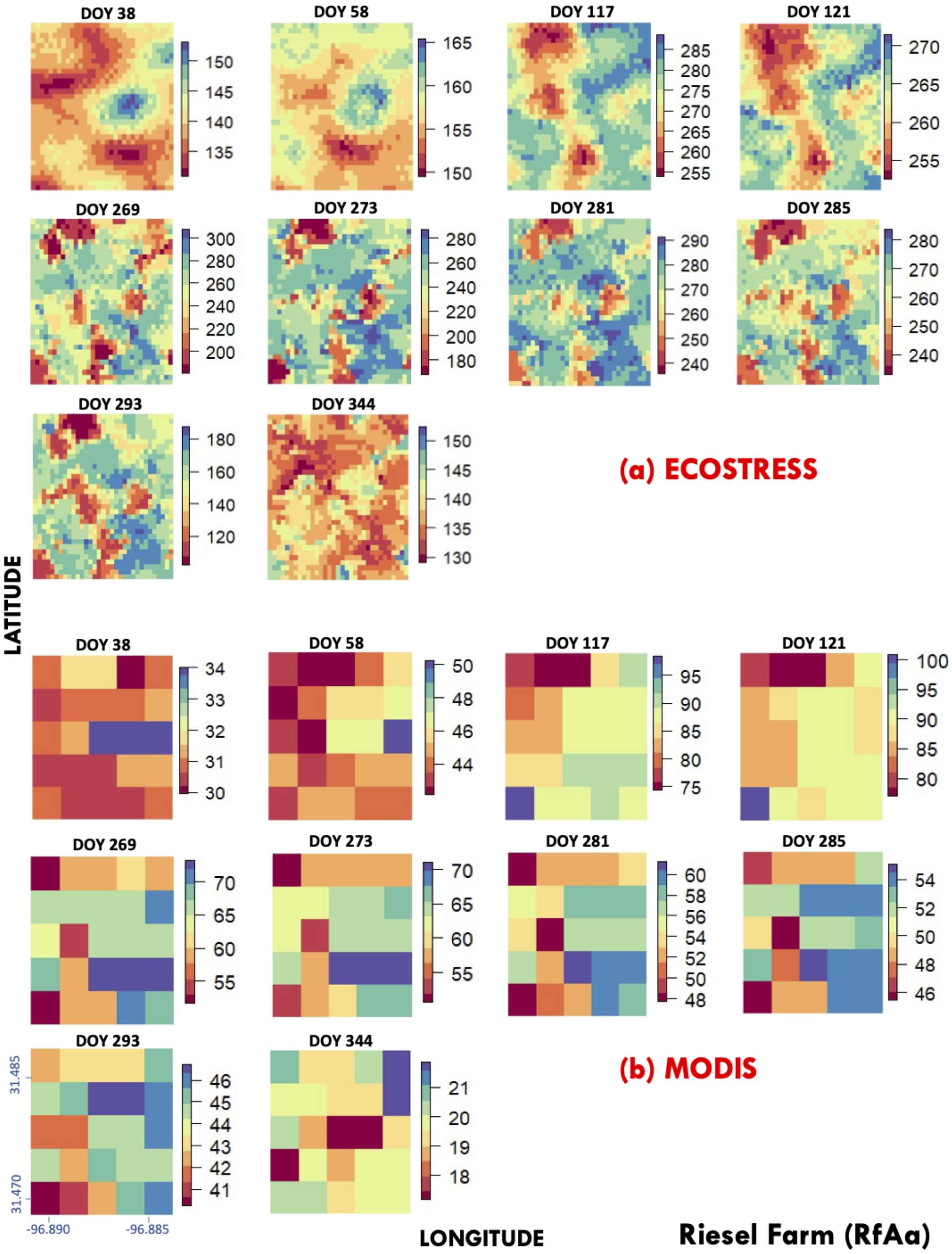

Figure 5.3: (a) The figure depicts the days of year for which ECOSTRESS ET ($W/m^2$) data is available for the Riesel Farm for the year 2020. The dimensions of the study region are $2.5/, km \times 2.5/, km$ centered on the Eddy-Covariance station. (b) MODIS ET ($W/m^2$) retrievals corresponding to the ECOSTRESS days for the study region.

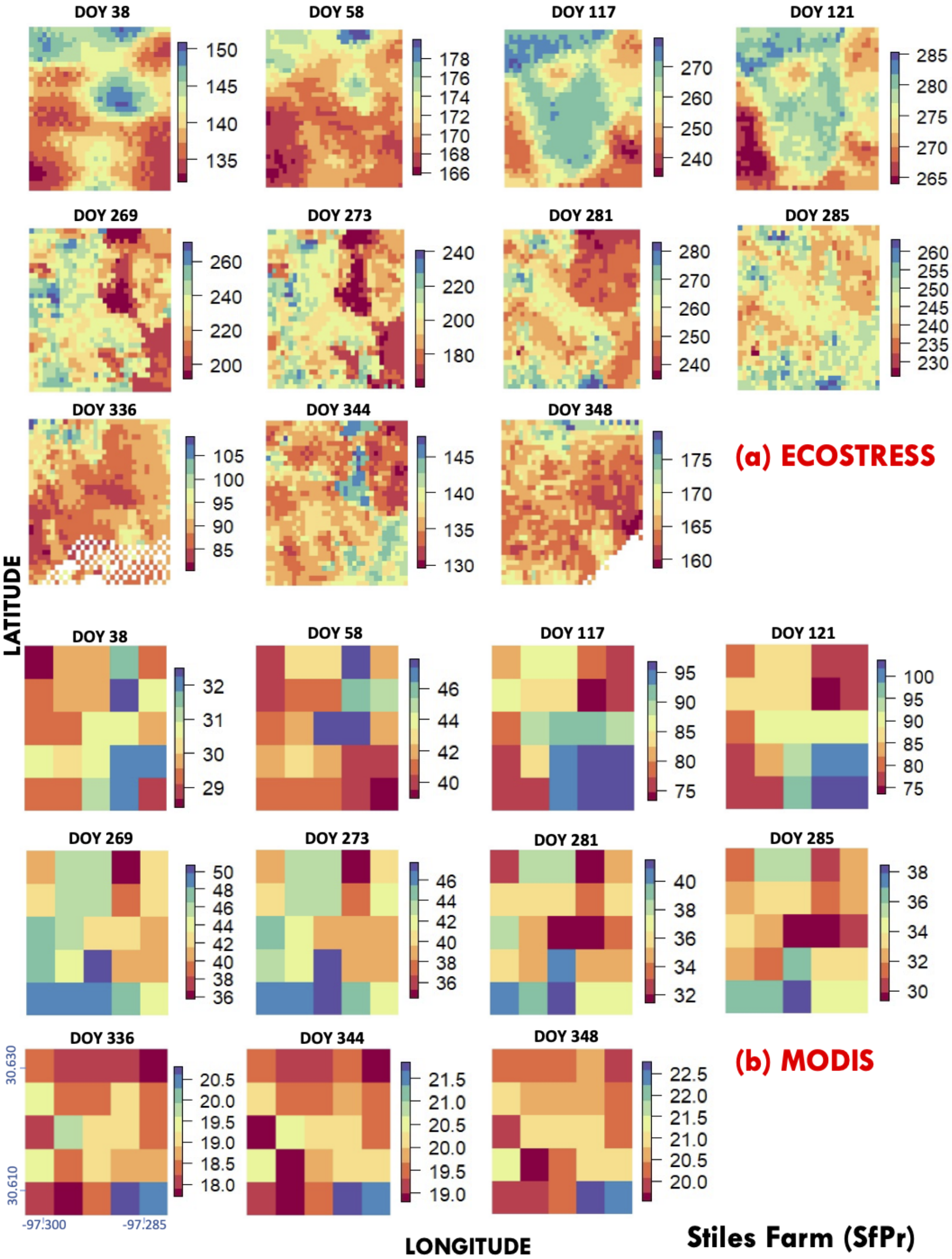

Figure 5.4: (a) The figure depicts the days of year for which ECOSTRESS ET ($W/m^2$) data is available for the Stiles Farm for the year 2020. The dimensions of the study region are $2.5/,km \times 2.5/,km$ centered on the Eddy-Covariance station. (b) MODIS ET ($W/m^2$) retrievals corresponding to the ECOSTRESS days for the study region.

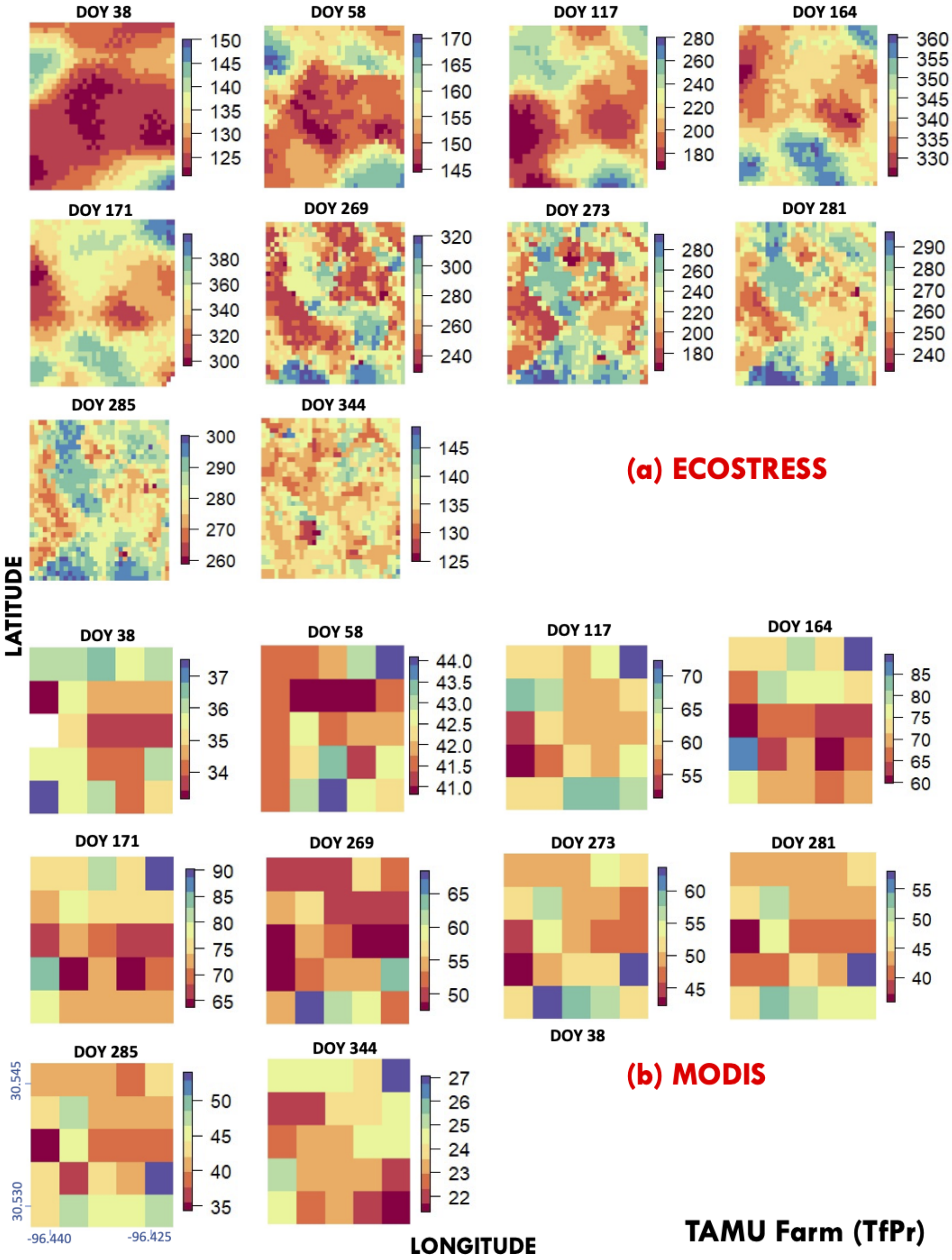

Figure 5.5: (a) The figure depicts the days of year for which ECOSTRESS ET ($W/m^2$) data is available for the TAMU Farm for the year 2020. The dimensions of the study region are $2.5/, km \times 2.5/, km$ centered on the Eddy-Covariance station. (b) MODIS ET ($W/m^2$) retrievals corresponding to the ECOSTRESS days for the study region.

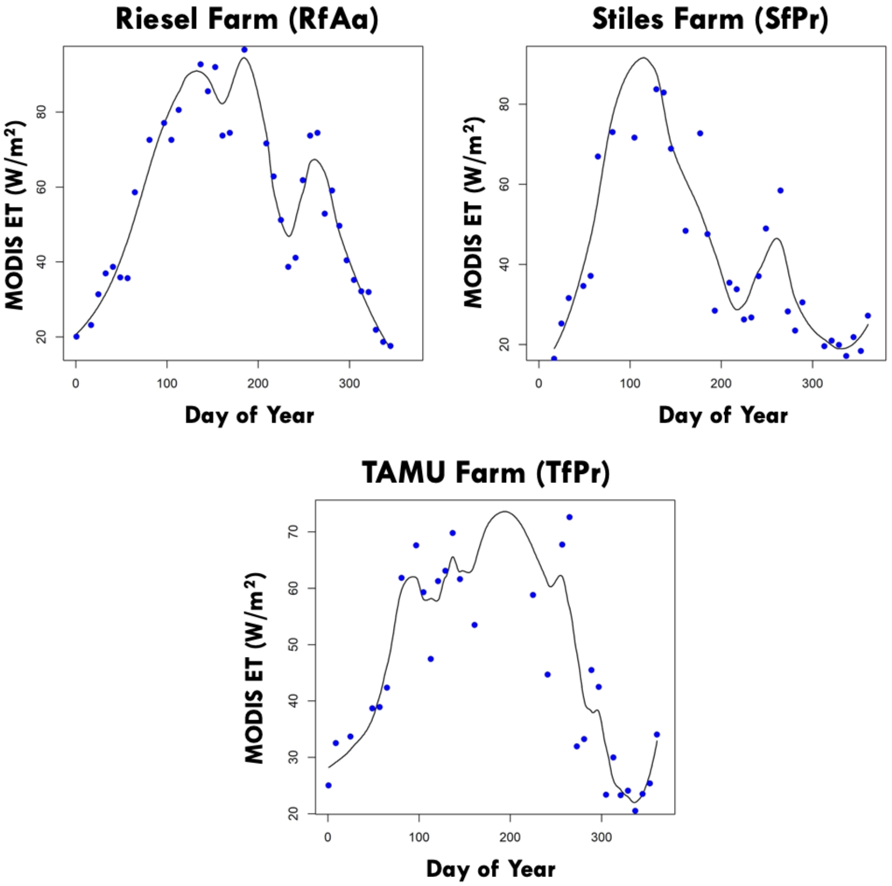

Figure 5.6: Smoothed ET plots for the eight-day averaged MODIS data for each site for 2020.

