A Bayesian Machine Learning Method to Explain the Error Characteristics of Global-Scale Soil Moisture Products

Hyunglok Kim[1,2*], Wade T. Crow[1], Wolfgang Wagner[3], Xiaojun Li[4], Venkataraman Lakshmi[5]

[1]United States Department of Agriculture - Agricultural Research Service, Hydrology and Remote Sensing Laboratory, Beltsville, MD 20705, USA

[2]School of Earth Sciences and Environmental Engineering, Gwangju Institute of Science and Technology, Gwangju, South Korea

[3]TU Wien (Technische Universität Wien), Department of Geodesy and Geoinformation, Wien, Austria

[4]INRAE, UMR1391 ISPA, Université de Bordeaux, France

[5]The Department of Engineering Systems and Environment, University of Virginia; Charlottesville, VA, 22904, USA.

*Correspondence to: hyunglokkim@gmail.com

June 29, 2023

Abstract

Estimating accurate surface soil moisture (SM) dynamics from space, and knowing the error characteristics of these estimates, is of great importance for the application of satellite-based SM data throughout many Earth Science/Environmental Engineering disciplines. Here, we introduce the Bayesian inference approach to analyze the error characteristics of widely used passive and active microwave satellite-derived SM data sets, at different overpass times, acquired from the Soil Moisture Active Passive (SMAP), Soil Moisture and Ocean Salinity (SMOS), and Advanced Scatterometer (ASCAT) missions. In particular, we apply Bayesian hierarchical modeling (BHM) and triple collocation analysis (TCA) to investigate the relative importance of different environmental factors and human activities on the accuracy of satellite-based data.

To start, we compare the BHM-based sensitivity analysis method to the classic multiple regression models using a frequentist approach, which includes complete pooling and no-pooling models that have been widely used for sensitivity analysis in the field of remote sensing and demonstrate the BHM's adaptability and great potential for providing insight into sensitivity analysis that can be used by various remote sensing research communities.

Next, we conduct an uncertainty analysis on BHM's model parameters using a full range of uncertainties to assess the association of various environmental factors with the accuracy of satellite-derived SM data. We focus on investigating human-induced error sources such as disturbed surface soil layers caused by irrigation activities on microwave satellite systems, naturally introduced error sources such as vegetation and soil organic matter, and errors related to the disregard of SM retrieval algorithmic assumptions - such as the thermal equilibrium passive microwave systems. Based on the BHM-based sensitivity analysis, we find that assessments of

SM data quality with single variable should be avoided, since numerous other factors simultaneously influence their quality. As such, this provides a useful framework for applying Bayesian theory to the investigation of the error characteristics of satellite-based SM data and other time-varying geophysical variables.

Keywords:

microwave satellite systems, remotely sensed soil moisture, Bayesian hierarchical model, triple collocation analysis, uncertainty analysis

1. Introduction

Since surface soil moisture (SM) controls the flow of water and energy and governs interactions between the land surface and atmosphere, obtaining accurate surface-level SM information is critical for understanding many Earth system processes (Hirschi *et al* 2014, Seneviratne *et al* 2006). Likewise, understanding the accuracy of SM data is essential for applying SM data to numerous research fields, such as predicting hydrologic extremes (e.g., droughts, floods, wildfires, and dust outbreaks), estimating water resources, and improving land surface models (LSMs) (Brocca *et al* 2019, Crow *et al* 2022, Reichle 2008).

Among the methods used to estimate surface SM - including, but not limited to, gravimetric sampling (Reynolds 1970), hand-held/in-situ electromagnetic sensors (Kim *et al* 2020a), and cosmic-ray neutron probes (Nguyen *et al* 2017) - microwave satellite systems are generally considered to be the most practical for obtaining temporal and spatial continuous SM data at large spatial scales (Cho *et al* 2017, Entekhabi *et al* 2010, Jackson *et al* 1996, Wagner *et al* 2007, Wigneron *et al* 2017, Kim and Lakshmi 2018). Such systems include passive microwave instruments such as the L-band radiometer on board the Soil Moisture Active Passive (SMAP) (Entekhabi *et al* 2010), the Microwave Imaging Radiometer with Aperture Synthesis (MIRAS) on board the Soil Moisture and Ocean Salinity (SMOS) (Kerr *et al* 2010), and the active microwave sensor Advanced Scatterometer (ASCAT) on board the MetOp-A (de-orbited in November 2021 after 15 years of service), MetOp-B, and MetOp-C satellites (Wagner *et al* 2013).

However, despite researchers' best efforts to obtain reliable SM information from satellite systems, we still encounter significant environmental/human-induced factors that decrease the quality of SM retrievals. For example, satellite-based SM data are vulnerable to

errors from sources such as dense vegetation canopy (Calvet *et al* 2011, Owe *et al* 2001), arid climatic conditions (Wagner *et al* 2022), radio frequency interference (RFI) (de Nijs *et al* 2015, Misra and Ruf 2012, Oliva *et al* 2012), soil properties which have been disturbed by irrigation activities (He *et al* 2021, Lawston *et al* 2017), and high amounts of soil organic matter (SOM) (Wigneron *et al* 2017). Although many previous studies have identified error sources that negatively impact SM data quality, there has been little consideration regarding the relative importance of these error sources, including human-created and environmental factors, in inferring the overall quality of SM data. Ideally, if we can identify robust relationships between the error variance of satellite-based SM and a given environmental condition, we can also use SM data more effectively.

Errors in satellite SM data are dependent upon the exact retrieval algorithm used and/or the satellite systems themselves. Therefore, knowing the relative accuracy of each SM data product is essential for making the best use of satellite-based SM retrievals – particularly in the common case where SM information is integrated from multiple sources.

Here, we seek to develop improved regression models to explore the relationship between various hydrogeological variables and the precision of satellite-based SM products across different land surface characteristics. To investigate a global-scale individual satellite-based SM retrieval's precision, we employ triple collocation analysis (TCA; see the methodology section for details). By building a land cover-specific hierarchical model based on a Bayesian approach, we seek to provide an enhanced description of the relationship between SM retrievals errors, as described by TCA and key environmental variables.

Traditional approaches to understanding relationships between independent and dependent variables often involve creating multiple linear regression models. These models, while easy to interpret, may not always produce reliable results due to factors such as inadequate data points for specific land surface types. Additionally, frequently used model parameter estimation methods like maximum likelihood estimation (MLE) come with limitations. For example, they assume fixed, unknown true parameter values and may fail to account for prior parameter information or fully capture uncertainty.

Given these challenges, there is a need for more flexible and robust modeling methods. Bayesian hierarchical modeling (BHM) can serve as such an approach (Wagenmakers, et al. 2008). It allows for the integration of prior knowledge about parameters, more precisely quantifies uncertainty, and adapts well to different data structures, making it particularly beneficial for datasets with naturally clustered observational units, such as different land cover types. The present research uses BHM to analyze error in satellite SM data and assess the relative importance of environmental factors on SM data quality, taking into account inherent uncertainties. In this study, by using a Bayesian approach, we aim to achieve more rigorous and reliable scientific inferences. This study distinguishes itself from current regression model approaches, which frequently result in overfitting or offer only a single fixed parameter for each environmental factor that influences SM data quality.

2. Data sets

Here, we focus on evaluating the error characteristics of three satellite-based SM data sets (using the most recent version of each data; last checked date October 2022): the dual-

114 channel algorithm (DCA) based SMAP L3 Version 8 SM product (O'Neill et al. 2021); the INRA-
115 CESBIO (IC) algorithm version 2 based SMOS SM; and the ASCAT SM product based on the TU
116 Wien algorithm (Wagner *et al* 2013). Please note that when investigating the baseline quality of
117 each SM product, we included all data without consideration of their data-quality flags. However,
118 we did mask areas where the RFI flag values for the SMOS-IC product were larger than 5 K (please
119 note that the criteria for the RFI flag may vary in future versions of the SMOS-IC algorithm). In
120 addition, we investigated the error characteristics of different overpass times for each product:
121 ascending, descending, and the combination of both.

122 To build the BHM, we used the fractional mean square error (*fMSE*) metric as the models'
123 response variable (i.e., dependent or target variable) calculated from TCA (please refer to the
124 methodology section below). The predictor variables (i.e., independent or feature variables)
125 came from various sources. To start, we used 21 daily time series of hydrometeorological and
126 radiation variables from the North American Land Data Assimilation System, version 2 (NLDAS-2)
127 (Xia *et al* 2012) and the Modern-Era Retrospective analysis for Research and Applications, version
128 2 (MERRA-2) (Gelaro *et al* 2017): time averaged, minimum, and maximum values of near-surface
129 wind speed (m/s), average rainfall rate (kg/m^{2}s), total precipitation rate (kg/m^{2}s), near-surface
130 air temperature (K), near-surface specific humidity (-), surface-incident shortwave radiation
131 (W/m^2), and surface-incident longwave radiation (W/m^2). In addition, the daily difference
132 between 2-m air temperature (AT) and soil temperature (ST) (surface level 1) ($|\Delta T(ST, AT)|$) (6
133 a.m. and 6 p.m.) was calculated from ERA5-Land global reanalysis data (Muñoz-Sabater *et al* 2021)
134 from 2015 to 2021, and the time averaged, minimum, and maximum values of ($|\Delta T(ST, AT)|$)
135 were computed. Please note that for the North American domain, we utilized NLDAS-2, while for

136 other regions, we used MERRA-2. This approach allowed us to ensure the highest quality data for
137 both areas.

138 Second, we simulated 24 daily hydrological variables using the Noah-
139 Multiparameterization LSM version 4.0.1 (Niu *et al* 2011): time averaged, minimum, and
140 maximum values of the latent heat flux (W/m^2), sensible heat flux (W/m^2), total
141 evapotranspiration (kg/m^2s), average surface temperature (K), surface radiative temperature (K),
142 soil temperature (K); LAI (-), and greenness (-). All the data sets were processed using the Land
143 Surface data Toolkit (LDT) (Arsenault *et al* 2018) and the Land Information System (LIS) (Kumar
144 *et al* 2006). The LDT and LIS are open-source tools and software libraries developed and
145 maintained by NASA for managing and analyzing remotely sensed and land surface data. They
146 are widely used in the field of Earth science, offering a comprehensive solution for global-scale
147 data analysis. The user guide and tutorials for this software are publicly available on a GitHub
148 page, which is noted in the Acknowledgements section. We intentionally omitted ERA5-Land and
149 SMAP L4 SM data products from our analysis to maximize the independence of our predictor
150 variables relative to response variables - note that both ERA5-land and SMAP L4 integrate some
151 form of satellite-based SM information. Instead, we opted to use an open loop simulation of the
152 Noah-MP land surface model (Noah-MP4.0.1) lacking any data assimilation.

