Uncertainty in Soil Moisture Retrievals: an Ensemble Approach using SMOS L-Band Microwave Data

Jan Quets[1], Gabriëlle J. M. De Lannoy[1,*], Amen Al Yaari[2], Steven Chan[3], Michael H. Cosh[4], Alexander Gruber[1], Rolf H. Reichle[5], Robin Van der Schalie[6], Jean-Pierre Wigneron[2]

[1] KU Leuven, Department of Earth and Environmental Sciences, 3001 Heverlee, Belgium*****

[2] ISPA, INRA Bordeaux, France

[3] NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109 USA

[4] USDA Agricultural Research Service-Hydrology and Remote Sensing Laboratory, Beltsville, MD USA

[5] Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, MD, USA

[6] VanderSat B.V, Haarlem, the Netherlands

*****Correspondence: jquets@gmail.com, gabrielle.delannoy@kuleuven.be

Received: 11/02/2019; Accepted: date; Published: date

Abstract: The uncertainty of surface soil moisture (SM) retrievals from satellite brightness temperature (TB) observations depends primarily on the choice of radiative transfer model (RTM) parameters, prior SM information and TB inputs. This paper studies the sensitivity of several established and experimental SM retrieval products from the Soil Moisture Ocean Salinity (SMOS) mission to these choices at 11 reference sites, located in 7 watersheds across the United States (US). Different RTM parameter sets cause large biases between retrievals. Whereas typical RTM parameter sets are calibrated for SM retrievals, it is shown

that a parameter set carefully optimized for TB forward modeling can also be used for retrieving SM. It is also shown that the inclusion of dynamic prior SM estimates in a Bayesian retrieval scheme can strongly improve SM retrievals, regardless of the choice of RTM parameters. The second part of this paper evaluates the ensemble uncertainty metrics for SM retrievals obtained by propagating a wide range of RTM parameters through the RTM, and the relationship with time series metrics obtained by comparing SM retrievals with *in situ* data. As expected for bounded variables, the total spread in the ensemble SM retrievals is smallest for wet and dry SM values and highest for intermediate SM values. After removal of the strong long-term SM bias associated with the RTM parameter values for individual ensemble members, the remaining anomaly ensemble SM spread shows higher values when SM deviates further from its long-term mean SM. This reveals higher-order biases (e.g. differences in variances) in the retrieval error, which should be considered when characterizing retrieval error. The time-average anomaly ensemble SM spread of 0.037 m^3/m^3 approximates the actual time series unbiased root-mean-square-difference of 0.042 m^3/m^3 between ensemble mean retrievals and *in situ* data across the reference sites.

Keywords: soil moisture; retrieval; radiative transfer model; parameters; ensemble; uncertainty

Highlights:

- wetter SM retrievals for higher roughness and vegetation parameter values
- good SM retrievals possible with RTM parameters calibrated for TB forward modeling
- improved SM retrievals with inclusion of model-based prior SM estimates
- similar SM mean - variance relationship sampled in time, space, or across an ensemble
- higher-order biases in SM retrieval error after long-term mean removal

1. Introduction

Surface soil moisture (SM) is an essential climate variable that can be estimated from space by analyzing the interaction of microwave radiation with the soil-water matrix, using either active or passive sensors. Currently, two passive L-band microwave missions are specifically designed to monitor SM: the Soil Moisture and Ocean Salinity (SMOS) and Soil Moisture Active Passive (SMAP) missions are producing global products of SM retrievals (Kerr et al. 2011, 2012, 2016, Entekhabi et al. 2010a, O'Neill et al. 2012).

The SM retrieval from satellite-based brightness temperature (TB), however, is not a well-posed problem: at each satellite overpass time, only a small set of TB observations is available and many unknowns need to be estimated, including SM, soil temperature and vegetation water content. The baseline SMAP SM retrieval uses single polarization TB observations at a fixed incidence angle of 40° (Entekhabi et al. 2010a) to estimate SM, using a radiative transfer model (RTM) and ancillary land surface information, including soil temperature, soil texture and vegetation conditions, as constraints. Because SMOS collects TB observations at a range of incidence angles (Fernandez-Moran et al. 2017, Wigneron et al. 2007), the operational SMOS retrieval algorithms could theoretically estimate other unknown variables and parameters besides SM. However, because the radiometric errors of the multi-angular SMOS TB observations are relatively large compared to those of SMAP, and because the multi-angular observations do not actually provide entirely independent information, it is not possible to retrieve many additional unknowns along with SM. The multi-angular SMOS TB observations are in practice employed to estimate only vegetation optical depth (VOD) in addition to SM, while obtaining the remaining unknown variables and parameters from auxiliary data, as is done for SMAP retrievals.

Many retrieval products result from a physically-based Bayesian inversion constrained by a simplified RTM structure, ancillary information (e.g. soil temperature, vegetation), look-up table parameter values (e.g. for soil parameters, scattering albedo, etc.), and prior SM

estimates (e.g. for non-dominant land fractions or to constrain the retrieval). Because of the inaccuracies in these constraints and the propagation of instrument errors, the resulting SM estimates will be corrupted by lumped 'retrieval errors'. Quantifying these errors is not trivial, the errors may contain both random and persistent error (biases, Reichle et al. 2004), and they vary in space and time.