### *5.5.1.2 Parameterization of the Dynamic ESTARFM for TWO stations*

To implement the ESTARFM algorithm for the three TWO sites, we need to employ parsimonious representation of the parameters in $\Theta$ since we have only 10-11 observed days from ECOSTRESS for the year 2020. We do the analysis and estimate the parameters separately for each TWO site. For each site we fix the size of the moving window to $1.25\,km \times 1.25\,km$ and assume that the number of clusters $n_c$ remain constant for the entire period. For any day $t_k$, we assume that the parameters $Q$, $R_k$, $\alpha_k$ and $\beta_k$ remain constant spatially in the $2.5\,km \times 2.5\,km$ study region. Temporally, we assume that the parameter values for $\alpha_k$ and $\beta_k$ remain constant for the growing crop period (when ET is increasing temporally) and assume different constant values for

Table 5.1: Estimated parameters for the Dynamic ESTARFM model for the three agricultural sites for the year 2020.

| Site | $SD_e$ | Range | Smoothness | $SD_o$ | $\alpha_k^r$ | $\beta_k^r$ | $\alpha_k^f$ | $\beta_k^f$ | clusters |
|---|---|---|---|---|---|---|---|---|---|
| Riesel Farm | 9.95 | 0.014 | 0.49 | 8.45 | 0.47 | 10.5 | 0.63 | 5.47 | 3 |
| Stiles Farm | 9.94 | 0.015 | 0.53 | 4.15 | 0.32 | -1.15 | 0.76 | -6.21 | 4 |
| TAMU Farm | 9.88 | 0.009 | 0.53 | 8.04 | 0.29 | -2.58 | 0.54 | -7.71 | 4 |

$\alpha_k$ and $\beta_k$ during the post ET peak period (when ET decreases temporally). The exact date where the peak ET occurs is estimated from the Eddy-Covariance data from the average of the peak ET period as observed from ECOSTRESS and MODIS for the years 2019 and 2020. For the spatial covariance matrix $Q$, we assume a temporally independent spatial process parameterized with a Matern covariance function with variance $\sigma^2$, range $\lambda$ and smoothness $\nu$. For the observed bias $\Delta_k$ existing between the ECOSTRESS and Eddy-Covariance platform, we assume a constant bias and estimate it from the observed differences in ECOSTRESS and Eddy-Covariance ET during the year 2019. It is therefore considered fixed when applying the Kalman Filter to the ET data for the year 2020. For the measurement error $\tau_k$ (equation 5.10), we assume independent and identically distributed measurement error with variance $\zeta^2$ which remains constant for the entire year, and therefore $R_k$ is parameterized as a diagonal matrix with all diagonal elements equal to $\zeta^2$. The estimated parameters for the three study sites are given in Table 5.1.

#### *5.5.1.3 Predictions*

We compute ET predictions using the original ESTARFM and the proposed dynamic ESTARFM and compare them to the daily observed ET from Eddy Covariance towers for the year 2020. Since, ECOSTRESS data is extremely sparse for the year 2020 (10-11 days), our objective is to estimate the mean trend of ET throughout the year using the sparsely observed ECOSTRESS data and the low resolution but (relatively) temporally dense MODIS data. We therefore find the mean trend of the observed ET data from Eddy-Covariance towers and use it for comparison with the computed predictions. The observed Eddy-Covariance data along with the mean trend, the observed ECOSTRESS values, the ESTARFM predictions and the dynamic ESTARFM predictions

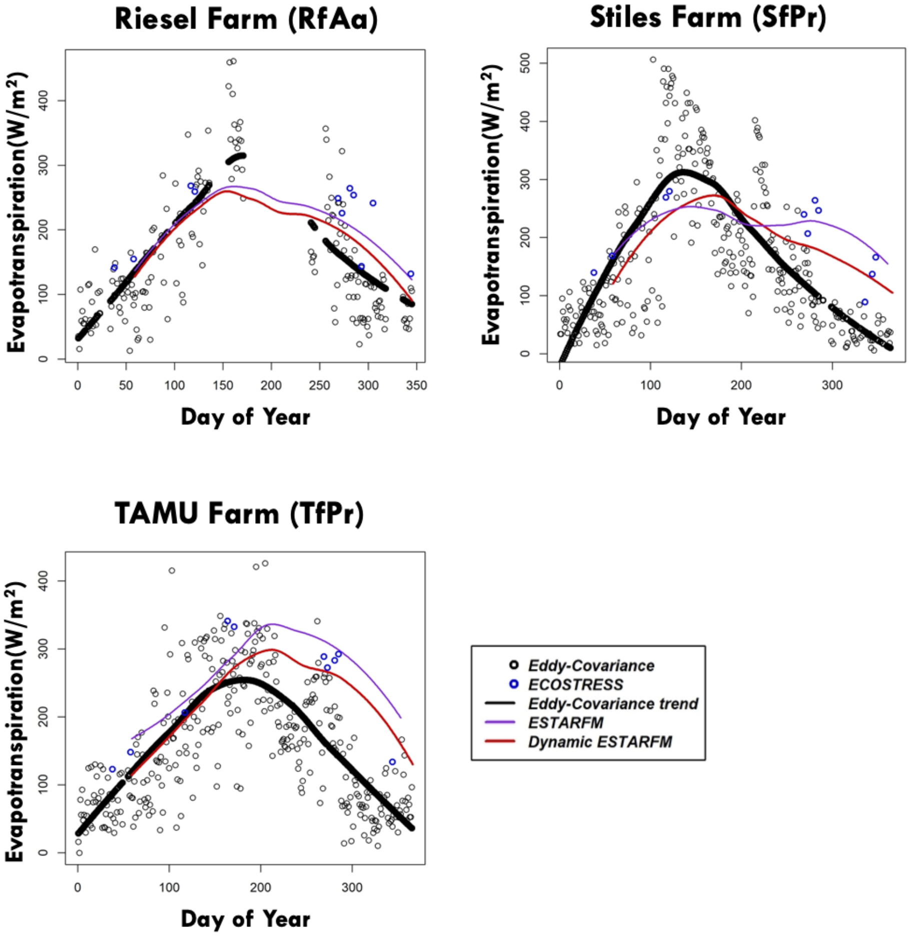

Figure 5.7: Comparison of ET predictions from the proposed Dynamic ESTARFM compared with the original ESTARFM approach. The Dynamic ESTARFM exhibits less errors when compared with the observed Eddy-Covariance Eton all the three sites for the year 2020.

are given in Figure 5.7. For the RfAa site, the dynamic ESTARFM ($RMSE = 100.78\,W/m^2$ ) performs marginally better than the original ESTARFM ($RMSE = 105.22\,W/m^2$). For the SfPr and TfPr sites, the dynamic ESTARFM does relatively better with RMSE values of $66.73\,W/m^2$ and $88.79\,W/m^2$ respectively compared with the original ESTARFM which has an RMSE of $87.9\,W/m^2$ and $117.54\,W/m^2$ for SfPr and TfPr respectively.

The dynamic ESTARFM performs better than the original ESTARFM approach for several reasons. First, the ESTARFM approach was developed for surface reflectance data and not for high level products such as ET which have the added uncertainty of the application of different retrieval algorithms. The dynamic ESTARFM approach accounts for these errors by applying a data-driven linear correction and also accounting for residual random error. This prevents the predictions from the dynamic ESTARFM from overfitting to the ECOSTRESS data. The dynamic ESTARFM also provides a framework for incorporating additional information (ET data from the year 2019 for this application) which helps to further minimize the errors and increase accuracy in predictions compared to the original ESTARFM which is totally unsupervised in its application.

Even for applications where the original ESTARFM performs well such as prediction of surface reflectance data, the dynamic ESTARFM can be used to find the optimal values of important inputs to the ESTARFM such as the number of clusters $n_c$ by disregarding any additional corrections ($\alpha_k = 1, \beta_k = 0$), fixing the error variance in equations 5.9 and 5.10 as very small, and maximizing the likelihood for different values of $n_c$. Such an approach will potentially give more accurate results and make the automatic implementation of the original ESTARFM feasible, though it will come at an added computational cost for estimating the optimal value of $n_c$.

### 5.5.2 Further Work- Soil moisture driven ET fusion

The purpose of this work was to extend the ESTARFM (and equivalently STARFM) approach to include uncertainty in the observed data and utilizing additional information from observed past data from 2019 to do corrections to the ESTARFM algorithm and improve ET predictions for the year 2020. But the applicability of the proposed work goes well beyond the present application. Since, the state-space framework of Dynamic ESTARFM allows addition of ancillary information, future work will focus on incorporating physical understanding of the underlying process into the data fusion process. The original ESTARFM algorithm is purely data-driven and the predictions are derived only from a remote sensing perspective. Adding domain understanding of the underlying processes to the ESTARFM approach can potentially improve predictions and thus the state-space approach can be a step forward into “Theory-Driven Data fusion” which combines the empirical nature of data fusion algorithms with the process understanding of the physical processes. As an example for ET estimation, future work can focus on combining the soil moisture (SM) - ET relationship with the ESTARFM data fusion approach. In such a case, the evolution model in equation 5.6 becomes:

$$x_{t_{k+1}}(s_i) = x_{t_k}(s_i) + (1 - \frac{w}{\delta_{tk}})f(SM(s_i), ET(s_i)) + \frac{w}{\delta_{tk}}E_{t_k}(s_i) + \eta(s_i) \tag{5.13}$$

where $\delta_{tk}$ refers to the time interval between the observed ECOSTRESS days $t_k$ and $t_{k+1}$, $w$ is a weighting factor, $f(SM(s_i), ET(s_i))$ refers to the function denoting the ET-SM relationship which has been studied at various scales [196, 197, 198], and $E_{t_k}(s_i)$ is the input from the ESTARFM approach. The weighting function w can be estimated from the observed data. The ESTARFM approach computes relatively accurate predictions when the underlying change in ET from dates $t_{k+1}$ and $t_k$ can be assumed to be linear but the predictions degrade as this change becomes non-linear which usually occurs when the time interval between successive ET retrievals increases. The weighting function $w$ is thus weighted by the time interval $\delta_{tk}$. Together with the same observation model in equation 5.10, the dynamic ESTARFM approach can thus effortlessly incorporate the ET-SM relationship using SM data from various remote sensing platforms. This can result in better predictions for both SM and ET for a given study area. Future work will focus on different applications of the ESTARFM approach constrained by such a theory-driven process understanding leading to increased predictive accuracy.

## 5.6 Conclusion

Remote sensing provides a cost-effective way to measure disparate environmental variables across the globe and thus in the past few years, there has been growing interest in data fusion techniques which can combine data from different remote sensing platforms. In this work, we propose a novel data fusion algorithm called the Dynamic ESTARFM which improves upon the existing STARFM/ESTARFM algorithm by accounting for measurement errors in the platforms, providing a framework to find optimal parameters based on observed data and increasing the potential for the inclusion of ancillary data as well as theory-driven data fusion in the hitherto unsupervised STARFM/ESTARFM algorithms. We apply the proposed technique to three agricultural sites in the Lower Brazos Basin, Texas and found that the Dynamic ESTARFM outperforms the original

ESTARFM approach. Future research will focus on addition of increased theory-based constraints in the Dynamic ESTARFM such as the one mentioned in Section 5.5.2. Better characterization of the measurement errors as well as a non-stationary spatial covariance structure [1] in the evolution model (equation 5.10) can also help in further increasing accuracy. The STARFM/ESTARFM approaches and by design the Dynamic ESTARFM are computationally intensive due to the moving window algorithm (equation 5.1) and future research will also focus on efficient ways to approximate this technique so that these data fusion algorithms can be applied at big spatio-temporal scales.

# 6. CONCLUSIONS

The deluge of remote sensing data in recent years presents an unprecedented opportunity to gain novel insights into multiscale dynamics of hydrological variables from local to global levels. Remote sensing platforms, however, suffer from several limitations such as incomplete spatio-temporal coverage, errors in retrievals, and retrieving data at a resolution which might be different than the one desired for a particular application. This study presented multiscale data fusion algorithms for hydrological variables such as soil moisture and evapotranspiration which account for the above limitations of remote sensing platforms while also incorporating the effect of static and dynamic physical controls on the spatio-temporal distribution of these variables.

In Chapter 2, a non-stationary geostatistical framework for soil moisture was proposed which accounts for the effect of physical controls such as soil texture, elevation and vegetation on the variance and spatial correlation of soil moisture. The framework was applied to a watershed exhibiting significant spatial heterogeneity in surface conditions in Winnipeg, Canada. It was seen that both vegetation and soil texture affect the variance/correlation of soil moisture governed by different stages of the crop-growing cycle and wetness conditions. The traditional assumption of stationarity in the variance/correlation of soil moisture was therefore found to be too simplistic.

In Chapter 3, the non-stationary algorithm proposed in Chapter 2 was utilized in a Bayesian Hierarchical framework to present a data fusion scheme combining multiscale soil moisture platforms. The data fusion framework accounted for uncertainty in remote sensing retrievals, change of support between different data platforms, effect of physical controls on soil moisture distribution, and uncertainty in parameters in the form of probability distributions. The data fusion framework was applied to the same watershed in Winnipeg, Canada to combine point, airborne and satellite platforms and it was shown that the data fusion algorithm computed satisfactory predictions from point to satellite scales while quantifying the retrieval errors in remote sensing platforms.

Chapter 4 focused on extending the data fusion framework presented in Chapter 3 to a “Big” data setting where the number of observations and/or the size of the study domain becomes massive, for instance when fusing remote sensing data at continental or global scales. An approximation to the data fusion framework was proposed to make the fusion scheme computationally efficient for Big datasets. The fusion scheme is then applied to combine point and satellite data soil moisture data across Contiguous US and five day soil moisture forecasts are computed. The effect of physical controls such as elevation, soil texture, vegetation and rainfall on the soil moisture distribution is quantified and predictions are validated using soil moisture data at multiple scales.

In Chapter 5, a data fusion framework is proposed for ET prediction at a fine spatial resolution when faced with the data limitation of fine resolution ET data being temporally sparse and only coarse resolution ET data being available frequently. Data fusion approaches exist for combining data for such scenarios and this work improves upon the existing fusion approaches by accounting for uncertainty in remote sensing retrievals, providing a framework to find optimal parameters based on the observed data, and allowing the inclusion of ancillary data and domain knowledge in the data fusion algorithms. The proposed fusion scheme is applied to three agricultural sites in the lower Brazos basin, Texas USA for the year 2020. It was found that the proposed fusion scheme was more accurate than the traditional data fusion approach when validated with daily Eddy-Covariance evapotranspiration data.