153 Third, we considered seven additional static variables: 1) topographic complexity (i.e.,
154 proxy for surface roughness) by taking the logarithm of the digital elevation model (DEM) data
155 obtained from the Shuttle Radar Topography Mission (SRTM), as described by Kim et al. (2015);
156 2) the diversity index (or Gini-Simpson index) (-) using the International Geosphere–Biosphere
157 Programme (IGBP) from the National Centers for Environmental Prediction (NCEP) land

158 classification map (17 IGBP data and three tundra landcover classes) (**Fig. 1(a)**); 3) the irrigation
159 fraction (%) from the Global Map of Irrigation Areas (GMIA) (Siebert et al., 2005; **Fig. 1(b)**); 4)
160 SOM from the International Soil Reference and Information Centre (ISRIC) (**Fig. 1(c)**); 5)
161 vegetation opacity (-) (or Tau) based on the SMAP Multi-Temporal Dual Channel Algorithm (MT-
162 DCA) (Konings *et al* 2017) (**Fig. 1(d)**); 6) the sand fraction (%) from the STATSGO-FAO soil texture
163 class map; 7) the slope from the SRTM DEM (%); 8) the average surface albedo based on Wang
164 et al.'s (2004) method, which uses data from the Moderate Resolution Imaging
165 Spectroradiometer (MODIS). In addition, the daily time averaged, minimum, and maximum
166 values of the brightness temperature (Tb)-RFI flag (K) from SMOS-IC data (a.m. and p.m.) (**Fig.**
167 **1(e)**) were collected. The RFI flag is represented by the Tb-RMSE in Kelvin (K), which is the root
168 mean square error (RMSE) value between the L-band Microwave Emission of the Biosphere (L-
169 MEB) model Tb and the measured Tb data. Wigneron et al. (2021) demonstrated that the TB-
170 RMSE is a simple and effective indicator of the actual RFI impact. In total, 67 predictors (**Table S1**)
171 were initially considered as predictor variables, and all variables were normalized using their
172 mean and standard deviation values.

173 Multicollinearity between predictor variables can result in less reliable statistical
174 inferences. Therefore, since several of the independent variables described above are likely to
175 be mutually correlated, we conducted a multicollinearity test before continuing. This test was
176 based on applying a variance inflation factor (VIF). We selected the 14 predictors whose VIF was
177 below 12 (**Table 1**). However, we included soil temperature and precipitation variables even
178 though their VIF were greater than 12, as they were the two most important hydrometeorological

variables of interest (note that high VIF for these two variables was not unexpected since some of our predictors were calculated from LSM).

The variables listed in **Table 1** have the potential to directly and/or indirectly impact the quality of SM data retrieved from both passive and active systems. For example, topographic complexity can serve as a proxy for surface roughness, which is a critical factor in retrieving SM data from the tau-omega model (Li *et al* 2020). Additionally, high sand fractions can impede the retrieval of SM by both passive and active satellite systems due to the subsurface scattering of microwave signals (Kim *et al* 2018). Furthermore, algorithms that assume a static state and a constant vertical SM distribution for L-band microwave radiometer-based SM systems can be adversely affected during and immediately after precipitation events due to the transient movement of water in the shallow subsurface. Additionally, during periods of heavy rainfall, both naturally emitted microwave signals observed by passive microwave sensors and microwave signals generated from active satellite systems will be affected by the presence of hydrometeors (Colliander *et al* 2020). Finally, RFI has a direct impact on L-band SM retrievals (Oliva *et al* 2012). The aim of this study is to examine the influence of these variables on the quality of SM data retrieved from both passive and active satellite systems.

All data were resampled into the Equal-Area Scalable Earth (EASE) grids (36-km × 36-km). In summary, a total of 100,766 data points and 9 commonly available land cover types were available in total. Note that we were forced to restrict our analysis to only 9 commonly available land cover types for three satellite-based SM products due to missing TCA values (i.e., the response variable) and resulting inadequate coverage of certain land cover types. This resulting

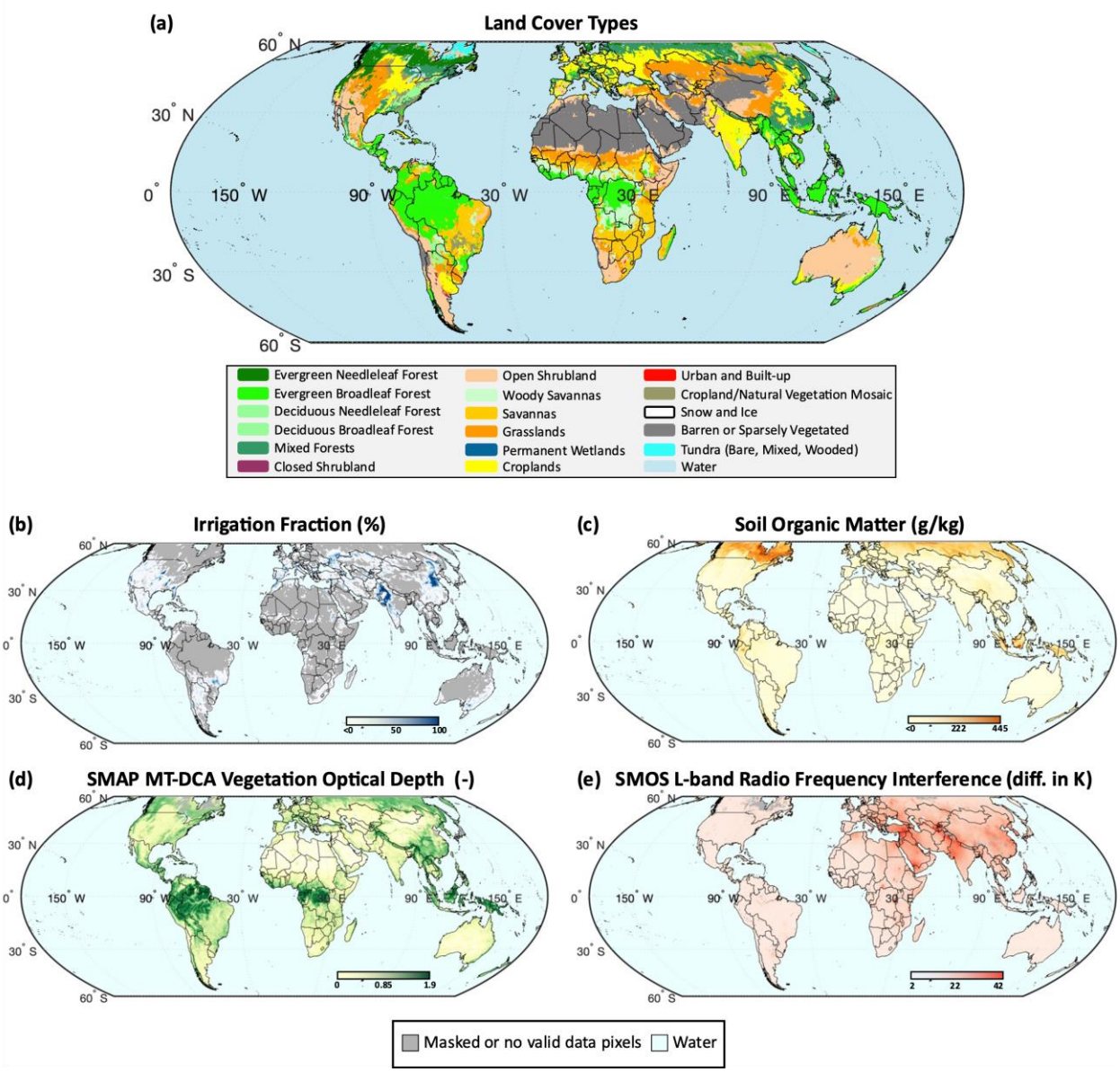
200 in a 100,766 × 14 predictor matrix with one categorical variable (i.e., 9 land cover types) for the

201 generation of a particular BHM.

202 **Table 1**. 14 Selected variables for the model development.

Variable Name (unit)	Factor	Data Source
Diversity (Gini-Simpson) Index (-)	Static	Calculated from the IGBP land classification data
Irrigation Fraction (-)	Static	Global Map of Irrigation Areas (GMIA)
Sand Fraction (-)	Static	STATSGO-FAO soil texture class map
Soil Organic Matter (g/kg)	Static	International Soil Reference and Information Centre (ISRIC)
Slope (-)	Static	Shuttle Radar Topography Mission (SRTM)
Topographic complexity (log(m))	Static	
SMAP Vegetation Opacity (Tau) (-)	Dynamic	SMAP (MT-DCA)
SMOS-IC Radio Frequency Interference (K) (a.m. and p.m.)	Dynamic	SMOS-IC (Version 2)
Soil Temperature (K)	Dynamic	NoahMP4.0.1
Sensible Heat Flux (W/m2)	Dynamic	
Surface Albedo (-)	Dynamic	Moderate Resolution Imaging Spectroradiometer (MODIS)
Total Precipitation (kg/m2s)	Dynamic	North American Land Data Assimilation System, phase 2 (NLDAS-2)
Near Surface Specific Humidity (-)	Dynamic	Modern-Era Retrospective analysis for Research and Applications, version 2(MERRA-2)
$\Delta T(AT, ST)$ (a.m. and p.m.)	Dynamic	ERA5-land

203



204

205 **Figure 1**. Maps of: (a) the 18-member land cover classification, (b) the irrigation fraction from

206 GMIA, (c) soil organic matter (SOM) from ISRIC, (d) vegetation optical depth (VOD) from SMAP

207 MT-DCA, and (e) radio frequency interference (RFI) RMSE from SMOS-IC (version 2).

3 Methods

3.1 Calculating the Error of satellite-based SM Data

To evaluate the uncertainties in satellite-based SM data sets on a global scale, we employed TCA to estimate the total random error (ε) variance of time-varying geophysical data. Among possible TCA-based error statistics, we selected the fractional mean square error (*fMSE*) as it provides straightforward insight on the precision of the data. Specifically, *fMSE* ranges from 0 to 1, whereas a value of lower than 0.5 indicates that the true SM signal is a larger component to the data than estimation noise.