Persistent differences between various SM retrievals and other products can be removed, e.g., by statistical rescaling (Yilmaz and Crow, 2013). By contrast, the random uncertainty associated with SM retrievals is traditionally rather poorly identified, even though it is crucial when judging a data set's quality and even more so when merging data, e.g. in a data assimilation system. Most often, an uncertainty estimate is based on a time series comparison of SM retrievals to *in situ* observations where available (Albergel et al. 2012, Jackson et al. 2010), other retrieval products (Burgin et al. 2017, Al Yaari et al. 2019), modeling products (Al Yaari et al. 2014), or a combination of these data sets within triple collocation analysis (Draper et al. 2013, Gruber et al. 2016, Chen et al. 2017). Alternatively, sensitivity experiments and error propagation techniques can reveal the SM retrieval error due to errors in ancillary inputs (such as RTM parameters or dynamic surface temperature) or radiometric TB errors (Konings et al. 2011, Parinussa et al. 2011). However, error propagation techniques do not completely describe retrieval errors as they do not incorporate errors arising from an unavoidably incomplete model structure, or errors associated with the retrieval algorithm.

In this study, we evaluate various established and experimental SMOS SM products and their sensitivity to various design choices in the retrieval algorithm by comparing time series of retrievals to *in situ* observations. Next, we derive ensemble-based SM retrieval skill metrics that vary in space and time. Such spatiotemporally variable SM error estimates are important for optimal data merging, data assimilation systems, risk assessments, and other applications that currently assume spatiotemporally constant SM errors. For example, it matters to know whether a wet SM estimate is very uncertain or not when used as input to

subsequent flood prediction (Alvarez-Garreton et al. 2015). This study will evaluate to which extent SM errors are ergodic, i.e. whether errors sampled in time, space and across ensembles behave similarly with varying SM.

The ensemble SM retrieval generation and verification used here is inspired by techniques developed for ensemble forecast modeling (Cheung, 2001). By running multiple versions of a retrieval algorithm on the same SMOS TB dataset, an ensemble of retrieved SM time series can be obtained. The challenge here is to design the ensemble in such a way that its SM distribution correctly represents the retrieval error. Good or reliable ensembles are composed of members that are statistically indistinguishable from the truth (Fortin et al. 2014). The reliability of ensemble sets can be assessed with various ensemble verification tools, such as rank diagrams and spread-skill relationships. A key difference between ensemble forecasting and ensemble retrieval is that the former enables a stochastic propagation of errors in time, via a prognostic model, whereas no temporal error propagation or error growth is involved during the SM retrieval, which is constrained by a diagnostic RTM. Also, time series of ensemble SM retrieval members will be correlated across members, possibly even more so than ensemble SM estimates obtained from land surface model (LSM) simulations (Crow et al. 2015), and almost certainly more so than errors in ensemble weather forecasts from dispersive atmospheric models (Toth and Kalnay, 1993).

The overall goals of this study are to (i) evaluate the sensitivity of physically-based SMOS SM retrieval products to different TB configurations, RTM parameterization and prior SM estimates, and compare experimental retrievals to established ones through time-integrated skill metrics using *in situ* data as reference, and (ii) quantify the ensemble-based SM retrieval uncertainty, and the associated retrieval error variance decomposition, as a function of time and space. Section 2 introduces the SMOS data, ancillary data and reference data, Section 3 discusses the retrieval algorithm, Section 4 presents the generation of

ensemble retrievals, time series skill metrics and ensemble verification metrics. Section 5 gives an overview of the results and section 6 summarizes the conclusions.

2. Data

2.1. Study sites

The SM retrievals are evaluated at 11 reference sites, where footprint-scale *in situ* SM measurements are available in the top 5 cm of the soil column. The 11 sites are located in 7 watersheds across the USA (Table 1; Figure 1), and have been used extensively for SM retrieval evaluations (Chan et al. 2016, Bindlish et al. 2018). The spatial extent of each of these sites is defined by the overlying 36-km grid cells on the Global Cylindrical Projection of the Equal Area Scalable Earth Grid version 2 (EASEv2) (Brodzik et al. 2012, Brodzik et al. 2014), which differ slightly from the "matchup" validation grid cells of the SMAP core validation sites (Colliander et al. 2017). Table 1 lists for each site the dominant International Geosphere–Biosphere Programme (IGBP) land cover class, the soil class according to De Lannoy et al. (2014), and the maximum number of spatially distributed SM sensors within the grid cell. For each 36-km reference site, the individual sensor measurements are averaged into a single SM time series, and hourly data are only retained when no more than a single sensor would drop out, when the temperature is > 4 ^{o}C, when no snow or flooding is present, and when the corresponding satellite data are deemed of sufficient quality (see Section 2.2), for the period 1 June 2010 - 1 June 2015.