This dissertation will motivate future “Theory-Driven Data Fusion” hydrological research studies combining knowledge of the underlying physical processes with multiscale Big data from disparate remote sensing platforms. Improved formulations of non-stationary data fusion schemes which are motivated by the underlying physical processes and covariate heterogeneity need to be explored. The objective of developing a data fusion scheme should not be motivated to just get a good fit to the data but also further our understanding of environmental processes across scales. The volume of remote sensing data is going to exponentially increase in the next decade and therefore further work in proposing physically-based computationally efficient fusion techniques becomes pertinent. Recently, Machine Learning techniques have been instrumental in improving predictive accuracy of various hydrological processes [129, 130, 131, 132, 133]. The existing Machine Learning algorithms, however, lack interpretability as well as are ill-equipped to handle multiscale data and account for meaurement errors in satellites [134]. Incorporating Machine Learning algorithms into physically motivated data fusion frameworks thus is a promising area of future research. Future work should also focus on coupling remote sensing driven data fusion schemes to better constrain large scale numerical models potentially enabling the scientific community to get one step closer to achieve the elusive mass and energy balance in Earth-System processes.

# REFERENCES

[1] D. Kathuria, B. Mohanty, and M. Katzfuss, “A Nonstationary Geostatistical Framework for Soil Moisture Prediction in the Presence of Surface Heterogeneity,” *Water Resources Research*, vol. 55, no. 1, 2019.

[2] D. Kathuria, B. Mohanty, and M. Katzfuss, “Multiscale Data Fusion for Surface Soil Moisture Estimation: A Spatial Hierarchical Approach,” *Water Resources Research*, vol. 55, no. 12, 2019.

[3] J. D. Bolten, W. T. Crow, X. Zhan, T. J. Jackson, and C. A. Reynolds, “Evaluating the utility of remotely sensed soil moisture retrievals for operational agricultural drought monitoring,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 3, pp. 57–66, oct 2010.

[4] S. I. Seneviratne, T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, and A. J. Teuling, “Investigating soil moisture–climate interactions in a changing climate: A review,” *Earth-Science Reviews*, vol. 99, pp. 125–161, oct 2010.

[5] J. Pastor and W. M. Post, “Influence of climate, soil moisture, and succession on forest carbon and nitrogen cycles,” *Biogeochemistry*, vol. 2, pp. 3–27, oct 1986.

[6] P. Falloon, C. D. Jones, M. Ades, and K. Paul, “Direct soil moisture controls of future global soil carbon changes: An important source of uncertainty,” *Global Biogeochemical Cycles*, vol. 25, pp. n/a–n/a, nov 2011.

[7] E. M. Fischer, S. I. Seneviratne, P. L. Vidale, D. Lüthi, and C. Schär, “Soil Moisture–Atmosphere Interactions during the 2003 European Summer Heat Wave,” *Journal of climate*, vol. 20, pp. 5081–5099, jan 2007.

[8] D. G. Miralles, A. J. Teuling, C. C. van Heerwaarden, and J. Vilà-Guerau de Arellano, “Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accu-

mulation," *Nature Geoscience*, vol. 7, pp. 345–349, jan 2014.

[9] N. Wanders, D. Karssenberg, A. de Roo, S. M. de Jong, and M. F. P. Bierkens, "The suitability of remotely sensed soil moisture for improving operational flood forecasting," *Hydrology and Earth System Sciences*, vol. 18, pp. 2343–2357, feb 2014.

[10] W. Cai, T. Cowan, P. Briggs, and M. Raupach, "Rising temperature depletes soil moisture and exacerbates severe drought conditions across southeast Australia," *Geophysical Research Letters*, vol. 36, oct 2009.

[11] D. Chaparro, M. Vall-Llossera, M. Piles, A. Camps, and C. Rüdiger, "Low soil moisture and high temperatures as indicators for forest fire occurrence and extent across the Iberian Peninsula.," *Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International*, p. 3325, jan 2015.

[12] J. K. Entin, A. Robock, K. Y. Vinnikov, S. E. Hollinger, S. Liu, and A. Namkhai, "Temporal and spatial scales of observed soil moisture variations in the extratropics," *Journal of Geophysical Research: Atmospheres*, vol. 105, pp. 11865–11877, may 2000.

[13] M. H. Cosh and W. Brutsaert, "Aspects of soil moisture variability in the Washita '92 study region," *Journal of Geophysical Research: Atmospheres*, vol. 104, pp. 19751–19757, aug 1999.

[14] N. Gaur and B. P. Mohanty, "Land-surface controls on near-surface soil moisture dynamics: Traversing remote sensing footprints," *Water Resources Research*, 2016.

[15] C. Joshi, B. P. Mohanty, J. M. Jacobs, and A. V. M. Ines, "Spatiotemporal analyses of soil moisture from point to footprint scale in two different hydroclimatic regions," *Water Resources Research*, vol. 47, jan 2011.

[16] D. Ryu and J. S. Famiglietti, "Multi-scale spatial correlation and scaling behavior of surface soil moisture," *Geophysical Research Letters*, vol. 33, apr 2006.

[17] G. Kim and A. Barros, "Downscaling of remotely sensed soil moisture with a modified fractal interpolation method using contraction mapping and ancillary data," *Remote Sensing of Environment*, vol. 83, pp. 400–413, oct 2002.

[18] B. P. Mohanty, J. S. Famiglietti, and T. H. Skaggs, "Evolution of soil moisture spatial structure in a mixed vegetation pixel during the Southern Great Plains 1997 (SGP97) Hydrology Experiment," *Water Resources Research*, vol. 36, pp. 3675–3686, dec 2000.

[19] M. E. Hawley, T. J. Jackson, and R. H. McCuen, "Surface soil moisture variation on small agricultural watersheds," *Journal of hydrology*, vol. 62, pp. 179–200, nov 1983.

[20] W. Korres, T. G. Reichenau, P. Fiener, C. N. Koyama, H. R. Bogena, T. Cornelissen, R. Baatz, M. Herbst, B. Diekkrüger, H. Vereecken, and K. Schneider, "Spatio-temporal soil moisture patterns – A meta-analysis using plot to catchment scale data," *Journal of hydrology*, vol. 520, pp. 326–341, sep 2015.

[21] H. Vereecken, J. A. Huisman, Y. Pachepsky, C. Montzka, J. van der Kruk, H. Bogena, L. Weihermüller, M. Herbst, G. Martinez, and J. Vanderborght, "On the spatio-temporal dynamics of soil moisture at the field scale," *Journal of Hydrology*, vol. 516, pp. 76–96, 2014.

[22] W. T. Crow, D. Ryu, and J. S. Famiglietti, "Upscaling of field-scale soil moisture measurements using distributed land surface modeling," *Advances in water resources*, vol. 28, pp. 1–14, sep 2005.

[23] D. Ryu and J. S. Famiglietti, "Characterization of footprint-scale surface soil moisture variability using Gaussian and beta distribution functions during the Southern Great Plains 1997 (SGP97) hydrology experiment," *Water resources research*, vol. 41, sep 2005.

[24] J. A. Huisman, J. J. J. C. Snepvangers, W. Bouten, and G. B. M. Heuvelink, "Monitoring temporal development of spatial soil water content variation," *Vadose Zone Journal*, vol. 2, p. 519, nov 2003.

[25] A. W. Western and G. Blöschl, “On the spatial scaling of soil moisture,” *Journal of hydrology*, vol. 217, pp. 203–224, jun 1999.

[26] A. W. Western, R. B. Grayson, and G. Blöschl, “SCALING OF SOIL MOISTURE : A Hydrologic Perspective,” *Annual review of earth and planetary sciences*, vol. 30, pp. 149–180, sep 2002.

[27] T. E. Ochsner, M. H. Cosh, R. H. Cuenca, W. A. Dorigo, C. S. Draper, Y. Hagimoto, Y. H. Kerr, E. G. Njoku, E. E. Small, and M. Zreda, “State of the Art in Large-Scale Soil Moisture Monitoring,” *Soil Science Society of America Journal*, vol. 77, p. 1888, jun 2013.

[28] H. Nguyen, N. Cressie, and A. Braverman, “Spatial statistical data fusion for remote sensing applications,” *Journal of the American Statistical Association*, vol. 107, pp. 1004–1018, apr 2012.

[29] R. G. Allen, L. S. Pereira, T. A. Howell, and M. E. Jensen, “Evapotranspiration information reporting: I. Factors governing measurement accuracy,” *Agricultural Water Management*, vol. 98, no. 6, pp. 899–920, 2011.

[30] H. J. Farahani, T. A. Howell, W. J. Shuttleworth, and W. C. Bausch, “E : p m m a,” vol. 50, no. 5, pp. 1627–1638, 2007.

[31] J. A. Otkin, M. C. Anderson, C. Hain, and M. Svoboda, “Examining the relationship between drought development and rapid changes in the evaporative stress index,” *Journal of Hydrometeorology*, vol. 15, no. 3, pp. 938–956, 2014.

[32] S. M. Vicente-Serrano, S. Beguería, and J. I. López-Moreno, “A multiscalar drought index sensitive to global warming: The standardized precipitation evapotranspiration index,” *Journal of Climate*, vol. 23, no. 7, pp. 1696–1718, 2010.

[33] D. E. Armanios and J. B. Fisher, “Measuring water availability with limited ground data: Assessing the feasibility of an entirely remote-sensing-based hydrologic budget of the Rufiji Basin, Tanzania, using TRMM, GRACE, MODIS, SRB, and AIRS,” *Hydrological Processes*, vol. 28, no. 3, pp. 853–867, 2014.

[34] Y. Chen, J. Xia, S. Liang, J. Feng, J. B. Fisher, X. Li, X. Li, S. Liu, Z. Ma, A. Miyata, Q. Mu, L. Sun, J. Tang, K. Wang, J. Wen, Y. Xue, G. Yu, T. Zha, L. Zhang, Q. Zhang, T. Zhao, L. Zhao, and W. Yuan, "Comparison of satellite-based evapotranspiration models over terrestrial ecosystems in China," *Remote Sensing of Environment*, vol. 140, pp. 279–293, 2014.

[35] M. C. Anderson, R. G. Allen, A. Morse, and W. P. Kustas, "Use of Landsat thermal imagery in monitoring evapotranspiration and managing water resources," *Remote Sensing of Environment*, vol. 122, pp. 50–65, 2012.

[36] R. M. Rabin, S. Stadler, P. J. Wetzel, D. J. Stensrud, and M. Gregory, "Observed effects of landscape variability on convective clouds," *Bulletin - American Meteorological Society*, vol. 71, no. 3, pp. 272–280, 1990.

[37] M. Drusch, "Initializing numerical weather prediction models with satellite-derived surface soil moisture: Data assimilation experiments with ECMWF's Integrated Forecast System and the TMI soil moisture data set," *Journal of Geophysical Research*, vol. 112, oct 2007.

[38] F. Fécan, B. Marticorena, and G. Bergametti, "Parametrization of the increase of the aeolian erosion threshold wind friction velocity due to soil moisture for arid and semi-arid areas," *Annales Geophysicae*, vol. 17, pp. 149–157, oct 1999.

[39] B. Laurent, B. Marticorena, G. Bergametti, J. F. Léon, and N. M. Mahowald, "Modeling mineral dust emissions from the Sahara desert using new surface properties and soil database," *Journal of Geophysical Research*, vol. 113, oct 2008.

[40] Z. Chen, B. P. Mohanty, and I. Rodriguez-Iturbe, "Space-time modeling of soil moisture," *Advances in water resources*, jun 2017.

[41] J. Kim and B. P. Mohanty, "A physically based hydrological connectivity algorithm for describing spatial patterns of soil moisture in the unsaturated zone," *Journal of Geophysical Research: Atmospheres*, vol. 122, pp. 2096–2114, jun 2017.

[42] F. Giorgi and R. Avissar, “Representation of heterogeneity effects in Earth system modeling: Experience from land surface modeling,” *Reviews of Geophysics*, vol. 35, pp. 413–437, nov 1997.

[43] W. T. Crow, “Impact of soil moisture aggregation on surface energy flux prediction during SGP’97,” *Geophysical Research Letters*, vol. 29, p. 1008, nov 2002.

[44] T. Wang, T. E. Franz, V. A. Zlotnik, J. You, and M. D. Shulski, “Investigating soil controls on soil moisture spatial variability: Numerical simulations and field observations,” *Journal of hydrology*, vol. 524, pp. 576–586, sep 2015.

[45] N. Gaur and B. P. Mohanty, “Evolution of physical controls for soil moisture in humid and subhumid watersheds,” *Water Resources Research*, 2013.

[46] C. Joshi and B. P. Mohanty, “Physical controls of near-surface soil moisture across varying spatial scales in an agricultural landscape during SMEX02,” *Water Resources Research*, vol. 46, no. 12, pp. 1–21, 2010.

[47] W. T. Crow, A. A. Berg, M. H. Cosh, A. Loew, B. P. Mohanty, R. Panciera, P. de Rosnay, D. Ryu, and J. P. Walker, “Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products,” *Reviews of Geophysics*, vol. 50, jun 2012.

[48] A. J. Teuling and P. A. Troch, “Improved understanding of soil moisture variability dynamics,” *Geophysical Research Letters*, vol. 32, mar 2005.

[49] J. E. Lawrence and G. M. Hornberger, “Soil moisture variability across climate zones,” *Geophysical Research Letters*, vol. 34, sep 2007.

[50] A. W. Western, R. B. Grayson, G. Blöschl, and D. J. Wilson, “Spatial variability of soil moisture and its implications for scaling.,” *Scaling methods in soil physics*, pp. 119–142, feb 2003.

[51] U. Rosenbaum, H. R. Bogena, M. Herbst, J. A. Huisman, T. J. Peterson, A. Weuthen, A. W. Western, and H. Vereecken, "Seasonal and event dynamics of spatial soil moisture patterns at the small catchment scale," *Water resources research*, vol. 48, feb 2012.

[52] Z. Hu, S. Islam, and Y. Cheng, "Statistical characterization of remotely sensed soil moisture images," *Remote Sensing of Environment*, vol. 61, pp. 310–318, nov 1997.

[53] N. N. Das and B. P. Mohanty, "Temporal dynamics of PSR-based soil moisture across spatial scales in an agricultural landscape during SMEX02: A wavelet approach," *Remote Sensing of Environment*, vol. 112, pp. 522–534, feb 2008.

[54] B. P. Mohanty and T. H. Skaggs, "Spatio-temporal evolution and time-stable characteristics of soil moisture within remote sensing footprints with varying soil, slope, and vegetation," *Advances in water resources*, vol. 24, pp. 1051–1067, jun 2001.

[55] K. Vanderlinden, H. Vereecken, H. Hardelauf, M. Herbst, G. Martínez, M. H. Cosh, and Y. A. Pachepsky, "Temporal stability of soil water contents: A review of data and analyses," *Vadose Zone Journal*, vol. 11, p. 0, feb 2012.

[56] B. Fang, V. Lakshmi, R. Bindlish, T. J. Jackson, M. Cosh, and J. Basara, "Passive microwave soil moisture downscaling using vegetation index and skin surface temperature.," *Vadose Zone Journal*, vol. 12, jun 2013.