The TCA-based error variance of individual satellite-based SM retrievals ($\sigma_{\varepsilon_i}^2$) and the variance of the individual data itself, σ_i^2, are used to calculate *fMSE*:

$$fMSE_i = \frac{\sigma_{\varepsilon_i}^2}{\sigma_i^2} = \frac{\sigma_{\varepsilon_i}^2}{\beta_i^2 \sigma_\Theta^2 + \sigma_{\varepsilon_i}^2} = \frac{1}{SNR + 1} \quad Eq.(1)$$

Where σ_Θ^2 is the variance of the true jointly observed SM signal; $fMSE_i$ is a normalized representation of the signal-to-noise ratio (SNR) ($\frac{\beta_i^2 \sigma_\Theta^2}{\sigma_{\varepsilon_i}^2}$). After removing the climatology of SM and under the TCA assumptions of error orthogonality and zero error-cross correlation (Gruber et al., 2016), $\sigma_{\varepsilon_i}^2$ can be calculated from:

$$\sigma_{\varepsilon_i}^2 = \sigma_i^2 - \frac{\sigma_{ij}\sigma_{ik}}{\sigma_{jk}} \quad Eq.(2)$$

Where *j* and *k* indicate other individual satellite-based SM retrievals, and σ_{xy} ($x \in \{i, j, k\}$, $y \in \{i, j, k\}$, and $x \neq y$) are the covariance between two different satellite-based SM retrievals. Here, we followed the calculation of the ensemble *fMSE* shown in Kim et al. (2021) using the most recent version of each SM data source systematically organized into different triplets. In

228 composing individual triplets, we emphasized combinations of a passive system, an active system,
229 and a model-based SM product to maximize the likelihood that each of the three products
230 contain mutually independent errors. We then calculated the ensemble *fMSE* of SMAP SM data
231 from the following triplets: 1) SMAP-ASCAT-model-based SM, 2) SMAP-AMSR2-model-based SM,
232 and 3) SMAP-ASCAT-AMSR2 - where model-based SM data (0 - 10 cm) was acquired from Global
233 Land Data Assimilation System Version 2 (hereafter GLDAS) (Rodell *et al* 2004). Likewise, for the
234 SMOS-IC ensemble *fMSE* calculation, the following triplets were used: 1) SMOS-ASCAT-GLDAS, 2)
235 SMOS-ASCAT-AMSR2, and SMOS-ASCAT-GLDAS. Finally, for the ASCAT ensemble *fMSE*
236 calculation, we used five additional triplets: 1) ASCAT-SMAP-GLDAS, 2) ASCAT-SMOS-GLDAS, 3)
237 ASCAT-AMSR2-GLDAS, 4) ASCAT-AMSR2-SMAP, and 5) ASCAT-AMSR2-SMOS. If the standard
238 deviation of *fMSE* of a given product, obtained across the set of triplets defined above, is larger
239 than 0.1, we discarded those *fMSE* values and assumed that they are biased due to the neglect
240 of one or more TCA assumptions. Please refer to Kim et al. (2021) for further details regarding
241 the calculation of the ensemble *fMSE*.
242
243 **3.2. Regression Model and Hierarchical Model Structures**
244 Our data structure has J groups indexed as j = 1, …, J=9 (i.e., 9 commonly available land
245 cover types), we have n observations of the response variable y_i ($fMSE_i$), i = 1, …, n with k
246 predictors in an n x k matrix $\boldsymbol{X}$. Let X_i be the i^{th} row of $\boldsymbol{X}$. Here, we have three commonly used
247 regression model structures for the standard regression models with a dummy variable (i.e., land
248 cover types): 1) a complete pooling model, 2) a no-pooling model, and 3) a partial pooling model.
249 However, each of these approaches has well-known limitations. For example, the complete

pooling model cannot provide groupwise error estimates and the no-pooling model gives poor and possibly extreme estimates for groups having a small sample size. Additional details regarding other regression model types are included in the supplementary material document.

Hierarchical linear modeling (HLM) is a special form of multiple linear regression used to analyze variances in outcome variables when the predictor variables are obtained from different groups. HLM uses available information in the data, i.e., the predictor variables, to better predict the group target or response values - even in small groups. A basic hierarchical model with varying intercepts and varying slope is given below:

$$y_i = \alpha_{j[i]} + \beta_{j[i]}\boldsymbol{X}_i + \epsilon_i \quad Eq.(3)$$

Where y_i is *fMSE* calculated from **Eq. (1)** ($i \in \{1, \dots, N\}$), $\boldsymbol{X}_i$ is the 14 predictors in **Table 1**, and $\alpha_{j[i]}$ and $\beta_{j[i]}$ are the parameters for each land cover type (*j*), and ϵ_i is normally distributed with mean 0 and variance of σ_ϵ^2:$\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. Please note that, we made the assumption that the error terms (ϵ_i) are independent and identically distributed (i.i.d). Specifically, these errors represent the discrepancies between our model's predictions and the actual values, and we assume these discrepancies are random, have a constant variance (σ^2), and are not correlated with each other or with the predictors in our model. The sign of β_j is essential for understanding the relative significance of every predictor in predicting *fMSE*. A large positive or negative β_j value for a predictor suggests a stronger association with the *fMSE* value, while a predictor with a value closer to zero is less strongly associated. Analyzing the β_j allows us to discern the significant variables that play a crucial role in impacting the accuracy of satellite-based SM data sets. In this study, HLM offers insights into the association between variables and the precision of satellite-based SM data across various land cover types. While insightful, HLM with MLE, like

classic regression models, does not account for parameter estimate uncertainty. Tools such as bootstrapping may estimate this uncertainty but are not universally applicable (Wagenmakers et al., 2008). To bridge this gap, we have employed Bayesian inference. This probabilistic model allows us to recover the full range of inferential solutions, contrasting with the singular deterministic estimate of classical regression. Rather than acquiring a single estimate for the model parameters (i.e., α and β), we propose that each model's parameters are drawn from a probability distribution using the Bayesian approach. This stands in contrast to ordinary least-squares regression, which only minimizes the residual sum of squares.

In this study, HLM offers insights into the association between variables and the precision of satellite-based SM data across various land cover types. While insightful, HLM with MLE, like classic regression models, does not account for parameter estimate uncertainty. Tools such as bootstrapping may estimate this uncertainty but are not universally applicable (Wagenmakers et al., 2008). To bridge this gap, we have employed Bayesian inference. This probabilistic model allows us to recover the full range of inferential solutions, contrasting with the singular deterministic estimate of classical regression. In other words, rather than acquiring a single estimate for the model parameters (e.g., $\beta_{j[i]}$), we can draw each model's parameters from a probability distribution using the Bayesian approach which enables us to estimate an unobserved population of parameters conditioned on the training inputs and outputs.

To sum up, our BHM approach, applied to **Eq. (3)**, presents a solution to the limitations of standard pooling or non-pooling models associated with the frequentist approach. By using **Eq. (4)**, we can infer the probability distribution of β_j in **Eq. (3)** from the underlying population of *fMSE* and ***X***. With credible intervals, we can then make reliable inferences about the relationship

between the hydrometeorological variables listed in **Table 1** and the accuracy of satellite-based SM data sets.

In addition, the Box-Cox transformation was applied to the independent variable *fMSE*, yielding $fMSE^{(\lambda)}$. This transformation reduced the skewness of the distribution, allowing $fMSE^{(\lambda)}$ to be better approximated by a normal distribution. The Box-Cox parameter lambda, λ, was estimated by minimizing a sum-of-squares misfit (**Fig. S1**).

The normal distribution for $fMSE^{(\lambda)}$ is characterized by location (μ) and scale (σ) parameters. The probability density function (PDF) of the univariate normal distribution is as follows:

$$fMSE_i^{(\lambda)} \sim N(\mu_i, \sigma^2) \quad \boldsymbol{Eq.(4-1)}$$

$$\mu_i = \alpha_{j[i]} + \beta_{j[i]}\boldsymbol{X}_{\mathrm{i}} \quad \boldsymbol{Eq.(4-2)}$$

where, σ is the standard deviation of the measurement error ϵ_i. The group-level random effect α_j and the group-specific coefficients β_j are assumed to each follow their own multivariate normal distribution. That is, the α_j for all groups form a multivariate normal distribution with a certain mean vector and covariance matrix, and similarly the β_j for all groups form another multivariate normal distribution (**Eq. (5)**) with its own mean vector and covariance matrix.

$$P(\boldsymbol{\beta}|\,\boldsymbol{\mu_\beta}, \boldsymbol{\Sigma}) = \frac{\exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\mu_\beta}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\mu_\beta}\right)\right\}}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}} \quad Eq.(5)$$

where $\boldsymbol{\mu_\beta}$ is the vector of the means; and $\boldsymbol{\Sigma}$ is the covariance matrix. In this manner, we model not just the variances of α_j and β_j, but also the covariances among the α_j's and among the β_j's,

315 respectively. Furthermore, although both α_j and β_j, are modeled as multivariate normal
316 distributions, our primary interest lies in the analysis of β_j. Therefore, **Eq. (5)** explicitly
317 incorporates the β term to represent this focus in our study.

318 Furthermore, because the scale parameter σ in **Eq. (4-1)** must be a positive value, σ is
319 assumed to be a Half-Cauchy distribution with infinite scale parameters (σ'). The distribution of
320 the Half-Cauchy log-likelihood is as follows:

321

322 $$P\left(\sigma \mid \sigma'\right) = \frac{2}{\pi\sigma'\left[1+\left(\frac{\sigma}{\sigma'}\right)^2\right]} \quad Eq.(6)$$

323

324 Please also note that, in our model, **Eq. (4)** defines the likelihood for each observation,
325 with the mean μ_i being modeled as a linear function of the predictors ($\boldsymbol{X}_i$), with coefficients that
326 vary by group (α_j and β_j) (**Eq. (3)**). Specifically, β is modeled by assuming a *priori* that they have
327 a zero mean matrix and their covariance matrices are provided by identification of the matrix
328 using the multivariate normal distribution function. This process provides a flexible family of prior
329 distributions for the matrix logarithm of the covariance structure (Sinay and Hsu, 2014). In
330 specific, this prior structure results in shrinkage of the β_j coefficients towards zero and each
331 other, depending on the covariance structure, while also assisting in the regularization process
332 and mitigating overfitting. Utilizing the No U-Turn Sampler (NUTS) (Hoffman and Gelman 2014)
333 method for posterior estimation, these priors could help enhance the mixing of chains (e.g.,
334 making the chains less likely to get stuck in one region) and reduce autocorrelation (e.g., ensuring

the chain moves more quickly and independently around the parameter space), leading to efficient sampling and accurate posterior estimation.

Based on the likelihood and prior distributions above, the joint posterior probability of the model parameter can be estimated using the Bayes theorem, which provides a principled way to calculate a conditional probability. Therefore, we then obtained the posterior distribution of the parameters $p(\boldsymbol{\beta} \mid fMSE^{(\lambda)}, X)$ from Bayes' theorem as shown below:

$$p\left(\boldsymbol{\beta} \mid fMSE^{(\lambda)}, \boldsymbol{X}\right) = \frac{p\left(fMSE^{(\lambda)} \mid \boldsymbol{\beta}, \boldsymbol{X}\right)P(\boldsymbol{\beta})}{p\left(fMSE^{(\lambda)} \mid \boldsymbol{X}\right)} = \frac{p\left(fMSE^{(\lambda)} \mid \boldsymbol{\beta}, \boldsymbol{X}\right)P(\boldsymbol{\beta})}{\int p(fMSE^{(\lambda)} \mid \boldsymbol{\beta}, \boldsymbol{X})P(\boldsymbol{\beta})\, d\boldsymbol{\beta}} \quad Eq.(7)$$

The complexity of the analytical form of $P\left(fMSE^{(\lambda)} \mid \boldsymbol{X}\right)$, which often does not belong to known distribution families nor conjugates with $P\left(fMSE^{(\lambda)} \mid \boldsymbol{\beta}, \boldsymbol{X}\right)$ (Chiu 1996), requires approximating the integrand through sampling to calculate the integral over marginal distributions $\left(fMSE^{(\lambda)} \mid \boldsymbol{X}\right) = \int P(fMSE^{(\lambda)} \mid \boldsymbol{\beta}, \boldsymbol{X})P(\boldsymbol{\beta})\, d\boldsymbol{\beta}$. This is accomplished through the Hamiltonian Monte Carlo (HMC) sampling approach (Hoffman and Gelman 2014) and variational procedures for initial point calculations (Blei et al 2017). In this study, the 792 parameters and hyperparameters of the non-centered hierarchical model were estimated using the NUTS sampler, with initial sampling points determined by the automatic differentiation variational inference method (ADVI) (Kucukelbir et al 2016). Differing from the frequentist approach, NUTS, a type of HMC sampling algorithm, generates the posterior distribution of unknown model parameters based on observed data and prior distribution in Bayesian inference, thereby producing a posterior distribution for parameters, such as β (**Eq. (5)**). This distribution enables

the estimation of summary statistics, inferences, and predictions, proving the utility of Bayesian machine learning in providing robust predictions, especially where data is limited. Our model's validation was facilitated through posterior predictive checks (PPCs), which employ the posterior distribution of model parameters to generate a predictive distribution for new observations, accounting for model parameter uncertainty and assessing the model's fit with the observed data.