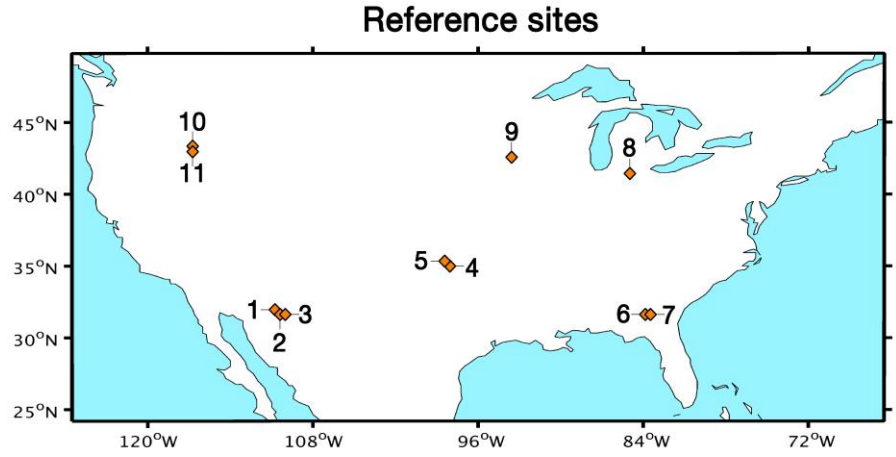


Figure 1. Reference sites with *in situ* SM time series located in the USA (see also Table 1).

Table 1. Overview of 11 reference sites with their label, watershed, USA state, vegetation class (vegcls; OS: open shrubland, G: grassland, CN: crop/natural mosaic, C: cropland), soil class (soilcls) according to De Lannoy et al. (2014), degree of vegetation heterogeneity within a site (based on optical vegetation information), start date of available grid-averaged data, and maximum number (N) of sensors measuring surface SM.

site	label	watershed	state	vegcls	soilcls	hetero-geneity	start date	max N
1	WG1	Walnut Gulch	Arizona	OS	209	high	2002/05/30	6
2	WG2	Walnut Gulch	Arizona	OS	184	high	2002/05/28	14
3	WG3	Walnut Gulch	Arizona	OS	208	high	2002/04/05	22
4	LW	Little Washita	Oklahoma	G	212	low	2007/01/01	15
5	FC	Fort Cobb	Oklahoma	G	136	low	2007/01/01	10
6	LR1	Little River	Georgia	CN	153	low	2004/07/24	15
7	LR2	Little River	Georgia	CN	153	low	2004/04/21	12
8	SJ	Saint-Joseph	Indiana	CN	211	low	2010/04/14	15
9	SF	South Fork	Iowa	C	195	low	2013/04/25	12
10	RC1	Reynold's Creek	Idaho	G	211	high	2002/03/20	6
11	RC2	Reynold's Creek	Idaho	G	209	high	2002/09/23	9

2.2. SMOS TB and established SM retrievals

Multi-angle Level 1C SMOS v620 TB observations are used as input to the retrieval algorithm to derive experimental SM retrievals in this paper. These TB data are preprocessed according to De Lannoy and Reichle (2016a), including quality filtering, and projecting the observed TBs onto the EASEv2 grid. The TBs are selected in 7 incidence angle bins, each of 1° width, centered on 30°, 35°, 40°, 45°, 50°, 55°, and 60°, for both horizontal (H) or vertical (V) polarization. The TB data are not fitted (i.e. unlike De Lannoy et al. 2015). In total, a maximum of 14 simultaneous TB observations (7 angles x 2 polarizations) are thus available

163 at a given overpass time and location. We use both ascending or descending orbital nodes,
164 with respective equator overpass times at 6 a.m. and 6 p.m. LT.

165 For comparison with the experimental retrievals of this study, we further use three
166 established (i.e. operational or widely used in research) SMOS SM retrieval products: the
167 SMOS L2 v620 (Kerr et al. 2016), SMOS-IC v103 (Fernandez-Moran et al. 2017, not yet
168 operational), and the SMOS retrievals of the ESA Climate Change Initiative (CCI) product
169 using the Land Parameter Retrieval Model (LPRM) (Van der Schalie et al. 2017). All data
170 are resampled to the EASEv2 36 km grid, which is close to the nominal 3 dB resolution of
171 about 40 km for SMOS. Only high-quality SM data are retained after a careful quality control
172 as described in De Lannoy and Reichle (2016a). For the specific study sites in this paper,
173 most of the SM retrieval data were retained, except during the winter at the northernmost
174 sites (RC, SJ, SF, Figure 1). The SMOS CCI LPRM dataset only provides retrieved SM at
175 ascending overpasses, whereas the other available datasets (established and experimental)
176 also comprise SM samples retrieved from descending nodes. Depending on the analysis, the
177 SM retrievals will thus either include both ascending and descending nodes, or be cross
178 masked to ascending orbits only for the common period 1 June 2010 - 1 June 2015.