[57] S. Ahmad, A. Kalra, and H. Stephen, "Estimating soil moisture using remote sensing data: A machine learning approach," *Advances in water resources*, vol. 33, pp. 69–80, sep 2010.

[58] P. K. Srivastava, D. Han, M. R. Ramirez, and T. Islam, "Machine learning techniques for downscaling SMOS satellite soil moisture using MODIS land surface temperature for hydrological application," *Water Resources Management*, vol. 27, pp. 3127–3144, feb 2013.

[59] A. K. Sahoo, G. J. M. De Lannoy, R. H. Reichle, and P. R. Houser, "Assimilation and downscaling of satellite observed soil moisture over the Little River Experimental Watershed in Georgia, USA," *Advances in water resources*, vol. 52, pp. 19–33, jun 2013.

[60] H. Lievens, S. K. Tomer, A. Al Bitar, G. J. M. De Lannoy, M. Drusch, G. Dumedah, H.-J. Hendricks Franssen, Y. H. Kerr, B. Martens, M. Pan, J. K. Roundy, H. Vereecken, J. P. Walker, E. F. Wood, N. E. C. Verhoest, and V. R. N. Pauwels, "SMOS soil moisture assimilation for improved hydrologic simulation in the Murray Darling Basin, Australia," *Remote Sensing of Environment*, vol. 168, pp. 146–162, jun 2015.

[61] M. Katzfuss and J. Guinness, "A general framework for Vecchia approximations of Gaussian processes," 2017.

[62] M. Katzfuss and J. Guinness, "A general framework for Vecchia approximations of Gaussian processes," pp. 1–22, 2017.

[63] M. Fuentes, "Spectral methods for nonstationary spatial processes," *Biometrika*, vol. 89, pp. 197–210, oct 2002.

[64] B. J. Reich, J. Eidsvik, M. Guindani, A. J. Nail, and A. M. Schmidt, "A class of covariate-dependent spatiotemporal covariance functions.," *The annals of applied statistics*, vol. 5, pp. 2265–2687, sep 2011.

[65] A. E. Gelfand, L. Zhu, and B. P. Carlin, "On the change of support problem for spatio-temporal data.," *Biostatistics*, vol. 2, pp. 31–45, jan 2001.

[66] N. Cressie and G. Johannesson, "Fixed rank kriging for very large spatial data sets," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, pp. 209–226, dec 2008.

[67] R. M. Lark, "Kriging a soil variable with a simple nonstationary variance model," *Journal of Agricultural, Biological and Environmental Statistics*, vol. 14, pp. 301–321, dec 2009.

[68] A. M.J-C. Wadoux, D. J. Brus, and G. B. M. Heuvelink, "Accounting for non-stationary variance in geostatistical mapping of soil properties," *Geoderma*, vol. 324, pp. 138–147, dec 2018.

[69] M. C. Peel, B. L. Finlayson, and T. A. McMahon, “Updated world map of the Köppen-Geiger climate classification,” *Hydrology and Earth System Sciences*, vol. 11, pp. 1633–1644, nov 2007.

[70] H. McNairn, T. J. Jackson, G. Wiseman, S. Belair, A. Berg, P. Bullock, A. Colliander, M. H. Cosh, S.-B. Kim, R. Magagi, M. Moghaddam, E. G. Njoku, J. R. Adams, S. Homayouni, E. Ojo, T. Rowlandson, J. Shang, K. Goita, and M. Hosseini, “The soil moisture active passive validation experiment 2012 (SMAPVEX12): prelaunch calibration and validation of the SMAP soil moisture algorithms,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, pp. 2784–2801, sep 2015.

[71] A. W. Western, S.-L. Zhou, R. B. Grayson, T. A. McMahon, G. Blöschl, and D. J. Wilson, “Spatial correlation of soil moisture in small catchments and its relationship to dominant spatial hydrological processes,” *Journal of hydrology*, vol. 286, pp. 113–134, sep 2004.

[72] N. Cressie, “Fitting variogram models by weighted least squares,” *Journal of the International Association for Mathematical Geology*, vol. 17, pp. 563–586, jun 1985.

[73] M. G. Genton, “Highly Robust Variogram Estimation,” *Mathematical Geology*, p. 30: 213, nov 1998.

[74] R. M. Lark, “A comparison of some robust estimators of the variogram for use in soil survey,” *European journal of soil science*, vol. 51, pp. 137–157, nov 2000.

[75] A. E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes, *Handbook of spatial statistics*. CRC press, 2010.

[76] M. L. Stein, *Interpolation of spatial data: some theory for kriging*. Springer Science Business Media, jun 2012.

[77] P. J. Diggle and P. J. Ribeiro, *Model-based Geostatistics*. New York, NY: Springer New York, jun 2007.

[78] J. Warnes and B. Ripley, “Problems with likelihood estimation of covariance functions of spatial gaussian processes,” *Biometrika*, vol. 74, no. 3, pp. 640–642, 1987.

[79] K. M. Mullen, “Continuous global optimization in R.,” *Journal of Statistical Software*, vol. 60, pp. 1–45, nov 2014.

[80] Y. Xiang, S. Gubian, B. Suomela, and J. Hoeng, “Generalized simulated annealing for global optimization: The GenSA package,” *R Journal*, 2013.

[81] W. R. M. Jr and J. S. Sekhon, “Genetic optimization using derivatives: the rgenoud package for R.,” *Journal of Statistical Software*, vol. 42, pp. 1–26, nov 2011.

[82] Y. Yang, “Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation,” *Biometrika*, vol. 92, no. 4, pp. 937–950, 2005.

[83] R. E. Kass and A. E. Raftery, “Bayes factors,” *Journal of the american statistical association*, vol. 90, no. 430, pp. 773–795, 1995.

[84] E. G. Njoku, W. J. Wilson, S. H. Yueh, S. J. Dinardo, F. K. Li, T. J. Jackson, V. Lakshmi, and J. Bolten, “Observations of soil moisture using a passive and active low-frequency microwave airborne sensor during SGP99,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, pp. 2659–2673, nov 2002.

[85] R. D. Koster, P. A. Dirmeyer, Z. Guo, G. Bonan, E. Chan, P. Cox, C. T. Gordon, S. Kanae, E. Kowalczyk, D. Lawrence, P. Liu, C.-H. Lu, S. Malyshev, B. McAvaney, K. Mitchell, D. Mocko, T. Oki, K. Oleson, A. Pitman, Y. C. Sud, C. M. Taylor, D. Verseghy, R. Vasic, Y. Xue, T. Yamada, and G. Team, “Regions of strong coupling between soil moisture and precipitation.,” *Science*, vol. 305, pp. 1138–1140, oct 2004.

[86] S. Tuttle and G. Salvucci, “Empirical evidence of contrasting soil moisture-precipitation feedbacks across the United States.,” *Science*, vol. 352, pp. 825–828, apr 2016.

[87] R. Koster, S. Schubert, and M. Suarez, “Analyzing the concurrence of meteorological droughts and warm periods, with implications for the determination of evaporative regime,” *Journal of Climate*, vol. 22, no. 12, pp. 3331–3341, 2009.

[88] S. I. Seneviratne, D. Lüthi, M. Litschi, and C. Schär, “Land-atmosphere coupling and climate change in Europe.,” *Nature*, vol. 443, pp. 205–209, feb 2006.

[89] M. H. Cosh, T. J. Jackson, R. Bindlish, and J. H. Prueger, “Watershed scale temporal and spatial stability of soil moisture and its role in validating satellite estimates.,” *Remote sensing of Environment*, vol. 92, pp. 427–435, jun 2004.

[90] D. Kathuria, B. P. Mohanty, and M. Katzfuss, “Multiscale Data Fusion for Surface Soil Moisture Estimation: A Spatial Hierarchical Approach,” *Water Resources Research*, vol. 55, no. 12, pp. 10443–10465, 2019.

[91] J. O. Skøien, G. Blöschl, and A. W. Western, “Characteristic space scales and timescales in hydrology,” *Water resources research*, vol. 39, jun 2003.

[92] T. J. Jackson, T. J. Schmugge, and J. R. Wang, “Passive microwave sensing of soil moisture under vegetation canopies.,” *Water Resources Research*, vol. 18, pp. 1137–1142, jun 1982.

[93] E. G. Njoku and D. Entekhabi, “Passive microwave remote sensing of soil moisture.,” *Journal of hydrology*, vol. 184, pp. 101–129, jun 1996.

[94] M. Neelam and B. P. Mohanty, “Global sensitivity analysis of the radiative transfer model,” *Water resources research*, vol. 51, pp. 2428–2443, jan 2015.

[95] S. C. Steele-Dunne, M. M. Rutten, D. M. Krzeminska, M. Hausner, S. W. Tyler, J. Selker, T. A. Bogaard, and N. C. van de Giesen, “Feasibility of soil moisture estimation using passive distributed temperature sensing,” *Water resources research*, vol. 46, jun 2010.

[96] J. Peng, A. Loew, O. Merlin, and N. E. C. Verhoest, “A review of spatial downscaling of satellite remotely sensed soil moisture,” *Reviews of Geophysics*, vol. 55, pp. 341–366, sep 2017.

[97] N. N. Das, D. Entekhabi, and E. G. Njoku, “An Algorithm for Merging SMAP Radiometer and Radar Data for High-Resolution Soil-Moisture Retrieval,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, pp. 1504–1512, may 2011.

[98] C. Montzka, T. Jagdhuber, R. Horn, H. R. Bogena, I. Hajnsek, A. Reigber, and H. Vereecken, “Investigation of SMAP Fusion Algorithms With Airborne Active and Passive L-Band Mi-

crowave Remote Sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, pp. 3878–3889, nov 2016.

[99] M. Piles, A. Camps, M. Vall-llossera, I. Corbella, R. Panciera, C. Rudiger, Y. H. Kerr, and J. Walker, "Downscaling SMOS-Derived Soil Moisture Using MODIS Visible/Infrared Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, pp. 3156–3166, nov 2011.

[100] J. Kim and T. S. Hogue, "Improving Spatial Soil Moisture Representation Through Integration of AMSR-E and MODIS Products," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, pp. 446–460, may 2012.

[101] O. Merlin, A. Chehbouni, G. Boulet, and Y. Kerr, "Assimilation of Disaggregated Microwave Soil Moisture into a Hydrologic Model Using Coarse-Scale Meteorological Data," *Journal of Hydrometeorology*, vol. 7, pp. 1308–1322, may 2006.

[102] Y. Shin and B. P. Mohanty, "Development of a deterministic downscaling algorithm for remote sensing soil moisture footprint using soil and vegetation classifications," *Water resources research*, vol. 49, pp. 6208–6228, may 2013.

[103] N. Gaur and B. P. Mohanty, "A Nomograph to Incorporate Geo-Physical Heterogeneity in Soil Moisture Downscaling," *Water resources research*, nov 2018.

[104] R. Akbar, D. J. Short Gianotti, G. D. Salvucci, and D. Entekhabi, "Mapped Hydroclimatology of Evapotranspiration and Drainage Runoff Using SMAP Brightness Temperature Observations and Precipitation Information," *Water Resources Research*, vol. 55, no. 4, pp. 3391–3413, 2019.

[105] N. E. C. Verhoest, M. J. van den Berg, B. Martens, H. Lievens, E. F. Wood, M. Pan, Y. H. Kerr, A. Al Bitar, S. K. Tomer, M. Drusch, H. Vernieuwe, B. De Baets, J. P. Walker, G. Dumedah, and V. R. N. Pauwels, "Copula-Based Downscaling of Coarse-Scale Soil Moisture Observations With Implicit Bias Correction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, pp. 3507–3521, nov 2015.

[106] N. Cressie and C. K. Wikle, *Statistics for spatio-temporal data*. John Wiley  Sons, jan 2015.

[107] A. A. Bitar, A. Mialon, Y. H. Kerr, F. Cabot, P. Richaume, E. Jacquette, A. Quesney, A. Mahmoodi, S. Tarot, M. Parrens, *et al.*, "The global smos level 3 daily soil moisture and brightness temperature maps," *Earth System Science Data*, vol. 9, no. 1, pp. 293–315, 2017.

[108] C. K. Wikle, "HIERARCHICAL BAYESIAN MODELS FOR PREDICTING THE SPREAD OF ECOLOGICAL PROCESSES," *Ecology*, apr 2003.

[109] L. M. Berliner, "Hierarchical bayesian time series models," in *Maximum entropy and bayesian methods* (K. M. Hanson and R. N. Silver, eds.), pp. 15–22, Dordrecht: Springer Netherlands, jan 1996.

[110] C. K. Wikle, "Hierarchical modeling with spatial data.," *Handbook of Spatial Statistics*, pp. 96–113, jun 2010.

[111] N. Cressie, "Aggregation and interaction issues in statistical modeling of spatiotemporal processes.," *Geoderma*, vol. 85, pp. 133–140, apr 1998.

[112] H. Nguyen, M. Katzfuss, N. Cressie, and A. Braverman, "Spatio-Temporal Data Fusion for Very Large Remote Sensing Datasets," *Technometrics : a journal of statistics for the physical, chemical, and engineering sciences*, vol. 56, pp. 174–185, aug 2014.

[113] P. S. Narvekar, D. Entekhabi, S.-B. Kim, and E. G. Njoku, "Soil Moisture Retrieval Using L-Band Radar Observations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, pp. 3492–3506, aug 2015.

[114] C. K. Wikle and L. M. Berliner, "Combining Information Across Spatial Scales," *Technometrics : a journal of statistics for the physical, chemical, and engineering sciences*, vol. 47, pp. 80–91, apr 2005.

[115] P. D. Hoff, *A First Course in Bayesian Statistical Methods*. New York, NY: Springer New York, may 2009.

[116] K. R. Bell, B. J. Blanchard, T. J. Schmugge, and M. W. Witczak, “Analysis of surface moisture variations within large-field sites,” *Water resources research*, vol. 16, pp. 796–810, nov 1980.

[117] G. Buttafuoco, A. Castrignano, E. Busoni, and A. Dimase, “Studying the spatial structure evolution of soil water content using multivariate geostatistics,” *Journal of Hydrology*, vol. 311, no. 1-4, pp. 202–218, 2005.

[118] N. S. Chauhan, S. Miller, and P. Ardanuy, “Spaceborne soil moisture estimation at high resolution: a microwave-optical/IR synergistic approach,” *International journal of remote sensing*, vol. 24, pp. 4599–4622, nov 2003.

[119] C. K. Wikle, A. Zammit-Mangion, and N. Cressie, *Spatio-Temporal Statistics with R*. CRC Press, 2019.

[120] P. Bauer, A. Thorpe, and G. Brunet, “The quiet revolution of numerical weather prediction,” *Nature*, vol. 525, no. 7567, pp. 47–55, 2015.

[121] G. Evensen, “Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics,” *Journal of Geophysical Research*, vol. 99, no. C5, 1994.