In summary, we transformed the *fMSE* data to a normal distribution through a Box-Cox transformation and confirmed the fit using the SSE method against 80 distribution candidates. We then built a hierarchical model where the mean is a function of predictors with group-level effects and group-specific coefficients, both following a multivariate normal distribution. We assumed a prior structure for the β coefficients based on a zero mean matrix and a covariance matrix identified from the multivariate normal distribution, while the scale parameter follows a Half-Cauchy distribution. Finally, the joint posterior probability was estimated using the Bayes' theorem and the NUTS sampler, with initial points obtained from the ADVI method. To estimate the posterior distribution for the β coefficients, we use the NUTS sampler, which generates a series of smart proposals through the parameter space. It starts at initial points defined by the ADVI method and proceeds with a trajectory until it appears to make a U-turn, ensuring efficient exploration. The iterative process of proposal and acceptance/rejection following the Metropolis-Hastings criterion results in a sequence of β coefficients samples representing the desired posterior distribution.

4. Results and discussion

4.1. Bayesian inference model evaluations

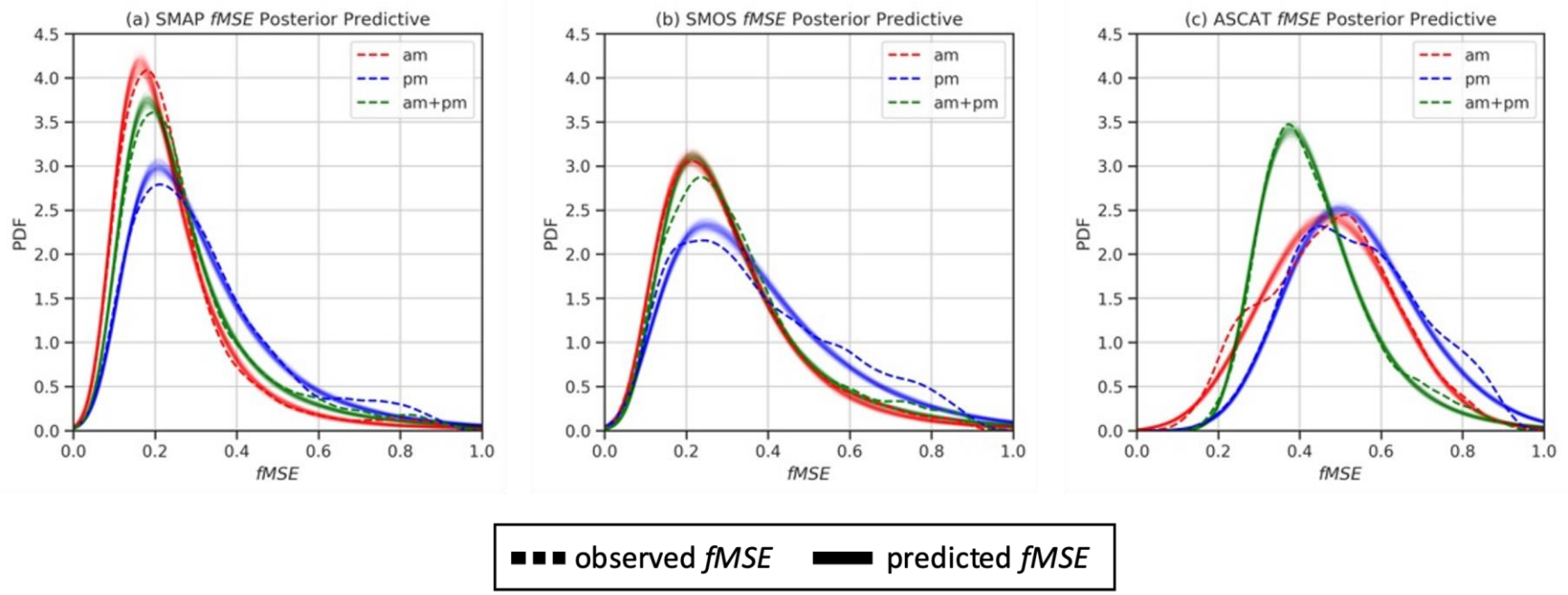
Fig. 2 shows the posterior predictive *fMSE* values (i.e., calculated from $\left(\overline{fMSE^{(\lambda)}}\lambda + 1\right)^{\frac{1}{\lambda}}$) for each satellite-based SM product for a.m. (solid red lines), p.m. (solid blue lines), and combined (solid green lines) overpasses. We have 2,000 HLMs for each product, obtained from 2,000 converged Bayesian HLMs using NUTS. Consequently, 2,000 PDF lines are used for the posterior predictive analysis of *fMSE* values calculated from 2,000 individual HLM for the SMAP (a.m., p.m., and a.m.+p.m.), SMOS (a.m., p.m., and a.m.+p.m.), and ASCAT (a.m., p.m., and a.m.+p.m.) cases.

The PDF of the predicted *fMSE* follow the observed *fMSE* data (dashed line for each product) remarkably well, indicating that the BHM can reasonably describe *fMSE* values over different land cover conditions based on the 14 chosen predictors. Please note that the Box-Cox transformed observed/predicted *fMSE* values were inversed to the original scale *fMSE* to illustrate these results. It is worth noting that the precision of SM data is improved (i.e., lower *fMSE*) if the a.m. and p.m. products are combined; however, at the same time, the predictive precision of *fMSE* from the Bayesian HLM can be reduced for the combined (a.m.+p.m.) SMAP and SMOS SM cases. This suggests that making an inference from the model parameters is harder with a.m.+p.m. data (i.e., understanding the impact of predictors on the precision of SM data is harder) because when the two data sets from passive microwave systems with different error characteristics are combined, the impact of error sources on SM precision can be blurred. For example, the model has trouble finding the relationship between the error characteristics of a.m.+p.m. SM data with time-averaged surface temperature.

On the other hand, the a.m.+p.m. case is a better fit for the active system (green lines in **Fig. 2(c)**). This could be because the ASCAT SM retrieval algorithm does not require land temperature inputs; therefore, its error estimation can be less sensitive to diurnal differences in

thermal conditions. Specifically, the 14 predictors that are being used might not be strong enough to describe the variability of the ASCAT a.m. or p.m. SM error. For example, it is hypothesized that the omission of predictors related to subsurface scattering, which is known to be one of the largest sources of error for ASCAT, can be effectively counterbalanced by averaging and combining a.m. and p.m. soil moisture (SM) data from three Metop satellites. This approach is postulated to mitigate the impact of subsurface scattering conditions on ASCAT SM, thereby significantly enhancing the model's capability to accurately describe the SM retrieval error using the current 14 predictors.

Additionally, ASCAT retrievals likely capture shallower, and thus higher-frequency, soil moisture dynamics than ~5-cm estimates from SMAP/SMOS and 10-cm estimates obtained from GLDAS (Wagner *et al* 2013, Wigneron *et al* 2017). Since TCA tends to punish outlier products, TCA results calculated for the triplets ASCAT/SMAP/GLDAS or ASCAT/SMOS/GDLAS triplets may therefore penalize ASCAT SM retrievals relative to SMAP and SMOS. Consequently, *fMSE*, calculated from the average of multiple ASCAT SM data per day (ASCAT H119/120 SM data contains SM retrievals obtained from the three Metop satellites), may lower a.m.+p.m. SM's *fMSE* versus the sole use of a.m. or p.m. data - since averaging ASCAT SM data reduces noise and smooths out high-frequency variability. Using these Bayesian HLMs, the association between each predictor and the precision of the SM data (i.e., *fMSE*) will be explored next, using a posterior marginal distribution of each predictor's parameter across different land cover types.



419

420 **Figure 2**. The posterior predictive of BHM-predicted *fMSE* using the NUTS method. Each PDF graph, **(a)**, **(b)**, and **(c)** shows the PDF for
421 the posterior predictive *fMSE* values (2,000 solid lines) from the Bayesian HLM produced by the NUTS method for the SMAP, SMOS,
422 and ASCAT products (grouped by a.m., p.m., and a.m.+p.m.). The observed *fMSE* probability PDF is indicated by the dashed lines.

4.2. The Usefulness of the Bayesian Hierarchical Modeling Approach

Our comparison results of BHM and frequentist approaches (i.e., the complete pooling model (CPM) and no-pooling model (NPM)) in **Figs. S2 – 17**, demonstrate that BHM can provide a more comprehensive probability distribution of $\boldsymbol{\beta}$ values than CPM or NPM and it offers a clearer picture of the associated uncertainty. This is because HLM offers several advantages over NPM, including enhancing parameter estimation through the judicious consideration of data from each land cover type and other land cover types. This leads to more precise parameter estimates, particularly in circumstances characterized by small sample sizes or noisy data. Finally, BHM also permits the incorporation of prior knowledge and assumptions about the data, which augments the estimation of $\boldsymbol{\beta}$ values and mitigates uncertainty. Since BHM can provide a complete distribution of β, offering a comprehensive understanding of the uncertainty of β, we use BHM for the remainder of the study to analyze the impact of 14 variables on *fMSE*.

Fig. 3 reveals a distinct difference in the association of VOD with SM data quality between a.m. and p.m. retrievals. As illustrated in **Figs. 3(a)**, **(b)**, **(c)**, and **(d)**, the association between vegetation matter and the quality of SMAP and SMOS SM retrievals varies depending on the overpass time of the satellites. One possible explanation for this is that the SM retrieval algorithms from these passive systems require an assumption of thermal equilibrium assumption, with 6 a.m. being an ideal time to achieve this status (Entekhabi et al. 2010). There is a higher likelihood of violating this assumption at 6 p.m. due to the potential temperature gradients resulting from vegetation which increase the impact of VOD on SM data quality during the late afternoon overpass time. However, the impact of VOD for the active system is different from that of the passive systems. The ASCAT SM retrieval algorithm (TU Wien algorithm) does not require

land temperature inputs; rather, SM retrievals are based on the relative backscatter values to historically maximum and minimum values. Our results, as seen in **Figs. 3(e)** and **(f)**, indicate that during the a.m. there is a significant positive association between vegetation dynamics and *fMSE* in the TU Wien SM retrieval algorithm (except ENF and CNV). However, this positive association is substantially weaker during the p.m. overpass time and is only evident for WS, Gr, OS, and EBF land cover types. This suggests that during p.m. overpasses, the effect of vegetation on *fMSE* is more challenging to determine, potentially due to greater day-to-day fluctuations in vegetation water content at 9 p.m. compared to 9 a.m.