179 *2.3. Land surface model SM and ancillary data*

180 Surface SM estimates are also simulated on the EASEv2 36 km grid with the Goddard
181 Earth Observing System version 5 (GEOS-5) Catchment Land Surface Model (CLSM,
182 Koster et al. 2000), using the same model settings as in De Lannoy and Reichle (2016a),
183 forced with Modern-Era Retrospective Analysis for Research and Applications (MERRA,
184 Rienecker et al. 2011) meteorology and gage-based precipitation corrections. The CLSM
185 model version used here is thus not the latest version, which is constantly being updated for
186 operational applications, such as e.g. the SMAP Level 4 products (Reichle et al. 2017). The
187 CLSM simulations and their input parameters are used in two ways: (i) surface SM output

(0-5 cm) is used for evaluation, and (ii) the seasonally varying climatological leaf area index input (LAI, Mahanama et al. 2015), soil texture input (De Lannoy et al. 2014) and dynamic surface soil temperature T_s output are taken as input to the experimental SM retrievals. Reichle et al. (2017) showed that the (absolute) bias and standard deviation of the errors in CLSM-based T_s estimates are less than 3 K at the reference sites used here.

3. Experimental soil moisture retrieval

3.1. τ-ω model

The τ-ω model (Mo et al. 1982) is a zero-order solution of radiative transfer equations and lies at the heart of most passive microwave SM retrieval algorithms. It is defined as:

$$TB_{p,a} = T_S\ e_{p,a}\Gamma_{p,a} + T_C\ \left(1-\Gamma_{p,a}\right)(1-\omega) + T_C\ \left(1-\Gamma_{p,a}\right)(1-\omega)\left(1-e_{p,a}\right)\Gamma_{p,a}, \quad (1)$$

with $TB_{p,a}$ [K] the top of canopy TB at polarization p (H or V) and incidence angle a (measured relative to nadir), T_s [K] the soil temperature, T_c [K] the vegetation temperature, $e_{p,a}$ [-] the soil emissivity, $\Gamma_{p,a}$ [-] the vegetation transmissivity and ω [-] the vegetation scattering albedo. The transmissivity $\Gamma_{p,a}$ is a function of the vegetation optical depth τ [-] at nadir and the incidence angle a, i.e. exp(-τ/cos(a)). Three TB components can be distinguished in Eq. (1): the upward emission by the soil surface which is attenuated by the vegetation layer (first term), the upward emission from the vegetation layer (second term), and the downward emission by the vegetation layer, reflected back upwards at the soil surface, and further attenuated by the vegetation layer (third term).

T_C is set equal to T_S in most operational SM retrieval algorithms (Wigneron et al. 2017), and also in our experimental retrievals. The soil emissivity ($e_{p,a}$) is modeled using the smooth reflectivity $R_{p,a}$, which depends on the dielectric constant and hence on SM (Calvet et al. 2011), the incidence angle a, and roughness parameters (h, N_{rp}):

$$e_{p,a} = 1 - R_{p,a}\exp(-h).\cos^{N_{rp}}(a) \quad (2)$$

212 The emissivity of the vegetation layer is expressed by the vegetation layer's absorptivity.
213 Indeed, emissivity, and absorptivity are identical under thermodynamic equilibrium
214 (Kirchoff 1860). The vegetation emissivity appears in the second and third terms of Eq. (1)
215 as that part of the vegetation attenuation not caused by scattering, i.e. $(1-\Gamma_{p,a})(1-\omega)$, where $1-$
216 $\Gamma_{p,a}$ signifies vegetation attenuation. The vegetation transmissivity $\Gamma_{p,a}$ ranges between 0 and
217 1, respectively indicating a vegetation layer completely opaque and completely transparent
218 for L-band microwave radiation.

219 The τ-ω model strongly simplifies radiative transfer processes, especially for areas with
220 denser vegetation. Structural errors can be partly compensated by choosing so-called
221 "effective" input parameters and parameterizations. The chosen ancillary inputs and
222 parameterizations of the above general τ-ω model unavoidably differ between our
223 experimental and established SMOS SM retrievals. For example, our experimental retrievals
224 utilize the CLSM soil temperature (T_s), whereas the established SMOS SM retrievals use
225 European Centre for Medium-Range Weather Forecasts (ECMWF) simulations for soil
226 temperature. We either calculate τ using a seasonally varying LAI climatology (Mahanama
227 et al. 2015, see below) or we retrieve τ together with SM using a TB polarization difference
228 approach. In contrast, all established SMOS products retrieve τ together with SM. We further
229 use the dielectric mixing model of Wang and Schmugge (1980) as in the SMOS CCI LPRM
230 retrievals, whereas the other established products all use Mironov et al. (2013). In line with
231 the operational SMOS L2 algorithm (Kerr et al. 2012), and unlike SMOS IC, we model the
232 soil roughness h as a function of SM, i.e. h decreases linearly between h_{max} and h_{min} as SM
233 values increase between the transition SM and soil porosity. These choices add uncertainty
234 to retrievals (e.g. soil temperature: Parinussa et al. 2016), but are not explicitly considered in
235 this paper.
236