[122] M. Ghil and P. Malanotte-Rizzoli, “Data Assimilation in Meteorology and Oceanography,” *Advances in Geophysics*, 1991.

[123] R. H. Reichle, D. B. McLaughlin, and D. Entekhabi, “Hydrologic data assimilation with the ensemble Kalman filter,” *Monthly Weather Review*, vol. 130, no. 1, pp. 103–114, 2002.

[124] B. Wang, X. Zou, and J. Zhu, “Data assimilation and its applications,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 21, pp. 11143–11144, 2000.

[125] H. Lievens, R. H. Reichle, Q. Liu, G. J. De Lannoy, R. S. Dunbar, S. B. Kim, N. N. Das, M. Cosh, J. P. Walker, and W. Wagner, “Joint Sentinel-1 and SMAP data assimilation to improve soil moisture estimates,” *Geophysical Research Letters*, 2017.

[126] M. Girotto, G. J. De Lannoy, R. H. Reichle, M. Rodell, C. Draper, S. N. Bhanja, and A. Mukherjee, "Benefits and pitfalls of GRACE data assimilation: A case study of terrestrial water storage depletion in India," *Geophysical Research Letters*, vol. 44, no. 9, pp. 4107–4115, 2017.

[127] G. Camps-valls, D. Tuia, and L. Bruzzone, "Advances in hyperspectral image classification," *IEEE Signal Processing Magazine*, no. January, pp. 45–54, 2013.

[128] T. Hengl, J. M. De Jesus, G. B. Heuvelink, M. R. Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, M. N. Wright, X. Geng, B. Bauer-Marschallinger, M. A. Guevara, R. Vargas, R. A. MacMillan, N. H. Batjes, J. G. Leenaars, E. Ribeiro, I. Wheeler, S. Mantel, and B. Kempen, "SoilGrids250m: Global gridded soil information based on machine learning," *PLoS ONE*, 2017.

[129] M. Jung, M. Reichstein, P. Ciais, S. I. Seneviratne, J. Sheffield, M. L. Goulden, G. Bonan, A. Cescatti, J. Chen, R. De Jeu, A. J. Dolman, W. Eugster, D. Gerten, D. Gianelle, N. Gobron, J. Heinke, J. Kimball, B. E. Law, L. Montagnani, Q. Mu, B. Mueller, K. Oleson, D. Papale, A. D. Richardson, O. Roupsard, S. Running, E. Tomelleri, N. Viovy, U. Weber, C. Williams, E. Wood, S. Zaehle, and K. Zhang, "Recent decline in the global land evapotranspiration trend due to limited moisture supply," *Nature*, vol. 467, no. 7318, pp. 951–954, 2010.

[130] H. Mao, D. Kathuria, N. Duffield, and B. P. Mohanty, "Gap Filling of High-Resolution Soil Moisture for SMAP/Sentinel-1: A Two-Layer Machine Learning-Based Framework," *Water Resources Research*, 2019.

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

[132] K. Fang, C. Shen, D. Kifer, and X. Yang, "Prolongation of SMAP to Spatiotemporally Seamless Coverage of Continental U.S. Using a Deep Learning Neural Network," *Geophysical*

*Research Letters*, vol. 44, no. 21, pp. 11,030–11,039, 2017.

[133] C. Shen, "A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists," *Water Resources Research*, vol. 54, no. 11, pp. 8558–8593, 2018.

[134] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, "Deep learning and process understanding for data-driven Earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, 2019.

[135] N. Cressie, "The origins of kriging," *Mathematical Geology*, vol. 22, no. 3, pp. 239–252, 1990.

[136] D. G. Krige, "A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand," *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 1952.

[137] F. Cecinati, M. A. Rico-Ramirez, G. B. Heuvelink, and D. Han, "Representing radar rainfall uncertainty with ensembles based on a time-variant geostatistical error modelling approach," *Journal of Hydrology*, 2017.

[138] M. Lanfredi, R. Coppola, M. D'Emilio, V. Imbrenda, M. Macchiato, and T. Simoniello, "A geostatistics-assisted approach to the deterministic approximation of climate data," *Environmental Modelling and Software*, vol. 66, pp. 69–77, 2015.

[139] R. M. Lark, "Towards soil geostatistics," *Spatial Statistics*, vol. 1, pp. 92–99, 2012.

[140] B. P. Mohanty, R. S. Kanwar, and R. Horton, "A Robust-Resistant Approach to Interpret Spatial Behavior of Saturated Hydraulic Conductivity of a Glacial Till Soil Under No-Tillage System," *Water Resources Research*, vol. 27, pp. 2979–2992, nov 1991.

[141] B. P. Mohanty, M. D. Ankeny, R. Horton, and R. S. Kanwar, "Spatial analysis of hydraulic conductivity measured using disc infiltrometers," *Water Resources Research*, vol. 30, pp. 2489–2498, sep 1994.

[142] B. P. Mohanty and R. S. Kanwar, “Spatial variability of residual nitrate-nitrogen under two tillage systems in central Iowa: A composite three-dimensional resistant and exploratory approach,” *Water Resources Research*, vol. 30, pp. 237–251, feb 1994.

[143] Z. Zhong and T. R. Carr, “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*, vol. 80, no. March 2018, pp. 61–75, 2019.

[144] P. Goovaerts, G. AvRuskin, J. Meliker, M. Slotnick, G. Jacquez, and J. Nriagu, “Geostatistical modeling of the spatial variability of arsenic in groundwater of southeast Michigan,” *Water Resources Research*, vol. 41, no. 7, pp. 1–19, 2005.

[145] M. D. Risser and C. A. Calder, “Regression-based covariance functions for nonstationary spatial modeling,” *Environmetrics*, 2015.

[146] C. G. Kaufman, M. J. Schervish, and D. W. Nychka, “Covariance tapering for likelihood-based estimation in large spatial data sets,” *Journal of the American Statistical Association*, 2008.

[147] D. Nychka, S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain, “A Multiresolution Gaussian Process Model for the Analysis of Large Spatial Datasets,” *Journal of Computational and Graphical Statistics*, 2015.

[148] A. V. Vecchia, “Estimation and Model Identification for Continuous Spatial Processes,” *Journal of the Royal Statistical Society: Series B (Methodological)*, 1988.

[149] J. Guinness, “Permutation and Grouping Methods for Sharpening Gaussian Process Approximations,” *Technometrics*, vol. 60, no. 4, pp. 415–429, 2018.

[150] M. Katzfuss, J. Guinness, W. Gong, and D. Zilber, “Vecchia Approximations of Gaussian-Process Predictions,” *Journal of Agricultural, Biological, and Environmental Statistics*, 2020.

[151] M. Katzfuss, “STAT 677 : Advanced Spatial Statistics,” pp. 1–62, 2018.

[152] R. D. Koster, L. Brocca, W. T. Crow, M. S. Burgin, and G. J. De Lannoy, "Precipitation estimation using L-band and C-band soil moisture retrievals," *Water Resources Research*, 2016.

[153] R. D. Koster, W. T. Crow, R. H. Reichle, and S. P. Mahanama, "Estimating Basin-Scale Water Budgets With SMAP Soil Moisture Data," *Water Resources Research*, vol. 54, no. 7, pp. 4228–4244, 2018.

[154] H. J. Diamond, T. R. Karl, M. A. Palecki, C. B. Baker, J. E. Bell, R. D. Leeper, D. R. Easterling, J. H. Lawrimore, T. P. Meyers, M. R. Helfert, G. Goodge, and P. W. Thorne, "U.S. climate reference network after one decade of operations status and assessment," *Bulletin of the American Meteorological Society*, vol. 94, no. 4, pp. 485–498, 2013.

[155] G. L. Schaefer, M. H. Cosh, and T. J. Jackson, "The USDA Natural Resources Conservation Service Soil Climate Analysis Network (SCAN)," *Journal of Atmospheric and Oceanic Technology*, vol. 24, no. 12, pp. 2073–2077, 2007.

[156] H. M. Barré, B. Duesmann, and Y. H. Kerr, "SMOS: The mission and the system," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 3, pp. 587–593, 2008.

[157] D. Entekhabi, E. G. Njoku, P. E. O'Neill, K. H. Kellogg, W. T. Crow, W. N. Edelstein, J. K. Entin, S. D. Goodman, T. J. Jackson, J. Johnson, J. Kimball, J. R. Piepmeier, R. D. Koster, N. Martin, K. C. McDonald, M. Moghaddam, S. Moran, R. Reichle, J. C. Shi, M. W. Spencer, S. W. Thurman, L. Tsang, and J. V. Zyl, "The Soil Moisture Active Passive (SMAP) Mission," *Proceedings of the IEEE*, vol. 98, no. 5, pp. 704–716, 2010.

[158] A. Colliander, T. J. Jackson, R. Bindlish, S. Chan, N. Das, S. B. Kim, M. H. Cosh, R. S. Dunbar, L. Dang, L. Pashaian, J. Asanuma, K. Aida, A. Berg, T. Rowlandson, D. Bosch, T. Caldwell, K. Caylor, D. Goodrich, H. al Jassar, E. Lopez-Baeza, J. Martínez-Fernández, A. González-Zamora, S. Livingston, H. McNairn, A. Pacheco, M. Moghaddam, C. Montzka, C. Notarnicola, G. Niedrist, T. Pellarin, J. Prueger, J. Pulliainen, K. Rautiainen, J. Ramos, M. Seyfried, P. Starks, Z. Su, Y. Zeng, R. van der Velde, M. Thibeault, W. Dorigo,

M. Vreugdenhil, J. P. Walker, X. Wu, A. Monerris, P. E. O'Neill, D. Entekhabi, E. G. Njoku, and S. Yueh, "Validation of SMAP surface soil moisture products with core validation sites," *Remote Sensing of Environment*, vol. 191, pp. 215–231, 2017.

[159] M. Pablos, M. Vall-llossera, M. Piles, A. Camps, C. Gonzalez-Haro, A. Turiel, C. J. Herbert, D. Chaparro, and G. Portal, "Influence of Quality Filtering Approaches in BEC SMOS L3 Soil Moisture Products," 2019.

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

[161] C. Daly, R. P. Neilson, and D. L. Phillips, "A statistical-topographic model for mapping climatological precipitation over mountainous terrain," *Journal of Applied Meteorology*, 1994.

[162] Soil Survey Staff, "Gridded Soil Survey Geographic (gSSURGO) Database for the Conterminous United States," 2020.

[163] T. Myneni, R., Knyazikhin, Y., Park, "MCD15A3H MODIS/Terra+Aqua Leaf Area Index/FPAR 4-day L4 Global 500m SIN Grid V006," 2015.

[164] X. Fan, Y. Liu, G. Gan, and G. Wu, "SMAP underestimates soil moisture in vegetation-disturbed areas primarily as a result of biased surface temperature data," *Remote Sensing of Environment*, vol. 247, no. November 2019, p. 111914, 2020.

[165] R. H. Reichle and R. D. Koster, "Bias reduction in short records of satellite soil moisture," *Geophysical Research Letters*, vol. 31, no. 19, pp. 2–5, 2004.

[166] S. Li, Y. Yu, D. Sun, D. Tarpley, X. Zhan, and L. Chiu, "Evaluation of 10 year AQUA/MODIS land surface temperature with SURFRAD observations," *International Journal of Remote Sensing*, vol. 35, no. 3, pp. 830–856, 2014.

[167] S. Westermann, M. Langer, and J. Boike, "Systematic bias of average winter-time land surface temperatures inferred from MODIS at a site on Svalbard, Norway," *Remote Sensing of Environment*, vol. 118, pp. 162–167, 2012.

[168] R. Klees, E. A. Zapreeva, H. C. Winsemius, and H. H. Savenije, “The bias in GRACE estimates of continental water storage variations,” *Hydrology and Earth System Sciences*, vol. 11, no. 4, pp. 1227–1241, 2007.

[169] G. Hu, L. Jia, and M. Menenti, “Comparison of MOD16 and LSA-SAF MSG evapotranspiration products over Europe for 2011,” *Remote Sensing of Environment*, vol. 156, pp. 510–526, 2015.

[170] N. M. Velpuri, G. B. Senay, R. K. Singh, S. Bohms, and J. P. Verdin, “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*, vol. 139, pp. 35–49, 2013.

[171] A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand, “Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets,” *Journal of the American Statistical Association*, 2016.

[172] Y. Sun and M. L. Stein, “Statistically and Computationally Efficient Estimating Equations for Large Spatial Datasets,” *Journal of Computational and Graphical Statistics*, 2016.

[173] N. N. Das, D. Entekhabi, R. S. Dunbar, S. Kim, S. Yueh, A. Colliander, P. E. O’Neill, and T. Jackson, “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*, 2018.

[174] D. Zilber and M. Katzfuss, “Vecchia-Laplace approximations of generalized Gaussian processes for big non-Gaussian spatial data,” 2019.

[175] C. O. Justice, J. R. Townshend, E. F. Vermote, E. Masuoka, R. E. Wolfe, N. Saleous, D. P. Roy, and J. T. Morisette, “An overview of MODIS Land data processing and product status,” *Remote Sensing of Environment*, vol. 83, no. 1-2, pp. 3–15, 2002.

[176] S. N. Goward, J. G. Masek, D. L. Williams, J. R. Irons, and R. J. Thompson, “The Landsat 7 mission: Terrestrial research and applications for the 21st century,” *Remote Sensing of Environment*, vol. 78, no. 1-2, pp. 3–12, 2001.

[177] J. B. Fisher, B. Lee, A. J. Purdy, G. H. Halverson, M. B. Dohlen, K. Cawse-Nicholson, A. Wang, R. G. Anderson, B. Aragon, M. A. Arain, D. D. Baldocchi, J. M. Baker, H. Barral, C. J. Bernacchi, C. Bernhofer, S. C. Biraud, G. Bohrer, N. Brunsell, B. Cappelaere, S. Castro-Contreras, J. Chun, B. J. Conrad, E. Cremonese, J. Demarty, A. R. Desai, A. De Ligne, L. Foltýnová, M. L. Goulden, T. J. Griffis, T. Grünwald, M. S. Johnson, M. Kang, D. Kelbe, N. Kowalska, J. H. Lim, I. Maïnassara, M. F. McCabe, J. E. Missik, B. P. Mohanty, C. E. Moore, L. Morillas, R. Morrison, J. W. Munger, G. Posse, A. D. Richardson, E. S. Russell, Y. Ryu, A. Sanchez-Azofeifa, M. Schmidt, E. Schwartz, I. Sharp, L. Šigut, Y. Tang, G. Hulley, M. Anderson, C. Hain, A. French, E. Wood, and S. Hook, "ECOSTRESS: NASA's Next Generation Mission to Measure Evapotranspiration From the International Space Station," *Water Resources Research*, vol. 56, no. 4, pp. 1–20, 2020.