In addition, over cropland, the relationship with uncertainties of satellite-based SM data is not solely determined by VOD and is linked to other environmental factors (please refer to our discussion pertaining to **Figs. 4** and **5** below). Additional results that are analogous to **Fig. 3**, but for different predictor variables and overpass times, are summarized in **Figs. S2 – S14**

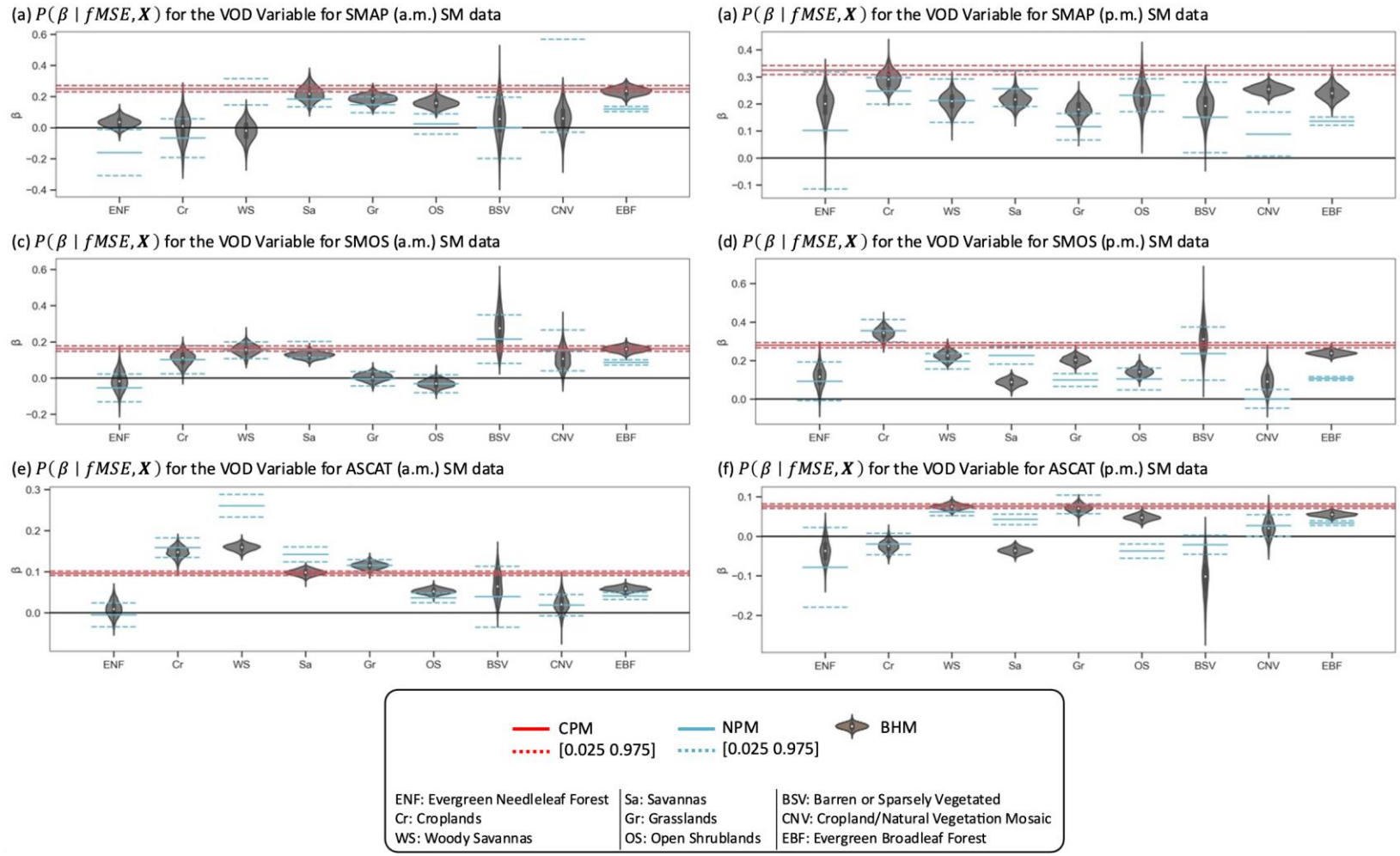


Figure 3. Parameter (β) estimations for the VOD variable from the complete pooling model (CPM), no-pooling model (NPM), and Bayesian hierarchical model (BHM) for (a, b, c) SMAP, (d, e, f) SMOS-IC, and (g, h, i) ASCAT for different overpass times.

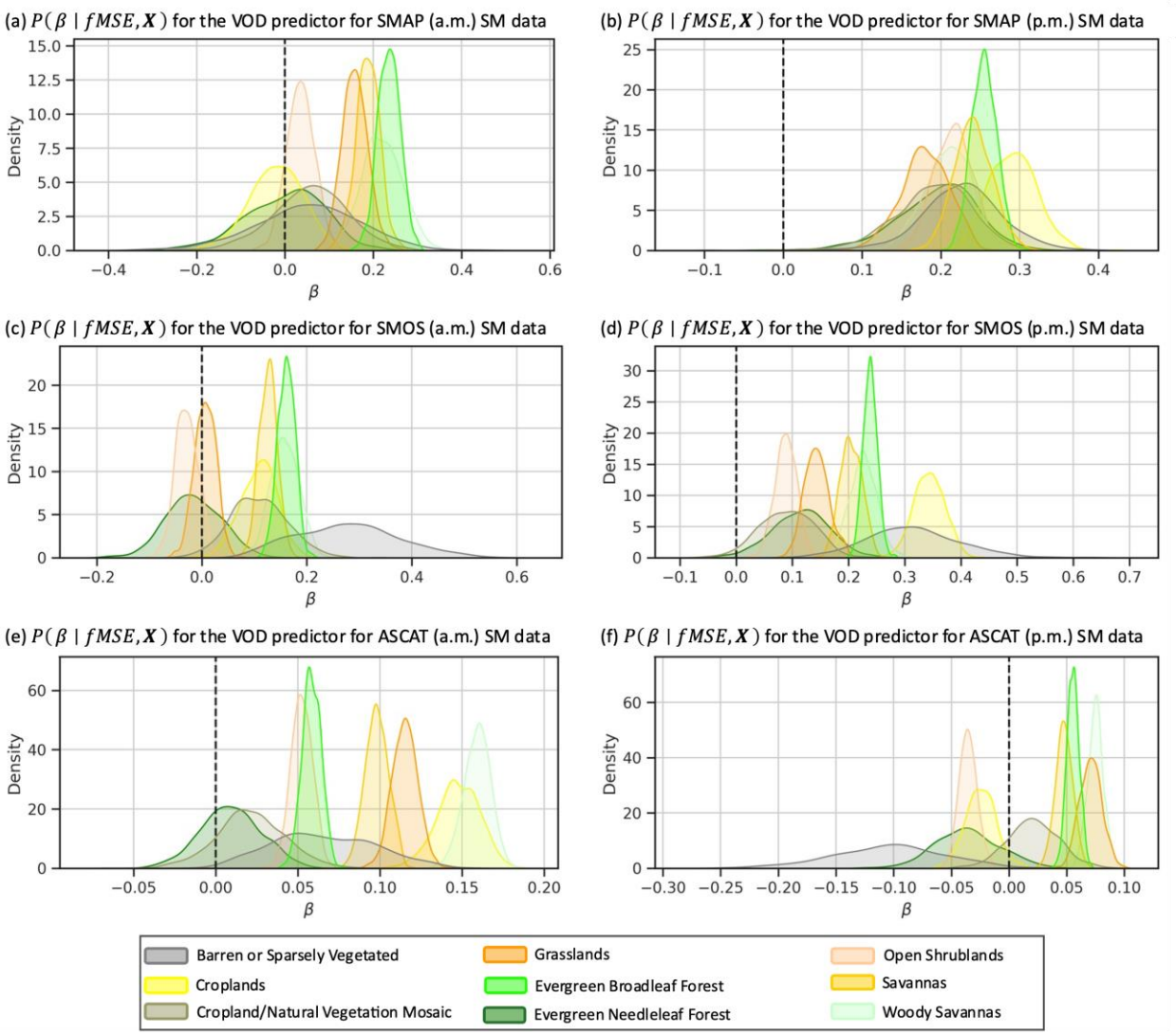
463 **4.3. Association of different error sources with the quality of satellite-based SM data**

464 Here, we examine the most interesting parameter sensitivity cases illustrated by the BHM.
465 In particular, we use 14 selected predictors to investigate the association of human activities (i.e.,
466 irrigation activities), vegetation mass (i.e., VOD), and SOM with the quality of SM retrieval data,
467 as these factors are generally considered to be major impediments to the retrieval of SM using
468 microwave satellite systems. (The results for the other nine predictors are included in the SI
469 document.) Even though we only included 14 predictor variables, the current framework can be
470 applied to analyze any other set of environmental factors. However, it should be stressed that, if
471 the posterior predictive (solid lines in **Fig. 2**) from different models do not correspond with the
472 observed *fMSE* (dashed solid lines in **Fig. 2**), the statistical inference from the Bayesian HLM will
473 be unreliable.

474 First, we evaluate the uncertainties of predictor variables by examining the credible
475 intervals of parameters to understand their associations with SM data quality across a range of
476 satellite products. Specifically, **Fig. 4** shows a correlation between SM retrievals for each satellite
477 system and the amount of vegetation mass over different land cover types. Different PDF lines
478 indicate different land cover types. $\boldsymbol{\beta}$ is the location parameter in **Eq. (7)** which shows the
479 distribution of μ in **Eq. (4)**. To determine the credibility of a predictor related to β, the
480 distribution of β should not cross zero or include zero within the ±95% density interval. In
481 addition, positive β values suggest that a factor tends to increase *fMSE* (i.e., degrades SM quality).
482 A wider distribution indicates greater uncertainty regarding a variable's impact on *fMSE*.

483 **Fig. 4** illustrates a strong association between vegetation and SM retrieval errors for most
484 land cover types. This correlation is seen as naturally emitted signals observed from passive

485 satellites, or the backscattered energy generated from active sensors, are very sensitive to
486 vegetation. This relationship is further supported by the $\boldsymbol{\beta}$ distributions that are relatively narrow
487 and do not cross zero. Nonetheless, the dispersed distribution (i.e., large σ) for barren or sparsely
488 vegetated (BSV) land cover types suggests a weaker association between VOD and error
489 characteristics in these areas, compared to areas with a narrow $\boldsymbol{\beta}$ distribution. This result also
490 implies that VOD may not play a major role in describing error characteristics for these land cover
491 types, because VOD does not vary significantly either spatially or temporally within these land
492 cover types. This finding underlines the need for careful consideration when building error
493 models primarily dependent on vegetation-related variables, especially for certain land cover
494 types. For land cover types where VOD does not appear strongly associated (i.e., where the
495 posterior $\boldsymbol{\beta}$ distribution is wide and includes zero), *fMSE* might be better characterized by other
496 variables. Please also refer to **Fig. S15** for an examination of the associations between other
497 environmental factors and satellite-based SM retrieval errors over arid environmental conditions.