3.2. τ-ω model parameterization

The main source of RTM uncertainty considered in this paper refers to the time invariant RTM parameters and the treatment of the vegetation. To account for the uncertainty in vegetation and roughness parameters in experimental SM retrievals, five common (spatially distributed) RTM parameter sets are tested to generate various sets of $TB_{p,a}^{sim}$ using Eq. (1). Each of these five parameter sets is combined with two options to estimate the vegetation optical depth τ, i.e. either using LAI input, or using the Microwave Polarization Difference Index (MPDI [-]; see below). The five sets of spatially distributed RTM parameters are based on literature:

- Lit1 values originate from the SMAP L2 and L3 SM retrieval algorithm (O'Neill, 2012). These are to a high degree inherited from the Hydros Observing System Simulation Experiment led by Crow et al. (2005), except for the vegetation class-dependent Leaf Equivalent Water Thickness (LEWT; Yilmaz et al. 2008 and references therein).
- Lit2 parameter values are compiled from literature using the Land Surface Microwave Emission Model (LSMEM), the L-band Microwave Emission of the Biosphere (L-MEB) Model, and the Community Microwave Emission Model (CMEM) (de Rosnay et al. 2009, Drusch et al. 2009, Grant et al. 2008, Wigneron et al. 2007, Saleh et al. 2007). These parameters are typically tuned over homogeneous areas (e.g. crop, Wigneron et al. 2007).
- Lit3 has all parameter values equal to those in Lit2, with the exception of N_{rp} (= 0), and the roughness parameters $h_{min} = h_{max} = 1.6$ adopted from the SMOS TB monitoring suite of the ECMWF using CMEM (Sabater et al. 2011).
- Lit4 values result from calibrating the forward RTM (using Eq. 3-4 below) to match observed SMOS TB for each SMOS pixel individually (De Lannoy et al. 2013). Such RTM parameters are also used in the SMAP Level 4 algorithm (Reichle et al. 2017).

- Lit5 values reflect the parameterization used in the SMOS-IC retrieval product (Fernandez-Moran et al. 2017). Unlike Lit2 and Lit3, these parameters are tuned for SM retrievals over heterogeneous SMOS pixels. Lit5 features higher $\omega = 0.1$ values for non-forested areas, also in line with the parameters used for SMOS CCI LPRM SM retrievals (Van der Schalie et al. 2016).

Lit1, Lit2 and Lit3 were earlier used in De Lannoy et al. (2013). The only difference is that the LEWT is adjusted in this paper, to compensate for a change in LAI climatology, while keeping the same level of Vegetation Water Content (VWC). Specifically, the LAI climatology used by De Lannoy et al. (2013) is from Advanced Very High Resolution Radiometer data and the LAI climatology used here is from Moderate Resolution Imaging Spectroradiometer data. With the exception of Lit4, all parameter sets are constant within each land cover class. It is important to note that Lit3 and Lit4 were originally designed for forward TB modeling only and not calibrated or validated for retrieving SM. This paper evaluates for the first time whether these parameter sets are suitable for retrieving SM.

The above five parameter sets are combined with two options to calculate τ. A first option uses ancillary vegetation index data, i.e. spatially distributed LAI [m²/m²] climatology values (Mahanama et al. 2015), to calculate the VWC [kg/m²]:

$$\mathrm{VWC} = \mathrm{LAI} \times \mathrm{LEWT} \quad [\mathrm{kg/m^2}], \tag{3}$$

with LEWT [kg/m²] depending on vegetation class. Values for τ_H and τ_V are derived from VWC and the polarization-dependent parameter b_p [$\mathrm{m^2/kg}$], so that:

$$\tau_p = b_p \times \mathrm{VWC} \quad [-]. \tag{4}$$

The second option is based on the LPRM algorithm, and does not rely on ancillary vegetation information. Instead, τ_H is estimated as a function (Meesters et al. 2005) of ω, the soil emissivity (e_H, e_V) and the Microwave Polarization Difference Index (MPDI [-]; Eq. 5) at an incidence angle a:

$$\mathrm{MPDI} = \frac{\mathrm{TB}_{\mathrm{V,a}}^{\mathrm{obs}} - \mathrm{TB}_{\mathrm{H,a}}^{\mathrm{obs}}}{\mathrm{TB}_{\mathrm{V,a}}^{\mathrm{obs}} + \mathrm{TB}_{\mathrm{H,a}}^{\mathrm{obs}}} \quad (5)$$

$$\tau_{\mathrm{H}} = f(\mathrm{MPDI}, \omega, \mathrm{e_H}, \mathrm{e_V}, \mathrm{a}). \quad (6)$$

Using τ_H, it is possible to calculate $\mathrm{TB}_{\mathrm{H,a}}^{\mathrm{sim}}$. When assuming $\tau_{H=}\tau_V$, $\mathrm{TB}_{\mathrm{V,a}}^{\mathrm{sim}}$ could also be simulated, but the latter is not used in this paper. In the original LPRM algorithm, soil roughness h is a function of τ_H, but this is not followed in our study in order to only focus on differences in τ estimation. Table 2 lists the key parameters for the 5 parameter sets, averaged over the 11 reference sites, and with an indication of which parameters are used along with either option to estimate τ.