[178] R. Torres, P. Snoeij, D. Geudtner, D. Bibby, M. Davidson, E. Attema, P. Potin, B. Ö. Rommen, N. Floury, M. Brown, I. N. Traver, P. Deghaye, B. Duesmann, B. Rosich, N. Miranda, C. Bruno, M. L'Abbate, R. Croci, A. Pietropaolo, M. Huchler, and F. Rostan, "GMES Sentinel-1 mission," *Remote Sensing of Environment*, vol. 120, pp. 9–24, 2012.

[179] X. Zhu, E. H. Helmer, F. Gao, D. Liu, J. Chen, and M. A. Lefsky, "A flexible spatiotemporal method for fusing satellite images with different resolutions," *Remote Sensing of Environment*, vol. 172, pp. 165–177, 2016.

[180] F. Gao, J. Masek, M. Schwaller, and F. Hall, "On the blending of the landsat and MODIS surface reflectance: Predicting daily landsat surface reflectance," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 8, pp. 2207–2218, 2006.

[181] T. Hilker, M. A. Wulder, N. C. Coops, N. Seitz, J. C. White, F. Gao, J. G. Masek, and G. Stenhouse, "Generation of dense time series synthetic Landsat data through data blending with MODIS using a spatial and temporal adaptive reflectance fusion model," *Remote Sensing of Environment*, vol. 113, no. 9, pp. 1988–1999, 2009.

[182] X. Zhu, J. Chen, F. Gao, X. Chen, and J. G. Masek, "An enhanced spatial and temporal

adaptive reflectance fusion model for complex heterogeneous regions," *Remote Sensing of Environment*, vol. 114, no. 11, pp. 2610–2623, 2010.

[183] C. M. Gevaert and F. J. García-Haro, "A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion," *Remote Sensing of Environment*, vol. 156, pp. 34–44, 2015.

[184] B. Zhukov, D. Oertel, F. Lanzl, and G. Reinhäckel, "Unmixing-based multisensor multiresolution image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 3 I, pp. 1212–1226, 1999.

[185] R. Zurita-Milla, J. G. Clevers, and M. E. Schaepman, "Unmixing-based landsat TM and MERIS FR data fusion," *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 3, pp. 453–457, 2008.

[186] B. Huang and H. Song, "Spatiotemporal reflectance fusion via sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 10 PART1, pp. 3707–3716, 2012.

[187] H. Song and B. Huang, "Spatiotemporal satellite image fusion through one-pair image learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 4, pp. 1883–1896, 2013.

[188] C. Cammalleri, M. C. Anderson, F. Gao, C. R. Hain, and W. P. Kustas, "A data fusion approach for mapping daily evapotranspiration at field scale," *Water Resources Research*, vol. 49, no. 8, pp. 4672–4686, 2013.

[189] Q. Mu, M. Zhao, and S. W. Running, "Improvements to a MODIS global terrestrial evapotranspiration algorithm," *Remote Sensing of Environment*, vol. 115, no. 8, pp. 1781–1800, 2011.

[190] M. C. Anderson, Y. Yang, J. Xue, K. R. Knipper, Y. Yang, F. Gao, C. R. Hain, W. P. Kustas, K. Cawse-Nicholson, G. Hulley, J. B. Fisher, J. G. Alfieri, T. P. Meyers, J. Prueger, D. D. Baldocchi, and C. Rey-Sanchez, "Interoperability of ECOSTRESS and Landsat for mapping

evapotranspiration time series at sub-field scales," *Remote Sensing of Environment*, vol. 252, no. October 2020, p. 112189, 2021.

[191] M. C. Anderson, J. M. Norman, J. R. Mecikalski, R. D. Torn, W. P. Kustas, and J. B. Basara, "A multiscale remote sensing model for disaggregating regional fluxes to micrometeorological scales," *Journal of Hydrometeorology*, vol. 5, no. 2, pp. 343–363, 2004.

[192] M. C. Anderson, J. M. Norman, J. R. Mecikalski, J. A. Otkin, and W. P. Kustas, "A climatological study of evapotranspiration and moisture stress across the continental United States based on thermal remote sensing: 1. Model formulation," *Journal of Geophysical Research Atmospheres*, vol. 112, no. 10, pp. 1–17, 2007.

[193] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications*. Springer International Publishing, 4 ed., 2017.

[194] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, vol. 82, pp. 35–45, mar 1960.

[195] G. Pastorello, C. Trotta, E. Canfora, H. Chu, D. Christianson, Y. W. Cheah, C. Poindexter, J. Chen, A. Elbashandy, M. Humphrey, P. Isaac, D. Polidori, A. Ribeca, C. van Ingen, L. Zhang, B. Amiro, C. Ammann, M. A. Arain, J. Ardö, T. Arkebauer, S. K. Arndt, N. Arriga, M. Aubinet, M. Aurela, D. Baldocchi, A. Barr, E. Beamesderfer, L. B. Marchesini, O. Bergeron, J. Beringer, C. Bernhofer, D. Berveiller, D. Billesbach, T. A. Black, P. D. Blanken, G. Bohrer, J. Boike, P. V. Bolstad, D. Bonal, J. M. Bonnefond, D. R. Bowling, R. Bracho, J. Brodeur, C. Brümmer, N. Buchmann, B. Burban, S. P. Burns, P. Buysse, P. Cale, M. Cavagna, P. Cellier, S. Chen, I. Chini, T. R. Christensen, J. Cleverly, A. Collalti, C. Consalvo, B. D. Cook, D. Cook, C. Coursolle, E. Cremonese, P. S. Curtis, E. D'Andrea, H. da Rocha, X. Dai, K. J. Davis, B. De Cinti, A. de Grandcourt, A. De Ligne, R. C. De Oliveira, N. Delpierre, A. R. Desai, C. M. Di Bella, P. di Tommasi, H. Dolman, F. Domingo, G. Dong, S. Dore, P. Duce, E. Dufrêne, A. Dunn, J. Dušek, D. Eamus, U. Eichelmann, H. A. M. ElKhidir, W. Eugster, C. M. Ewenz, B. Ewers, D. Famulari,

S. Fares, I. Feigenwinter, A. Feitz, R. Fensholt, G. Filippa, M. Fischer, J. Frank, M. Galvagno, M. Gharun, D. Gianelle, B. Gielen, B. Gioli, A. Gitelson, I. Goded, M. Goeckede, A. H. Goldstein, C. M. Gough, M. L. Goulden, A. Graf, A. Griebel, C. Gruening, T. Grünwald, A. Hammerle, S. Han, X. Han, B. U. Hansen, C. Hanson, J. Hatakka, Y. He, M. Hehn, B. Heinesch, N. Hinko-Najera, L. Hörtnagl, L. Hutley, A. Ibrom, H. Ikawa, M. Jackowicz-Korczynski, D. Janouš, W. Jans, R. Jassal, S. Jiang, T. Kato, M. Khomik, J. Klatt, A. Knohl, S. Knox, H. Kobayashi, G. Koerber, O. Kolle, Y. Kosugi, A. Kotani, A. Kowalski, B. Kruijt, J. Kurbatova, W. L. Kutsch, H. Kwon, S. Launiainen, T. Laurila, B. Law, R. Leuning, Y. Li, M. Liddell, J. M. Limousin, M. Lion, A. J. Liska, A. Lohila, A. López-Ballesteros, E. López-Blanco, B. Loubet, D. Loustau, A. Lucas-Moffat, J. Lüers, S. Ma, C. Macfarlane, V. Magliulo, R. Maier, I. Mammarella, G. Manca, B. Marcolla, H. A. Margolis, S. Marras, W. Massman, M. Mastepanov, R. Matamala, J. H. Matthes, F. Mazzenga, H. McCaughey, I. McHugh, A. M. McMillan, L. Merbold, W. Meyer, T. Meyers, S. D. Miller, S. Minerbi, U. Moderow, R. K. Monson, L. Montagnani, C. E. Moore, E. Moors, V. Moreaux, C. Moureaux, J. W. Munger, T. Nakai, J. Neirynck, Z. Nesic, G. Nicolini, A. Noormets, M. Northwood, M. Nosetto, Y. Nouvellon, K. Novick, W. Oechel, J. E. Olesen, J. M. Ourcival, S. A. Papuga, F. J. Parmentier, E. Paul-Limoges, M. Pavelka, M. Peichl, E. Pendall, R. P. Phillips, K. Pilegaard, N. Pirk, G. Posse, T. Powell, H. Prasse, S. M. Prober, S. Rambal, Ü. Rannik, N. Raz-Yaseef, D. Reed, V. R. de Dios, N. Restrepo-Coupe, B. R. Reverter, M. Roland, S. Sabbatini, T. Sachs, S. R. Saleska, E. P. Sánchez-Cañete, Z. M. Sanchez-Mejia, H. P. Schmid, M. Schmidt, K. Schneider, F. Schrader, I. Schroder, R. L. Scott, P. Sedlák, P. Serrano-Ortíz, C. Shao, P. Shi, I. Shironya, L. Siebicke, L. Šigut, R. Silberstein, C. Sirca, D. Spano, R. Steinbrecher, R. M. Stevens, C. Sturtevant, A. Suyker, T. Tagesson, S. Takanashi, Y. Tang, N. Tapper, J. Thom, F. Tiedemann, M. Tomassucci, J. P. Tuovinen, S. Urbanski, R. Valentini, M. van der Molen, E. van Gorsel, K. van Huissteden, A. Varlagin, J. Verfaillie, T. Vesala, C. Vincke, D. Vitale, N. Vygodskaya, J. P. Walker, E. Walter-Shea, H. Wang, R. Weber, S. Westermann, C. Wille, S. Wofsy, G. Wohlfahrt,

S. Wolf, W. Woodgate, Y. Li, R. Zampedri, J. Zhang, G. Zhou, D. Zona, D. Agarwal, S. Biraud, M. Torn, and D. Papale, "The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data," *Scientific data*, vol. 7, no. 1, p. 225, 2020.

[196] E. Daly, A. Porporato, and I. Rodriguez-Iturbe, "Coupled dynamics of photosynthesis, transpiration, and soil water balance. Part I: Upscaling from hourly to daily level," *Journal of Hydrometeorology*, vol. 5, no. 3, pp. 546–558, 2004.

[197] I. Rodriguez-Iturbe, A. Porporato, L. Rldolfi, V. Isham, and D. R. Cox, "Probabilistic modelling of water balance at a point: The role of climate, soil and vegetation," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 455, no. 1990, pp. 3789–3805, 1999.

[198] P. J. Wetzel and J.-T. Chang, "Concerning the Relationship between Evapotranspiration and Soil Moisture," *Journal of Applied Meteorology and Climatology*, vol. 26, no. 1, pp. 18–27, 1987.

[199] H. Haario, E. Saksman, and J. Tamminen, "An Adaptive Metropolis Algorithm," *Bernoulli*, vol. 7, p. 223, aug 2001.

# APPENDIX A

# BAYESIAN SCHEME FOR PARAMETER INFERENCE

For the SHM defined in Section 3.5, we perform the Bayesian parameter inference as follows. Let $\mathbf{HX}_{\boldsymbol{\eta}} = \tilde{\mathbf{X}}_{\boldsymbol{\eta}}$ and $\mathbf{HX}_{\beta} = \tilde{\mathbf{X}}_{\beta}$. We carry out a Gibbs sampler for all parameters in the model, by sampling from the following FCD.

## A.1 Update for $\beta$

Let $\mathbf{z}_{\boldsymbol{\beta}} = \mathbf{z} - \tilde{\mathbf{X}}_{\boldsymbol{\eta}}\boldsymbol{\eta}$. Then $[\mathbf{z}_{\boldsymbol{\beta}}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}] = \mathcal{N}_n(\mathbf{z}_{\boldsymbol{\beta}}|y, \Sigma)\mathcal{N}_n(y|\mu, C)d\mathrm{y}$. It can be shown that this is equal to $\mathcal{N}_n(\mathbf{z}_{\boldsymbol{\beta}}|\boldsymbol{\mu}_{\mathbf{z}_{\boldsymbol{\beta}}}, \boldsymbol{\Sigma}_{\mathbf{H}})$, where $\boldsymbol{\mu}_{\mathbf{z}_{\boldsymbol{\beta}}} = \mathbf{HX}_{\boldsymbol{\beta}}\boldsymbol{\beta} = \tilde{\mathbf{X}}_{\boldsymbol{\beta}}\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\mathbf{H}} = \mathbf{HCH'} + \Sigma$. If the prior distribution of $\boldsymbol{\beta}$ is assumed to be $N_{n_\beta}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$, then it can be shown that:

$$[\boldsymbol{\beta}|\mathbf{z}_{\boldsymbol{\beta}}, \boldsymbol{\gamma}, \boldsymbol{\eta}] = \mathcal{N}_{n_\beta}(\boldsymbol{\mu}_{\boldsymbol{\beta}_{\mathbf{FCD}}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{\mathbf{FCD}}}) \tag{A.1}$$

where $\boldsymbol{\mu}_{\boldsymbol{\beta}_{\mathbf{FCD}}} = (\boldsymbol{\Sigma}_{\boldsymbol{\beta}}{}^{-1} + \tilde{\mathbf{X}}'_{\boldsymbol{\beta}}\boldsymbol{\Sigma}_{\mathbf{H}}{}^{-1}\tilde{\mathbf{X}}_{\boldsymbol{\beta}})^{-1}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}{}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} + \tilde{\mathbf{X}}'_{\boldsymbol{\beta}}\boldsymbol{\Sigma}_{\mathbf{H}}{}^{-1}\mathbf{z}_{\boldsymbol{\beta}})$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_{\mathbf{FCD}}} = (\boldsymbol{\Sigma}_{\boldsymbol{\beta}}{}^{-1} + \tilde{\mathbf{X}}'_{\boldsymbol{\beta}}\boldsymbol{\Sigma}_{\mathbf{H}}{}^{-1}\tilde{\mathbf{X}}_{\boldsymbol{\beta}})^{-1}$.

## A.2 Update for $\boldsymbol{\eta}$

Again, putting $\mathbf{z}_{\boldsymbol{\eta}} = \mathbf{z} - \tilde{\mathbf{X}}_{\boldsymbol{\beta}}\boldsymbol{\beta}$ and for a prior distribution of $\boldsymbol{\eta}$ assumed to be $N_{n_\eta}(\boldsymbol{\mu}_{\boldsymbol{\eta}}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}})$, it can be shown that:

$$[\boldsymbol{\eta}|\mathbf{z}_{\boldsymbol{\eta}}, \boldsymbol{\beta}, \boldsymbol{\eta}] = \mathcal{N}_{n_\eta}(\boldsymbol{\mu}_{\boldsymbol{\eta}_{\mathbf{FCD}}}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}_{\mathbf{FCD}}}) \tag{A.2}$$

where $\boldsymbol{\mu}_{\boldsymbol{\eta}_{\mathbf{FCD}}} = (\boldsymbol{\Sigma}_{\boldsymbol{\eta}}{}^{-1} + \tilde{\mathbf{X}}'_{\eta}\boldsymbol{\Sigma}_{\mathbf{H}}{}^{-1}\tilde{\mathbf{X}}_{\eta})^{-1}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}{}^{-1}\boldsymbol{\mu}_{\boldsymbol{\eta}} + \tilde{\mathbf{X}}'_{\eta}\boldsymbol{\Sigma}_{\mathbf{H}}{}^{-1}\mathbf{z}_{\eta})$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_{\mathbf{FCD}}} = (\boldsymbol{\Sigma}_{\boldsymbol{\eta}}{}^{-1} + \tilde{\mathbf{X}}'_{\eta}\boldsymbol{\Sigma}_{\mathbf{H}}{}^{-1}\tilde{\mathbf{X}}_{\eta})^{-1}$.