498

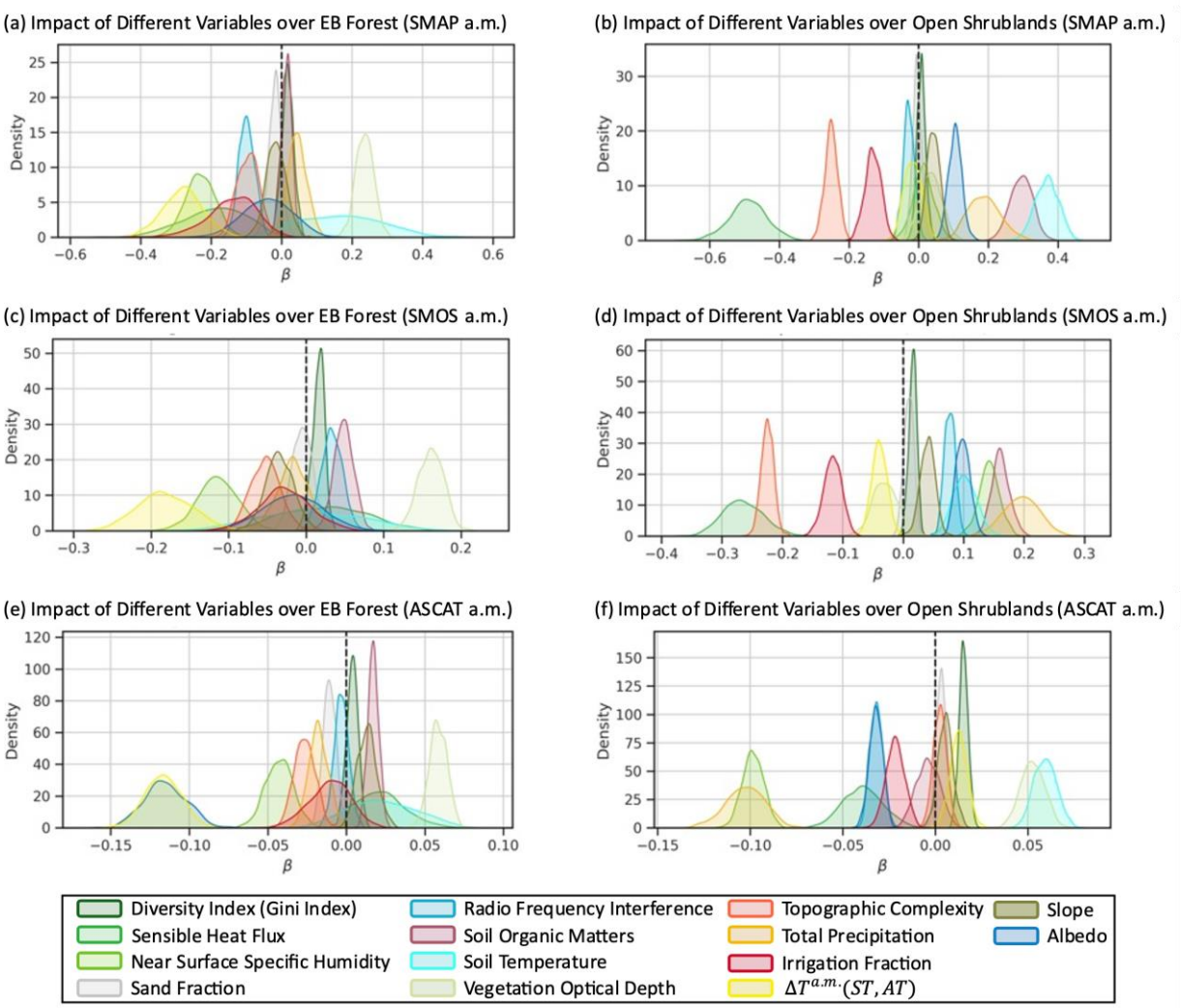
499 **Figure 4**. Posterior marginal distributions for the β parameter assigned vegetation optical depth
500 (VOD) over 9 land cover types. This figure illustrates the relationship between the amount of
501 vegetation matter and error magnitudes in *fMSE* for the a.m. (left) and p.m. (right) overpasses of
502 (a, b) SMAP, (c, d) SMOS, and (e, f) ASCAT SM data across 9 different land cover types. If the
503 distribution of the β parameter is on the positive (negative) side, it indicates that higher
504 vegetation matter is associated with higher (lower) *fMSE* over the corresponding land cover type.

Fig. 5 displays results analogous to **Fig. 4**, but it illustrates the associations between all 14 environmental factors and the quality of satellite-based SM data over evergreen broadleaf forest (hereafter EB forest) and open shrublands land cover types. Here, we only include the a.m. data, but all other results are also included in the supplementary material document (**Figs. S15** to **S23**). Over EB forest and open shrublands areas, temperature and VOD emerge as two primary factors associated with increased SM data errors for both passive and active microwave systems (a.m.) (see the neon fluorescent blue- and light -green-colored PDF lines, respectively, in **Fig. 5**).

In addition, for the passive systems, BHM results point to a correlation between SOM and the diminished quality of SM retrievals. This could stem from the possibility that high SOM can decrease the soil's respective dielectric constant, potentially resulting in greater porosity than anticipated in a SM retrieval algorithm. Current passive microwave SM products (including both SMOS and SMAP) use dielectric constant models based on soil clay content for inversion without considering SOM (Wigneron *et al* 2017). Our result is aligned with previous studies' findings (Zhang et al 2019, Li et al 2022) that suggest a higher fraction of SOM in soil can introduce time-varying SM errors into current passive-microwave SM retrievals algorithms. Li et al. (2022) showed that the higher the fraction of SOM in soil, the lower the performance of SM retrievals from SMOS SM. This degradation was evident in increased bias, higher RMSD (ubRMSD), and decreased correlation coefficient (R) compared to in-situ SM data. A plausible explanation for this degradation is that SOM increases the number of micropores and macropores in the soil by adhering soil particles together, which, in turn, affects the soil properties, including structure. Therefore, SOM affects the soil dielectric properties, while current SMAP and SMOS-IC SM retrieval algorithms use a clay-based dielectric constant model that does not consider the

527 presence of SOM. Similarly, Zhang et al. (2019) showed that, generally, the more the organic
528 carbon in soils, the lower the performance metrics of SMAP SM data (i.e., higher bias and RMSE,
529 and lower correlation coefficient). However, it should be noted that higher bias and time-varying
530 error could be due to the effect of soil freezing and thawing process, since in-situ SM sites with
531 higher SOM tend to be located at higher elevations (**Figs. 1(a)** and **1(c)**). For the active ASCAT
532 system, it appears that uncertainties arising from temperature and VOD are two major factors
533 correlated with an increase in error in SM retrieval data over these land cover types.

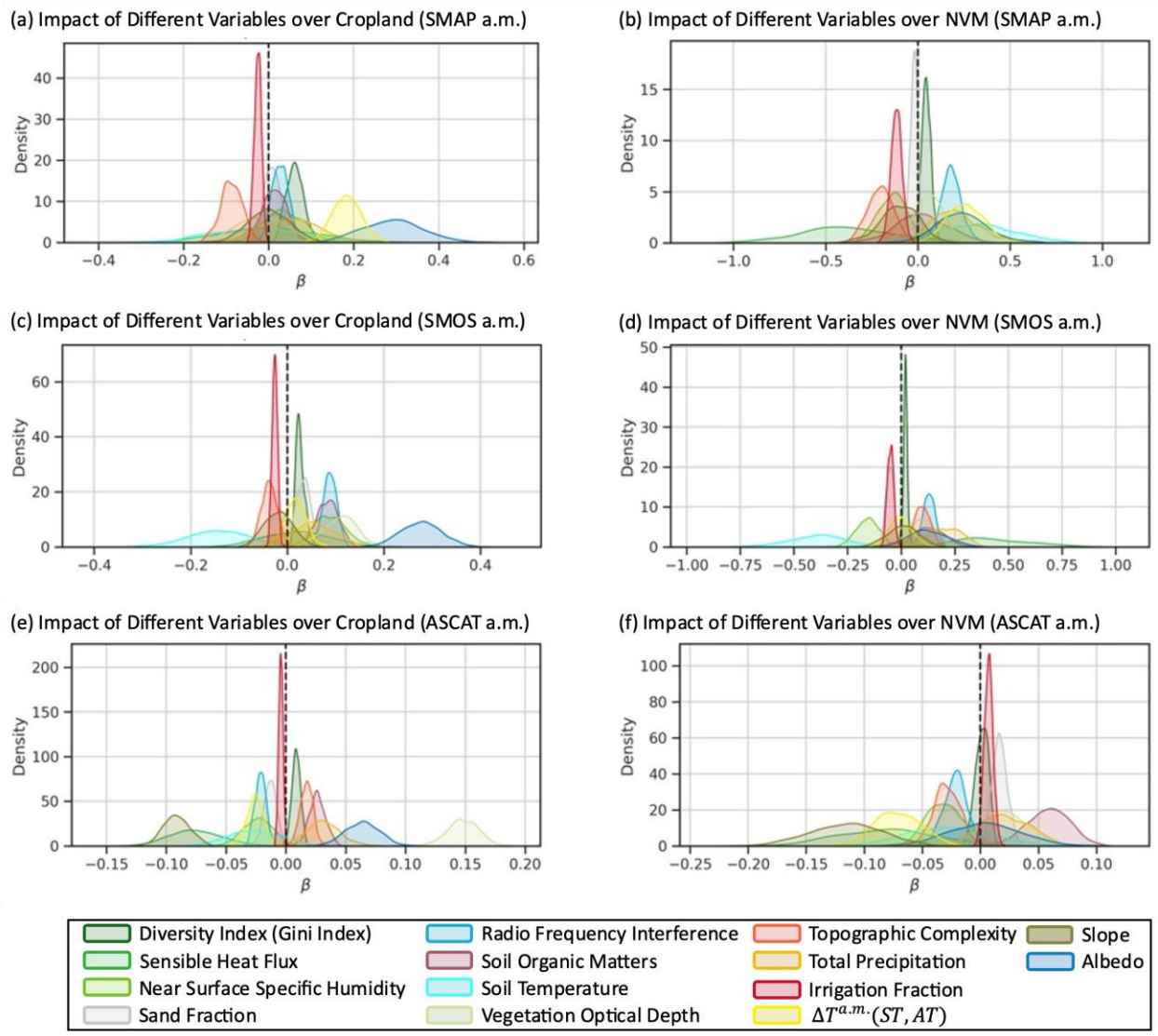
534 It is also intriguing to note the strong similarities between the two passive systems (SMOS
535 and SMAP), as opposed to the active system (ASCAT), in terms of associations with environmental
536 factors. Despite the differences in retrieval systems and algorithms for SMOS and SMAP, the
537 degree and pattern of changes in errors in passive SM retrievals display considerable similarities
538 across various land cover types (see **Fig. 5**). This could potentially be attributed to SMAP and
539 SMOS utilizing similar frequencies and overpass times. This observation underscores the value of
540 integrating SM data from diverse systems (e.g., merging across active and passive systems) when
541 attempting to filter random noise.