Table 2. RTM parameters (h_{min}, h_{max}, b_V, b_H, ω, LEWT, N_{rH}, and N_{rV}, all unitless) averaged over the 11 reference sites for the 5 sets of spatially distributed parameter maps (Lit1, Lit2, Lit3, Lit4 and Lit5), ± standard deviation across the 11 reference sites. In case τ is calculated (option 1), more parameters are used in the SM retrieval than when is τ retrieved (option 2).

	Option 1 &2			**Option 1**					**Option 2**
	h_{min}	**h_{max}**	**ω**	**b_V**	**b_H**	**LEWT**	**N_{rV}**	**N_{rH}**	**N_{rV}=N_{rH}**
Lit1	0.13±0.02	0.13±0.02	0.05±0.01	0.11±0.01	0.11±0.01	0.28±0.03	0±0.00	0±0.00	1
Lit2	0.46±0.29	0.46±0.29	0.05±0.00	0.21±0.06	0.21±0.06	0.86±0.05	-0.36±0.50	0.64±0.50	1
Lit3	1.66±0.00	1.66±0.00	0.05±0.00	0.21±0.06	0.21±0.06	0.86±0.05	0±0.00	0±0.00	1
Lit4	0.19±0.21	0.53±0.29	0.05±0.04	0.23±0.12	0.29±0.10	0.86±0.05	-0.36±0.50	0.64±0.50	1
Lit5	0.17±0.04	0.17±0.04	0.10±0.02	0.21±0.06	0.21±0.06	0.86±0.05	-1±0.00	-1±0.00*	1

* N_{rH} is set to 1 in SMOS IC version 105.

3.3 Optimization scheme

For each reference site and overpass time, SM is retrieved as the SM value that results in the best match between simulated $\mathrm{TB}_{\mathrm{p,a}}^{\mathrm{sim}}$ (Eq. 1) and SMOS observed $\mathrm{TB}_{\mathrm{p,a}}^{\mathrm{obs}}$ by simply minimizing a cost function CF(SM) for a pre-selected range of SM values $\in \{0.005, 0.010, 0.015, 0.020, \ldots, 1\}$ [m³/m³]:

$$\mathrm{CF(SM)} = \begin{bmatrix} \mathrm{TB}_{\mathrm{H},a_1}^{\mathrm{obs}} - \mathrm{TB}_{\mathrm{H},a_1}^{\mathrm{sim}}(x,\mathrm{SM}) \\ \mathrm{TB}_{\mathrm{H},a_2}^{\mathrm{obs}} - \mathrm{TB}_{\mathrm{H},a_2}^{\mathrm{sim}}(x,\mathrm{SM}) \\ \vdots \\ \mathrm{TB}_{\mathrm{V},a_n}^{\mathrm{obs}} - \mathrm{TB}_{\mathrm{V},a_n}^{\mathrm{sim}}(x,\mathrm{SM}) \end{bmatrix}^{\mathrm{T}} \left(\sigma_{\mathrm{TB_{obs}}}^2 \mathrm{ECM}\right)^{-1} \begin{bmatrix} \mathrm{TB}_{\mathrm{H},a_1}^{\mathrm{obs}} - \mathrm{TB}_{\mathrm{H},a_1}^{\mathrm{sim}}(x,\mathrm{SM}) \\ \mathrm{TB}_{\mathrm{H},a_2}^{\mathrm{obs}} - \mathrm{TB}_{\mathrm{H},a_2}^{\mathrm{sim}}(x,\mathrm{SM}) \\ \vdots \\ \mathrm{TB}_{\mathrm{V},a_n}^{\mathrm{obs}} - \mathrm{TB}_{\mathrm{V},a_n}^{\mathrm{sim}}(x,\mathrm{SM}) \end{bmatrix}$$

$$+ \alpha \frac{\left(\mathrm{SM}_{\mathrm{prior}} - \mathrm{SM}\right)^2}{\sigma^2_{\mathrm{SM}_{\mathrm{prior}}}} \quad (8)$$

with the first term referring to the observation likelihood and the second term referring to the prior constraint, which can be included (α=1) or not (α=0). Furthermore, $\sigma_{\mathrm{TB}_{\mathrm{obs}}}$[K] is the TB observation error standard deviation, ECM [-] is the residual TB error correlation matrix, $\mathrm{SM}_{\mathrm{prior}}$ [m³/m³] is the prior guess and $\sigma_{\mathrm{SM}_{\mathrm{prior}}}$ [K] is the prior SM error standard deviation. The total number *C* of observed-minus-simulated TB pairs is at least 1 and at most 14, depending on the selected number of TB species. A species refers to a TB estimate at a specific combination of polarization p (H- or V-polarization) and incidence angle a (one of 7 different angles). $\mathrm{TB}^{\mathrm{sim}}_{\mathrm{p,a}}(x, \mathrm{SM})$ signifies the TB simulated using parameterization *x* (five parameter maps and two options for vegetation parameterization, Section 3.2) and using a particular input SM.