## A.3 Update for $\gamma$

Unlike $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, closed form for FCD for $\boldsymbol{\gamma}$ cannot be derived. We therefore use a Metropolis Hastings update for $[\boldsymbol{\gamma}|\boldsymbol{\beta}, \boldsymbol{\eta}, \mathbf{z}]$ using a Gaussian proposal distribution. To ensure proper mixing, we adopt the Adaptive-Metropolis algorithm by [199] which adapts the covariance of the proposal distribution at any iteration $t$, $Cov_t^{prop}$ using past proposals after a burn-in period $t_0$ such that:

$$Cov_t^{prop} = \begin{cases} Cov_0^{prop} & \text{t} \leq t_0 \\ s_d cov(\gamma_0, ..., \gamma_{t-1}) + s_d \epsilon I_d & \text{t} > t_0 \end{cases} \tag{A.3}$$

where $Cov_0^{prop}$ is the covariance of the initial proposal distribution, $d$ is the number of parameters in $\boldsymbol{\gamma}$, $s_d$ is the scaling parameter equal to $(2.4)^2/d$, $\epsilon > 0$ is a constant chosen to ensure non-singularity of $Cov_t^{prop}$, $\mathbf{I_d}$ is the identity matrix. In the present study, we choose $\epsilon = 0.001$ and $t_0 = 1000$ for all the analyzed days.

# APPENDIX B

# DERIVATION OF *VECCHIA-MULTISCALE*

## B.1 Parameter estimation

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

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

$$\lambda_{ij} = -\frac{w_{i,j}}{\sigma^2_{z_i|z_{1:i-1}}} \tag{B.2}$$

where $w_{ij}$ equals $w_i = \Sigma^{-1}_{1:i-1} C(\boldsymbol{A_{1:i-1}}, A_i)$ for $j = 1, ..., i-1$, 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 $C(\boldsymbol{A_{1:i-1}}, A_i)_j \approx (h^{\kappa}_{A_j})^T C(\mathcal{G}_{A_j}, \mathcal{G}_{A_i}) h^{\kappa}_{A_i}$ and $C(\boldsymbol{A_{1:i-1}}, \boldsymbol{A_{1:i-1}})_{jk} (h^{\kappa}_{A_j})^T C(\mathcal{G}_{A_j}, \mathcal{G}_{A_k}) h^{\kappa}_{A_k}$ for $j, k = 1, ..., i-1$ 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 4.7.

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

If $\mu(A_i) = X(A_i)^T \beta$ where $X(A_i) = \{X^1(A_i), ..., X^p(A_i)\}$ is a vector of covariates of length $p$ for pixel $A_i$. Then $\mu_{A_i} = X_{A_i}\beta$ in equation 4.9 where $X_{A_i}$ is the matrix of covariates associated with the points associated with $y_{A_i}$ in equation 4.7. Then $\mu_z$ (equation 4.8) can be written as $\tilde{X}\tilde{\beta}$ where the $i^{th}$ row of $\tilde{X}$ is given as $\{h^{\kappa}_{A_i} X^1_{A_i}, ..., h^{\kappa}_{A_i} X^p_{A_i}, \delta(A_i)\}$. The parameter vector $\tilde{\beta}$ can be profiled out by using the profile-likelihood:

$$-2log(\hat{f}(z(\mathcal{A})|\theta) = -2log(det(\hat{\Lambda})) + (\hat{\Lambda}(z - \tilde{X}\tilde{\beta}))^T\hat{\Lambda}(z - \tilde{X}\tilde{\beta})) + nlog(2\pi) \tag{B.3}$$

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

$$\tilde{\beta}_{MLE} = [(\hat{\Lambda}\tilde{X})^T(\hat{\Lambda}\tilde{X})]^{-1}(\hat{\Lambda}\tilde{X})^T(\hat{\Lambda}z) \tag{B.4}$$

## B.2 Prediction Algorithm

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

$$\begin{aligned} E[y^{pred}|z] &= X^{pred}\hat{\beta} + \Sigma_{21}^{comp}(\Sigma_{11}^{comp})^{-1}(z - X\beta) \\ &= -(\Lambda_{22}^{comp})^{-1}\Lambda_{21}^{comp}z \approx -(\hat{\Lambda}_{22}^{comp})^{-1}\hat{\Lambda}_{21}^{comp}(z - X\beta) \end{aligned} \tag{B.5}$$

where $\hat{\Lambda}^{comp}$ is the sparse approximation of $\Lambda^{comp}$ calculated following Section B.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, $-\Lambda_{22}^{-1}\Lambda_{21}(z-z^*)+y^{pred*}$ 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 $-\Lambda_{22}^{-1}\Lambda_{21}(z - z^*) + y^{pred*}$ five thousand times and approximate the prediction variance.

# APPENDIX C

# DIFFERENT PERMUTATIONS FOR *VECCHIA-MULTISCALE*, SIMULATION RESULTS AND SUPPORTING INFORMATION

## C.1 Illustration of different permutations for Vecchia-Multiscale

We illustrate the effect of different permutations (Figure C.1 and C.2) by applying the eight permutations to the hypothetical example in Figure 4.2 (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 C.1 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 C.1 (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 C.1 (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 C.1 (h)).

The *Joint-Maxmin* ordering (Figure C.1 (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 C.1 (i)-(k)). The subvector $\boldsymbol{A_{m_i}}$ now consist of both far and near data surrounding $A_i$ (Figure C.1 (l)). Though *Joint-Random* (Figure C.1 (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 C.2. 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 4.4). 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 C.2 (Column I) since we have 40 point data, $\{A_1, \ldots, A_{40}\}$ always represent point data in "Separate-" permutations. Now for any areal pixel $A_i$ (which for "Separate-" permutations in this example represent $\{A_1, \ldots, A_{140}\}$), $\boldsymbol{A_{1:i-1}}$ 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 C.2, Column IV).

## C.2 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 resulting out of the eight permutations better approximates the exact likelihood. Similar to the hypothetical example in Figure 4.2 (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 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 C.1. 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 $\mathcal{D}$.

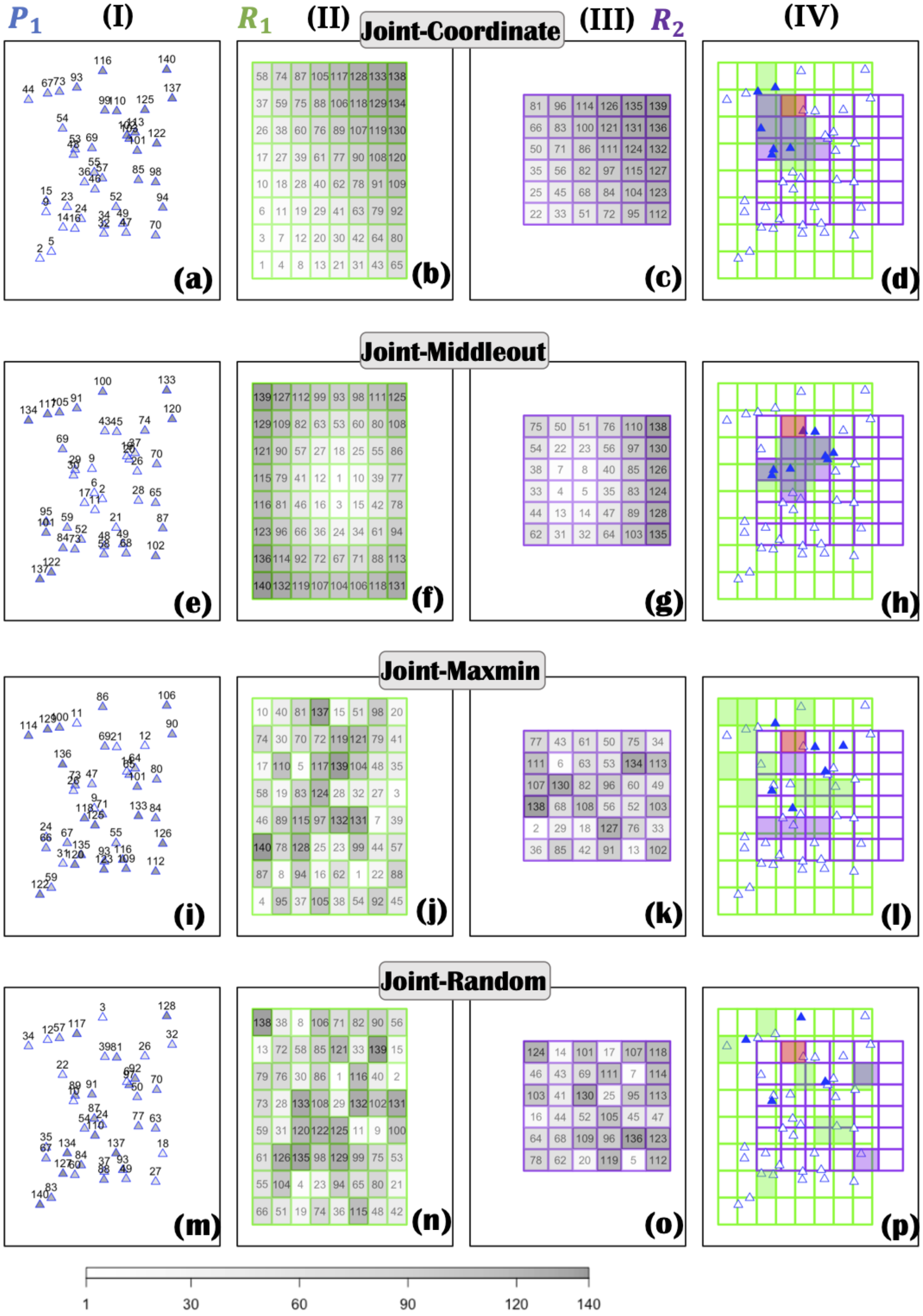

Figure C.1: Illustration of the "Joint-" Permutations applied on the example from Figure 4.2 (a) in the main text consisting of 40 point data $P_1$ and 100 areal pixels in $R_1$ (64 pixels) and $R_2$ (36 pixels). Numbers in columns (I) to (III) represent the ordering number in the vector $\mathcal{A} = \{A_1, \ldots, A_{140}\}$ 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 $A_{m_i}$ (equation 4.11) comprising color-filled blue triangles, and color-filled green and purple pixels, for a randomly chosen pixel $A_i$ (color-filled red) for $m = 20$.

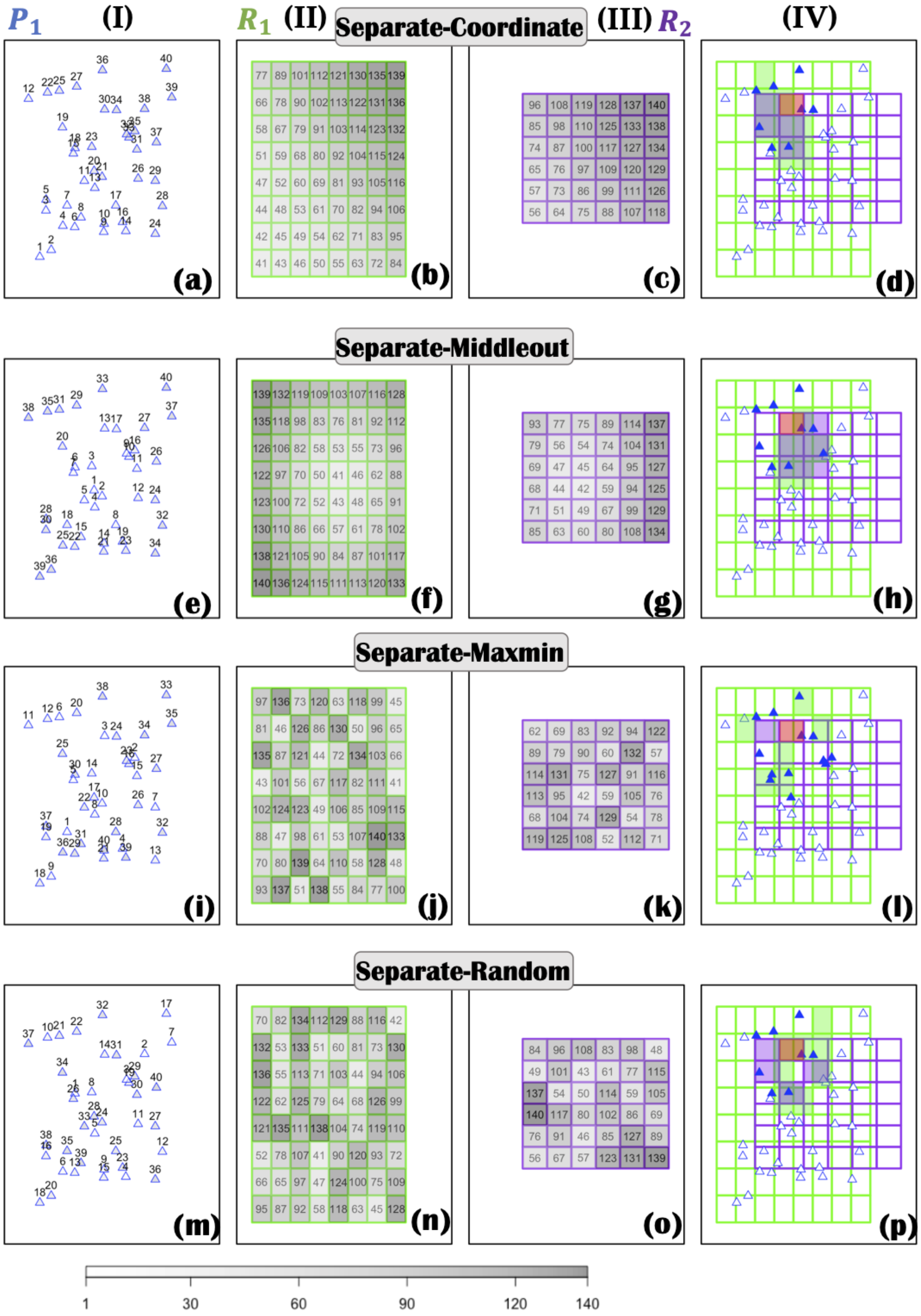

Figure C.2: Illustration of the “Separate-“ Permutations applied on the example from Figure 4.2 (a) in the main text consisting of 40 point data $P_1$ and 100 areal pixels in$R_1$ (64 pixels) and $R_2$ (36 pixels). Numbers in columns (I) to (III) represent the ordering number in the vector $\mathcal{A} = \{A_1, \ldots, A_{140}\}$ 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 $A_{m_i}$ (equation 4.11) comprising color-filled blue triangles, and color-filled green and purple pixels, for a randomly chosen pixel $A_i$ (color-filled red) for $m = 20$.