542

543 **Figure 5**. The posterior marginal distribution of β parameters for evergreen broadleaf (EB) forest
544 and open shrublands areas for all 14 predictive variables. This figure describes the associations
545 between each of these 14 variables and the errors (i.e., *fMSE*) in (a, b) SMAP (a.m.), (c, d) SMOS
546 (a.m.), and (e, f) ASCAT (a.m.) SM data over the EB forest and open shrublands land cover types.
547 If the distribution of the β parameter falls on the positive side, it indicates that an increase in the
548 corresponding variable is associated with a higher *fMSE* over EB Forest or Open Shrublands.
549 Conversely, if the distribution is on the negative side, this suggests that an increase in the
550 corresponding variable is associated with a lower *fMSE* over EB Forest or Open Shrublands.

Upon examining **Figs. 5** and **Figs. S15-S23**, we can derive insights concerning the association between the intensity of irrigation (represented by red-colored PDF lines) and the quality of SM retrievals for each satellite system. When compared to other environmental factors, the association between irrigation amounts and SM retrieval data quality is relatively slight (i.e., all distributions cross zero with a dispersed distribution shape) across all satellite systems. Furthermore, intensive irrigation does not necessarily correlate with an increase in the uncertainty of SM retrievals from space. This result supports previous findings that satellite-based SM data can play a significant role in detecting the irrigation signals and can be used to improve quality of LSM simulations through the data assimilation (Kim *et al* 2020b, Kwon *et al* 2022, Lawston *et al* 2017, Lei *et al* 2020). In fact, as shown in **Fig. 6**, there are many other factors, other than irrigation fraction, that contribute significantly to increasing errors in SM (a.m.) data over cropland/NVM areas. Furthermore, it was found that over cropland and NVM, observations with less difference between AT and ST ($|\Delta T(ST, AT)|$ were associated with lower *fMSE* for the SMAP/SMOS SM retrieval systems (yellow-colored PDF lines). This result might be related to the assumption of the passive SM retrieval algorithms which require the thermal equilibrium status (Entekhabi *et al* 2010). However, once again, it is worth noting that the lower ($|\Delta T(ST, AT)|$) values do not indicate better quality of SM data across all land cover types. Finally, RFI is another factor increasing errors in passive SM retrieval systems because many croplands are close to RFI sources over East Asia (**Figs. 1(b)** and **(e)**). As shown in **Fig. 6**, it is the combined effect of these multiple factors, rather than one specific error source in isolation, that causes difficulty in retrieving SM over potentially irrigated areas (i.e., cropland and NVM) with passive SM retrieval.



572

573 **Figure 6**. The posterior marginal distribution of β parameters for croplands and natural
574 vegetation mosaic (NVM) areas for 14 variables. This figure describes the associations between
575 each of these 14 variables and the error (i.e., *fMSE*) in (a, b) SMAP, (c, d) SMOS, and (e, f) ASCAT
576 SM data over croplands and croplands/natural vegetation mosaic (NVM). If the distribution of
577 the β parameter falls on the positive side, it indicates that an increase in the corresponding
578 variable is associated with a higher *fMSE* over cropland or NVM. Conversely, if the distribution is
579 on the negative side, this suggests that an increase in the corresponding variable is associated
580 with a lower *fMSE* over cropland or NVM.

Here, we only included results for 9 (originally 17 excluding water) land cover types. The main reason for omitting other land cover types was due to not obtaining sufficient *fMSE* values from the current TCA method. Land-cover types we omitted included polar or cold regions and urban areas where the SM dynamics are relatively meaningless due to snow cover, frozen soil conditions (because signals can be more vulnerable to water in snow, and the relationship between the dielectric constant of frozen water to water content is unreliable), and impervious surface conditions in highly urbanized areas (Wagner 1998).

Finally, despite the apparent ability of the BHM to successfully reproduce ASCAT *fMSE* over valid TCA points, it is still unclear whether or not it is explicitly capturing sub-scattering impacts on ASCAT precision. Further study of this question is required - but will likely require the availability of new predictor variables for subsurface scattering strength (SCS). Therefore, the impact of the SCS on ASCAT SM error characteristics will be investigated in a future study once independent data is available that describes conditions under which SCS is likely.

5. Conclusions

Here, we introduce a novel approach for sensitivity analysis of satellite-based SM error characteristics using a Bayesian hierarchical modeling (BHM) approach for regression parameters and hyper-parameters plus a No-U-Turn Sampler (i.e., sampling approach). We then applied the approach to estimate the credible intervals for 14 selected environmental factors over different land cover types. In this way, we investigated the error characteristics of the three mostly-widely used satellite-based SM data for the nighttime, daytime, and combined overpass times, and

illustrated the advantages and versatility of the BHM approach versus classical regression approaches for the error sensitivity analysis of satellite-based SM retrievals.

We focused on analyzing the impact of irrigation amounts, vegetation mass, and soil organic matter (SOM) on the quality of satellite-based SM data. Results suggest a strong association between vegetation and the errors in satellite-based SM retrievals; however, it is important to note that the quality of SM data cannot be inferred solely from single error sources, as it is also linked to many other factors. By comparing BHM with the classical regression model with the frequentist approach for sensitivity analysis, we demonstrated how CPM or NPM with the frequentist approach can lead to different/misleading results -- for instance, based on CPM, one can draw opposing inferences about the impact of vegetation on the quality of the satellite-based SM data. In addition, we also found an association between SOM and challenges in retrieving SM information from passive microwave sensors. Moreover, we observed that the combined presence of signal attenuations from vegetation and RFI seems to be correlated with further difficulties in SM information retrieval. Over potentially irrigated areas such as croplands and natural vegetation mosaic, the degree of irrigation may not be used for inferring the quality of SM data, as other factors (e.g., thermal equilibrium status, albedo, RFI) control the quality of SM data over these areas. Although we could only include 14 predictors with the SM variable in the current analysis, our approach is highly general, and many other predictors and time-varying geophysical variables can be added by individual researchers according to their own research needs.

Lastly, it is essential to adopt a streamlined approach that employs fewer but more effective predictor variables and improved land cover maps, along with simplified statistical

models, to address overfitting and incorporate critical aspects such as subsurface scattering parameters that were previously omitted. Additionally, it is important to acknowledge that the triplets used to calculate fMSE may be inherently more favorable to SMAP and SMOS due to the particular combinations of sensors and datasets employed, which could have influenced the outcome. By refining the regression models to encompass these vital components and critically examining the selection of triplets, the robustness and reliability of the results can be enhanced. This refined approach is expected to foster a deeper understanding of the processes in question and lead to more accurate interpretations of the interactions between microwave signals and various factors, including subsurface properties. Including subsurface scattering parameters is particularly crucial for establishing a more accurate error prediction model.

Acknowledgments

The authors thank the teams from NASA, USDA, USGS, ESA, and INRA for making their data sets publicly available. Hyunglok Kim acknowledges the Future Investigators in NASA Earth and Space Science and Technology (FINESST) Award (#80NSSC19K1337) and the AGU Horton Grant. The USDA is an equal opportunity employer and provider. Data supporting the conclusions of this study are properly cited and publicly available. SMAP L3 and IGBP land classification data sets are available at https://nsidc.org/data/SPL3SMP. ASCAT data is available at https://land.copernicus.vgt.vito.be. SMOS-IC data is available at https://ib.remote-sensing.inrae.fr. The Land Information System Framework (LISF) source code is available at https://github.com/NASA-LIS/LISF. The boxplot and violin plot is available at https://www.mathworks.com/matlabcentral/fileexchange/91790-al_goodplot-boxblot-violin-plot and MT-DCA Tau data is available at https://zenodo.org/record/5579549#.YyOdbyHMIqs (Feldman et al. 2021). Online HTML-based documentation reflecting the master branch of NASA-LIS/LISF is available on https://nasa-lis.github.io/LISF/ and GitHub page (https://github.com/NASA-LIS/LISF).

Disclaimer

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the US Government.

References

Arsenault K R, Kumar S V, Geiger J V, Wang S, Kemp E, Mocko D M, Beaudoing H K, Getirana A, Navari M, Li B, Jacob J, Wegiel J and Peters-Lidard C D 2018 The Land surface Data Toolkit (LDT v7.2) – a data fusion environment for land data assimilation systems *Geosci. Model Dev.* **11** 3605–21

Brocca L, Filippucci P, Hahn S, Ciabatta L, Massari C, Camici S, Schüller L, Bojkov B and Wagner W 2019 SM2RAIN–ASCAT (2007–2018): global daily satellite rainfall data from ASCAT soil moisture observations *Earth Syst. Sci. Data* **11** 1583–601

Calvet J-C, Wigneron J-P, Walker J, Karbou F, Chanzy A and Albergel C 2011 Sensitivity of Passive Microwave Observations to Soil Moisture and Vegetation Water Content: L-Band to W-Band *IEEE Trans. Geosci. Remote Sensing* **49** 1190–9

Cho E, Su C-H, Ryu D, Kim H and Choi M 2017 Does AMSR2 produce better soil moisture retrievals than AMSR-E over Australia? *Remote Sensing of Environment* **188** 95–105

Colliander A, Jackson T J, Berg A, Bosch D D, Caldwell T, Chan S, Cosh M H, Collins C H, Martínez-Fernández J, McNairn H, Prueger J H, Starks P J, Walker J P and Yueh S H 2020 Effect of Rainfall Events on SMAP Radiometer-Based Soil Moisture Accuracy Using Core Validation Sites *Journal of Hydrometeorology* **21** 255–64

672 Cosh M H, Jackson T J, Bindlish R, Famiglietti J S and Ryu D 2005 Calibration of an impedance
673 probe for estimation of surface soil water content over large regions *Journal of*
674 *Hydrology* **311** 49–58

675 Crow W T, Dong J and Reichle R H 2022 Leveraging Pre-Storm Soil Moisture Estimates for
676 Enhanced Land Surface Model Calibration in Ungauged Hydrologic Basins *Water*
677 *Resources Research* **58** Online:
678 https://onlinelibrary.wiley.com/doi/10.1029/2021WR031565

679 Crow W T, Han E, Ryu D, Hain C R and Anderson M C 2017 Estimating annual water storage
680 variations in medium-scale (2000–10 000 km²) basins using
681 microwave-based soil moisture retrievals *Hydrol. Earth Syst. Sci.* **21** 1849–62

682 Entekhabi D, Njoku E G, O'Neill P E, Kellogg K H, Crow W T, Edelstein W N, Entin J K, Goodman S
683 D, Jackson T J, Johnson J, Kimball J, Piepmeier J R, Koster R D, Martin N, McDonald K C,
684 Moghaddam M, Moran S, Reichle R, Shi J C, Spencer M W, Thurman S W, Tsang L and Zyl
685 J V 2010 The Soil Moisture Active Passive (SMAP) Mission *Proceedings of the IEEE* **98**
686 704–16