$\sigma_{\mathrm{TB}_{\mathrm{obs}}}$[K] includes both the SMOS radiometer instrument error of 4 K for every TB species, as well as the TB representativeness error, which includes discrepancies between modeled and observed TB due to sampling volume mismatches and due to RTM error. The latter should theoretically vary with the model structure and parameterization (Section 3.2) but is set to 4.5 K here for simplicity (De Lannoy and Reichle, 2016a,b), so that $\sigma_{\mathrm{TB}_{\mathrm{obs}}} = \sqrt{(4\ \mathrm{K})^2 + (4.5\ \mathrm{K})^2} = 6\ \mathrm{K}$.

The ECM reflects the, probably, correlated nature of the errors in various TB deviations. The $\mathrm{TB}^{\mathrm{sim}}_{\mathrm{p,a}}$ in particular may exhibit error correlations between TB at different incidence angles and polarizations. Such error correlations will reduce the weight of the observational term relative to the prior term in the CF (if the latter is included) and they restrain the increase in the observation likelihood weight with increasing numbers of TB species. Specifically, we assume an error correlation of 0.5 between TB species for a unit distance in incidence angle or polarization. Our choice for this value of 0.5 accounts for the fact that the TB deviations

include *both* TB measurement and forecast errors, where the former are assumed to be not correlated and the latter likely highly correlated. A difference of incidence angle of 5° between two TB species i and j is considered as one distance unit $d_{i,j}$. A difference in polarization between any pair of TB species counts as an additional unit of distance. The cross-correlation $\rho_{i,j}$ between the residual time series associated with TB species i and j is modeled to geometrically decrease with increasing $d_{i,j}$:

$$\rho_{i,j} = (0.5)^{d_{i,j}}, \quad (9)$$

In addition to assessing the sensitivity to RTM parameters (x), this study will thus look into two design aspects of the retrieval algorithm: the TB configuration, i.e. number of incidence angles and polarizations (C) and the prior SM constraint (α).

4. SM sensitivity and ensemble retrievals

4.1 SM retrieval skill and sensitivity

4.1.1 SM retrieval realizations

By varying select aspects of the retrieval algorithm, a range of SM retrievals is generated, always constrained by a deterministic relationship between TB and SM. The latter constraint can be strong (RTM is assumed to be perfect) or weak (RTM parameters are perturbed).

First, the sensitivity of the SM skill to the TB configuration (C) in the CF is studied via experimental retrievals using either both polarizations and all incidence angles, or combinations of single-angle single-polarization and multi-angle single-polarization TB. For reference, in the SMOS L2 and SMOS IC retrievals, the multi-angle and dual-polarization TB observations are simultaneously used (Kerr et al. 2012, Fernandez-Moran et al. 2017) assuming zero cross correlations of TB errors. In the SMOS CCI LPRM product, retrievals are an average of scaled retrievals at each individual angle (after quality control). Retrievals derived from a single TB species (i.e. only one polarization and a single incidence angle) are

typically not used for SMOS SM retrievals, because of the high error for individual TB measurements, but this approach is used for SMAP SM retrievals and also for other missions, such as the Advanced Microwave Scanning Radiometer - Earth Observing System (AMSR-E/2) (Imaoka et al. 2012).

Second, the effect of the prior SM guess (SM_{prior},$\sigma_{SM_{prior}}$) on the Bayesian inversion is quantified. The first case (α=0) effectively sets the prior term in $CF(SM)$ to zero. This unconstrained case mimics the settings of operational SMAP L2, SMOS L2 and SMOS CCI LPRM retrievals, where either no prior SM estimates are involved in the CF minimization, or the prior constraint is technically included but effectively turned off by using a large error standard deviation. The SMOS IC retrieval algorithm uses a constant prior SM of 0.2 m^3/m^3 with an intentionally large uncertainty $\sigma_{SM_{prior}}$=0.2 m^3/m^3 (Fernandez-Moran et al. 2017). In a second case (α=1), spatio-temporally varying CLSM SM is used as prior SM with $\sigma_{SM_{prior}}$ =0.02 m^3/m^3, i.e. with a relatively low error estimate that balances the 6 K observation error to ensure a stronger impact of the prior SM on the final retrieval.

Third, different RTM parameterizations (x) are used, i.e. 5 RTM parameter sets are used along with calculating τ as input (option 1) and 5 RTM parameter sets are tested with retrieval of τ (option 2, Section 3.2). For the case where τ is calculated as input (option 1), the model constraint is further weakened by perturbing each of the 5 RTM parameter sets. This is done by multiplying each one of 4 parameters (i.e. h_{min}, or h_{max}, or b_v&b_h simultaneously, or ω) by either ×0.50, ×0.75, ×1.25, or ×1.50, i.e. decreasing and increasing the parameter values by 25% and 50% of their original values. This leads to 17 realizations (1 without and 16 with perturbations) for each of the 5 RTM parameter sets. The corresponding 17 SM realizations associated with each of the 5 RTM parameters could be considered an SM "ensemble". However to avoid confusion with the verified SM ensembles (Section 4.2.1) we will refer to (averages of) multiple SM "realizations" in the sensitivity study.