As mentioned in Section $4.4.2$, 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 $\hat{f}(z(\mathcal{A})\theta)$ over the exact likelihood $f(z(\mathcal{A})|\theta)$, 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 C.3. 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.

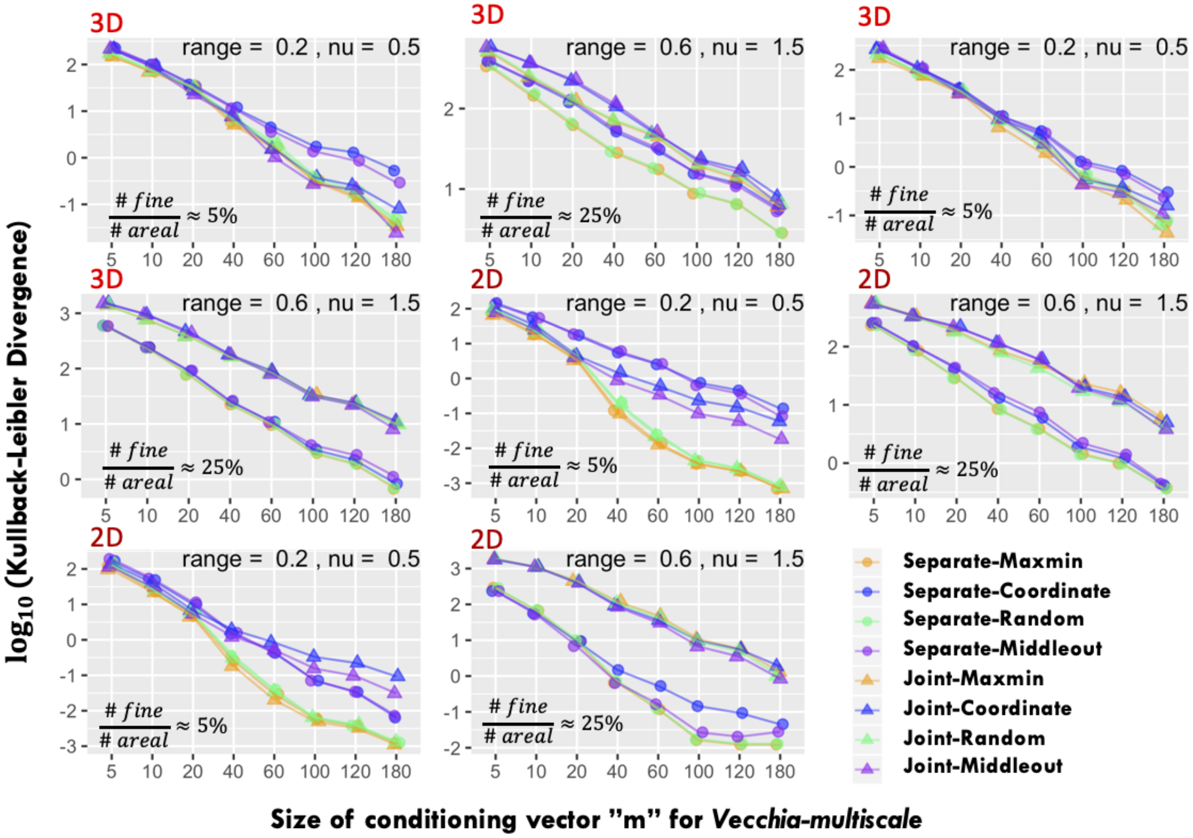

Figure C.3: Representative simulations comparing the (log) KL-Divergence of the approximations over the true likelihood for measurement error variance equal to 0.05. A lower KL-Divergence denotes a better approximation. For the majority of the simulation settings, the *Separate-Maxmin* and the *Separate-Random* lead to better approximation of the exact likelihood.

Table C.1: Data setting for simulations

| 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 | - |

## C.3 Supporting Information for Chapter 4

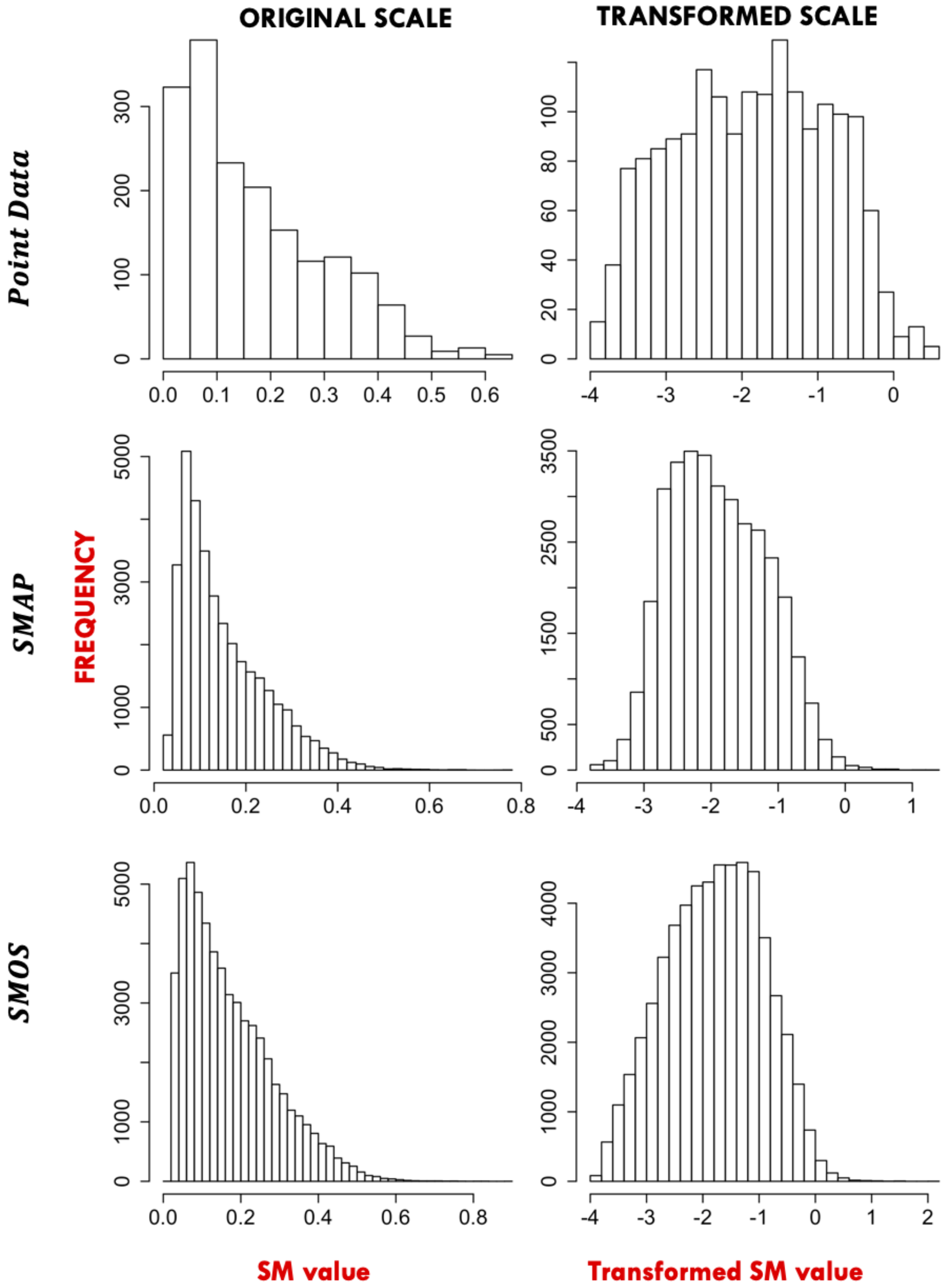

Figure C.4: Histograms of point soil, SMAP and SMOS soil moisture data for July 06-20, 2017. On the original scale soil moisture exhibits considerable skewness but on the logit scale the soil moisture distribution becomes less skewed making the Gaussian assumption tenable.

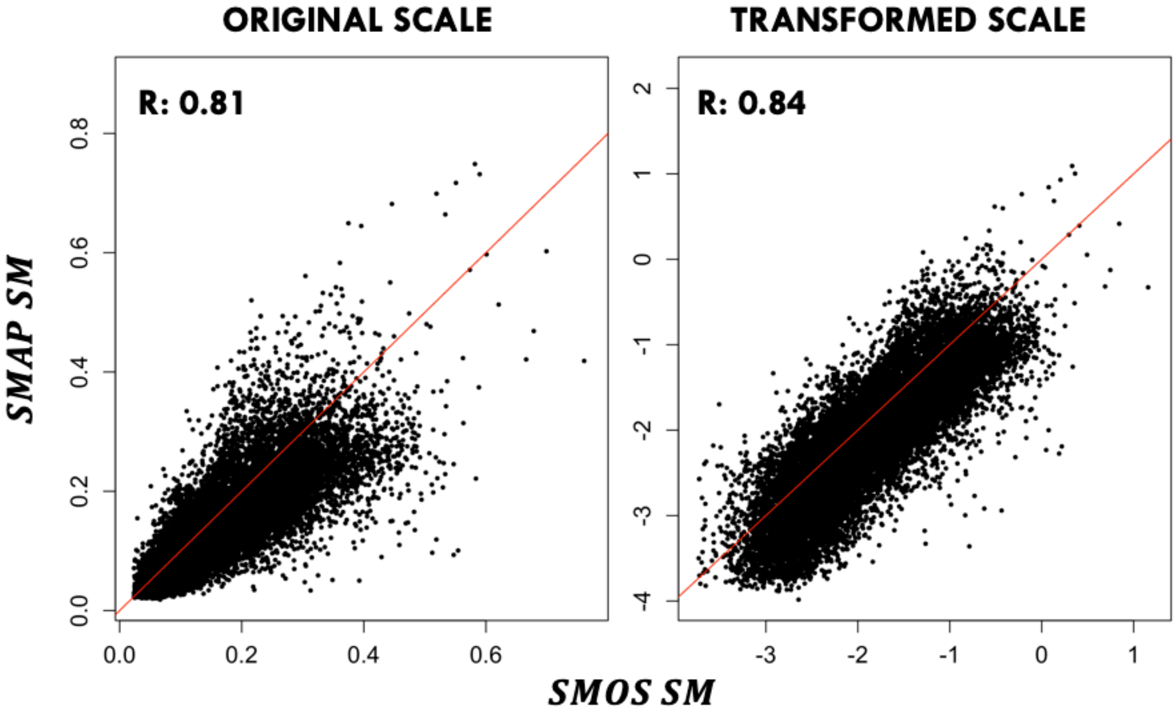

Figure C.5: Overlapping SMOS and SMAP pixels for July 06-20, 2017. The SMOS pixels are bilinearly interpolated to the overlapping SMAP pixels for this exploratory analysis. The red line denotes the 1:1 line. The transformed scale results in a slightly better correlation (R) between the two datasets. On the transformed scale, it can also be seen that there is a bias between SMOS and SMAP datasets for the analyzed time period.

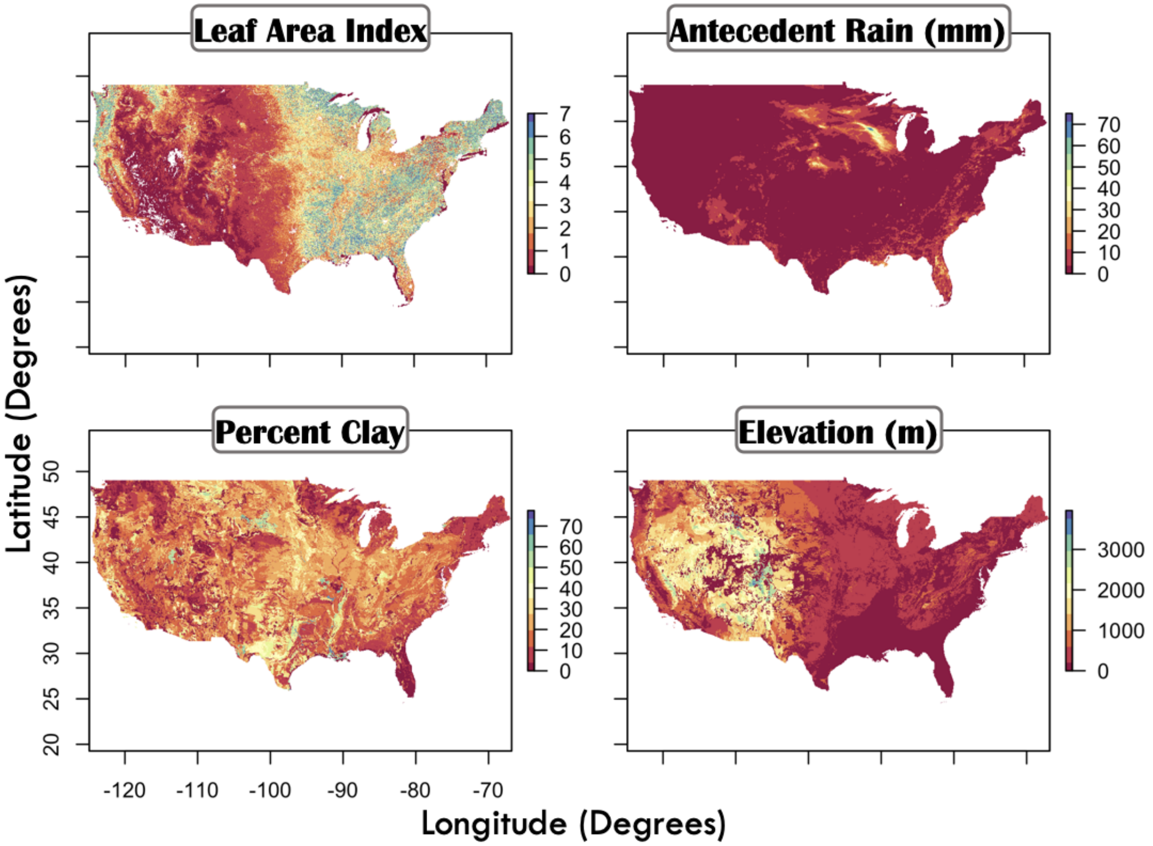

Figure C.6: Covariate plots for July 06, 2020 for Contiguous US (CONUS). All the four covariates exhibit considerable heterogeneity across CONUS.