687 Gelaro R, McCarty W, Suárez M J, Todling R, Molod A, Takacs L, Randles C A, Darmenov A,
688 Bosilovich M G, Reichle R, Wargan K, Coy L, Cullather R, Draper C, Akella S, Buchard V,
689 Conaty A, da Silva A M, Gu W, Kim G-K, Koster R, Lucchesi R, Merkova D, Nielsen J E,
690 Partyka G, Pawson S, Putman W, Rienecker M, Schubert S D, Sienkiewicz M and Zhao B
691 2017 The Modern-Era Retrospective Analysis for Research and Applications, Version 2
692 (MERRA-2) *J. Climate* **30** 5419–54

693 He L, Chen J M, Mostovoy G and Gonsamo A 2021 SMAP improves global soil moisture
694 simulation in a land surface scheme and reveals strong irrigation signals over farmlands
695 *Geophys Res Lett* Online: https://onlinelibrary.wiley.com/doi/10.1029/2021GL092658

696 Hirschi M, Mueller B, Dorigo W and Seneviratne S I 2014 Using remotely sensed soil moisture
697 for land–atmosphere coupling diagnostics: The role of surface vs. root-zone soil
698 moisture variability *Remote Sensing of Environment* **154** 246–52

699 Hoffman M D and Gelman A 2014 The No-U-Turn Sampler: Adaptively Setting Path Lengths in
700 Hamiltonian Monte Carlo *Journal of Machine Learning Research* **15** 1593–623

701 Jackson T J, Schmugge J and Engman E T 1996 Remote sensing applications to hydrology: soil
702 moisture *Hydrological Sciences Journal* **41** 517–30

703 Kerr Y H, Waldteufel P, Wigneron J-P, Delwart S, Cabot F, Boutin J, Escorihuela M-J, Font J, Reul
704 N, Gruhier C, Juglea S E, Drinkwater M R, Hahne A, Martín-Neira M and Mecklenburg S
705 2010 The SMOS Mission: New Tool for Monitoring Key Elements ofthe Global Water
706 Cycle *Proc. IEEE* **98** 666–87

Kim H, Cosh M H, Bindlish R and Lakshmi V 2020a Field evaluation of portable soil water content sensors in a sandy loam *Vadose zone j.* **19** Online: https://onlinelibrary.wiley.com/doi/abs/10.1002/vzj2.20033

Kim H and Lakshmi V 2018 Use of Cyclone Global Navigation Satellite System (cygnss) Observations for Estimation of Soil Moisture *Geophys. Res. Lett.* **45** 8272–82

Kim H, Parinussa R, Konings A G, Wagner W, Cosh M H, Lakshmi V, Zohaib M and Choi M 2018 Global-scale assessment and combination of SMAP with ASCAT (active) and AMSR2 (passive) soil moisture products *Remote Sensing of Environment* **204** 260–75

Kim H, Wigneron J-P, Kumar S, Dong J, Wagner W, Cosh M H, Bosch D D, Collins C H, Starks P J, Seyfried M and Lakshmi V 2020b Global scale error assessments of soil moisture estimates from microwave-based active and passive satellites and land surface models over forest and mixed irrigated/dryland agriculture regions *Remote Sensing of Environment* **251** 112052

Konings A G, Piles M, Das N and Entekhabi D 2017 L-band vegetation optical depth and effective scattering albedo estimation from SMAP *Remote Sensing of Environment* **198** 460–70

Kumar S, Peterslidard C, Tian Y, Houser P, Geiger J, Olden S, Lighty L, Eastman J, Doty B and Dirmeyer P 2006 Land information system: An interoperable framework for high resolution land surface modeling *Environmental Modelling & Software* **21** 1402–15

Kwon Y, Kumar S V, Navari M, Mocko D M, Kemp E M, Wegiel J W, Geiger J V and Bindlish R 2022 Irrigation characterization improved by the direct use of SMAP soil moisture anomalies within a data assimilation system *Environ. Res. Lett.* **17** 084006

Lawston P M, Santanello J A and Kumar S V 2017 Irrigation Signals Detected From SMAP Soil Moisture Retrievals: Irrigation Signals Detected From SMAP *Geophys. Res. Lett.* **44** 11,860-11,867

Lei F, Crow W T, Kustas W P, Dong J, Yang Y, Knipper K R, Anderson M C, Gao F, Notarnicola C, Greifeneder F, McKee L M, Alfieri J G, Hain C and Dokoozlian N 2020 Data assimilation of high-resolution thermal and radar remote sensing retrievals for soil moisture monitoring in a drip-irrigated vineyard *Remote Sensing of Environment* **239** 111622

Li X, Al-Yaari A, Schwank M, Fan L, Frappart F, Swenson J and Wigneron J-P 2020 Compared performances of SMOS-IC soil moisture and vegetation optical depth retrievals based on Tau-Omega and Two-Stream microwave emission models *Remote Sensing of Environment* **236** 111502

Misra S and Ruf C S 2012 Analysis of Radio Frequency Interference Detection Algorithms in the Angular Domain for SMOS *IEEE Trans. Geosci. Remote Sensing* **50** 1448–57

741 Muñoz-Sabater J, Dutra E, Agustí-Panareda A, Albergel C, Arduini G, Balsamo G, Boussetta S,
742 Choulga M, Harrigan S, Hersbach H, Martens B, Miralles D G, Piles M, Rodríguez-
743 Fernández N J, Zsoter E, Buontempo C and Thépaut J-N 2021 ERA5-Land: a state-of-the-
744 art global reanalysis dataset for land applications *Earth Syst. Sci. Data* **13** 4349–83

745 Nguyen H H, Kim H and Choi M 2017 Evaluation of the soil water content using cosmic-ray
746 neutron probe in a heterogeneous monsoon climate-dominated region *Advances in*
747 *Water Resources* **108** 125–38

748 de Nijs A H A, Parinussa R M, de Jeu R A M, Schellekens J and Holmes T R H 2015 A
749 Methodology to Determine Radio-Frequency Interference in AMSR2 Observations *IEEE*
750 *Trans. Geosci. Remote Sensing* **53** 5148–59

751 Niu G-Y, Yang Z-L, Mitchell K E, Chen F, Ek M B, Barlage M, Kumar A, Manning K, Niyogi D,
752 Rosero E, Tewari M and Xia Y 2011 The community Noah land surface model with
753 multiparameterization options (Noah-MP): 1. Model description and evaluation with
754 local-scale measurements *J. Geophys. Res.* **116** D12109

755 Oliva R, Daganzo E, Kerr Y H, Mecklenburg S, Nieto S, Richaume P and Gruhier C 2012 SMOS
756 Radio Frequency Interference Scenario: Status and Actions Taken to Improve the RFI
757 Environment in the 1400–1427-MHz Passive Band *IEEE Trans. Geosci. Remote Sensing* **50**
758 1427–39

759 Owe M, de Jeu R and Walker J 2001 A methodology for surface soil moisture and vegetation
760 optical depth retrieval using the microwave polarization difference index *IEEE Trans.*
761 *Geosci. Remote Sensing* **39** 1643–54

762 Reichle R H 2008 Data assimilation methods in the Earth sciences *Advances in Water Resources*
763 **31** 1411–8

764 Reynolds S G 1970 The gravimetric method of soil moisture determination Part I A study of
765 equipment, and methodological problems *Journal of Hydrology* **11** 258–73

766 Rodell M, Houser P R, Jambor U, Gottschalck J, Mitchell K, Meng C-J, Arsenault K, Cosgrove B,
767 Radakovich J, Bosilovich M, Entin J K, Walker J P, Lohmann D and Toll D 2004 The Global
768 Land Data Assimilation System *Bull. Amer. Meteor. Soc.* **85** 381–94

769 Seneviratne S I, Lüthi D, Litschi M and Schär C 2006 Land–atmosphere coupling and climate
770 change in Europe *Nature* **443** 205–9

771 Siebert S, Doll P and Hoogeveen J 2005 Development and validation of the global map of
772 irrigation areas *Hydrology and Earth System Sciences* 13

773 Vaz C M P, Jones S, Meding M and Tuller M 2013 Evaluation of Standard Calibration Functions
774 for Eight Electromagnetic Soil Moisture Sensors *Vadose Zone Journal* **12** vzj2012.0160

Wagner W 1998 SOIL MOISTURE RETRIEVAL FROM ERS SCATTEROMETER DATA *Inst. für Photogrammetrie u. Fernerkundung d. Techn. Univ.* **49**

Wagner W, Hahn S, Kidd R, Melzer T, Bartalis Z, Hasenauer S, Figa-Saldaña J, de Rosnay P, Jann A, Schneider S, Komma J, Kubu G, Brugger K, Aubrecht C, Züger J, Gangkofner U, Kienberger S, Brocca L, Wang Y, Blöschl G, Eitzinger J and Steinnocher K 2013 The ASCAT Soil Moisture Product: A Review of its Specifications, Validation Results, and Emerging Applications *metz* **22** 5–33

Wagner W, Lemoine G, Borgeaud M and Rott H 1999 A study of vegetation cover effects on ERS scatterometer data *IEEE Trans. Geosci. Remote Sensing* **37** 938–48

Wagner W, Lindorfer R, Melzer T, Hahn S, Bauer-Marschallinger B, Morrison K, Calvet J-C, Hobbs S, Quast R, Greimeister-Pfeil I and Vreugdenhil M 2022 Widespread occurrence of anomalous C-band backscatter signals in arid environments caused by subsurface scattering *Remote Sensing of Environment* **276** 113025

Wagner W, Naeimi V, Scipal K, de Jeu R and Martínez-Fernández J 2007 Soil moisture from operational meteorological satellites *Hydrogeol J* **15** 121–31

Wang Z, Zeng X, Barlage M, Dickinson R E, Gao F and Schaaf C B 2004 Using MODIS BRDF and Albedo Data to Evaluate Global Model Land Surface Albedo *J. Hydrometeor* **5** 3–14

Wigneron J-P, Jackson T J, O'Neill P, De Lannoy G, de Rosnay P, Walker J P, Ferrazzoli P, Mironov V, Bircher S, Grant J P, Kurum M, Schwank M, Munoz-Sabater J, Das N, Royer A, Al-Yaari A, Al Bitar A, Fernandez-Moran R, Lawrence H, Mialon A, Parrens M, Richaume P, Delwart S and Kerr Y 2017 Modelling the passive microwave signature from land surfaces: A review of recent results and application to the L-band SMOS & SMAP soil moisture retrieval algorithms *Remote Sensing of Environment* **192** 238–62

Xia Y, Mitchell K, Ek M, Cosgrove B, Sheffield J, Luo L, Alonge C, Wei H, Meng J, Livneh B, Duan Q and Lohmann D 2012 Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow: VALIDATION OF MODEL-SIMULATED STREAMFLOW *J. Geophys. Res.* **117** n/a-n/a