380 4.1.2 Time series skill metrics

381 The similarity between SM retrieval time series v(t) and *in situ* ground data *w*(t)
382 (t=1,…N), or the SM retrieval skill, is quantified by four skill metrics, commonly used in
383 performance assessments (Entekhabi et al. 2010b). Even though the reference *in situ* data are
384 of good quality, it is important to realize that they are not the 'truth' and also contain errors.
385 The bias B quantifies the long-term mean error (Eq. 10), the unbiased Root Mean Squared
386 Difference ubRMSD (Eq. 11) captures the remaining error, and the correlation coefficient R
387 (Eq. 12) and the anomaly correlation coefficient Ranom (Eq. 13) focus on similarity in time
388 series variability:

389 $$\mathrm{B} = \frac{\sum_{t=1}^{\mathrm{N}}(\mathrm{w}(t)-v(t))}{\mathrm{N}}, \tag{10}$$

390 $$\mathrm{ubRMSD} = \sqrt{\frac{\sum_{t=1}^{\mathrm{N}}(\mathrm{w}(t)-v(t))^2}{\mathrm{N}} - \mathrm{B}^2}\ , \tag{11}$$

391 $$\mathrm{R} = \frac{\sum_{\mathrm{t}=1}^{\mathrm{N}}(\mathrm{w}(\mathrm{t})-\overline{\mathrm{w}})(\mathrm{v}(\mathrm{t})-\overline{\mathrm{v}})}{\sum_{\mathrm{t}=1}^{\mathrm{N}}(\mathrm{w}(\mathrm{t})-\overline{\mathrm{w}})^2 \sum_{\mathrm{t}=1}^{\mathrm{N}}(\mathrm{v}(\mathrm{t})-\overline{\mathrm{v}})^2}\ , \tag{12}$$

392 $$\mathrm{Ranom} = \frac{\sum_{\mathrm{t}=1}^{\mathrm{N}}(\mathrm{w}_{\mathrm{anom}}(\mathrm{t})-\overline{\mathrm{w}}_{\mathrm{anom}})(\mathrm{v}_{\mathrm{anom}}(\mathrm{t})-\overline{\mathrm{v}}_{\mathrm{anom}})}{\sum_{\mathrm{t}=1}^{\mathrm{N}}(\mathrm{w}_{\mathrm{anom}}(\mathrm{t})-\overline{\mathrm{w}}_{\mathrm{anom}})^2 \sum_{\mathrm{t}=1}^{\mathrm{N}}(\mathrm{v}_{\mathrm{anom}}(\mathrm{t})-\overline{\mathrm{v}}_{\mathrm{anom}})^2}\ . \tag{13}$$

393 The Ranom is calculated between the retrieved time series anomalies $v_{\mathrm{anom}}(t)$ and the *in*
394 *situ* time series anomalies $\mathrm{w}_{\mathrm{anom}}(t)$. Single overbars (as in $\overline{v}$, $\overline{\mathrm{w}}$, $\overline{v}_{\mathrm{anom}}$ and $\overline{\mathrm{w}}_{\mathrm{anom}}$) indicate
395 time-averaged values. The metrics are calculated for each of the 11 reference sites
396 individually, accompanied by a 95% confidence interval, and then spatially averaged across
397 the 11 sites, similarly as in De Lannoy and Reichle (2016b). The anomalies are calculated
398 relative to the multi-year, 30-day moving window, average SM and thus reflect both short-
399 term and inter-annual deviations from the long-term mean seasonal cycle. These (zero-mean)
400 anomalies differ from the zero-mean errors quantified in the ubRMSD, for which only the
401 constant long-term mean is removed.

The skills are computed for three established SMOS SM retrieval products (Section 2.2), CLSM simulated SM (Section 2.3) and various experimental SM retrievals. Furthermore, the time series skill of multi-realization average SM retrievals is investigated. In theory (Appendix A), when the errors in K realizations are uncorrelated, then the ubRMSD for an average retrieval would reduce to $1/\sqrt{K}$ of the ubRMSD for a single realization (for large K). The use of average multi-realization retrievals would thus be an interesting approach to reduce the uncertainties of satellite retrievals. However, due to the multiple constraints in the retrieval algorithm, SM retrieval realizations are always partly correlated, which limits the potential of the average retrievals to decrease the ubRMSD.

4.2 Ensemble uncertainty

4.2.1 Candidate ensembles and verification

The time series metrics above quantify temporally averaged errors, using *in situ* data as reference. These metrics aim at describing the probability density function of a retrieval at a given location by sampling realizations in time. We now attempt to sample the ensemble probability density function (EPDF) at each time step *t* (when observed SMOS TB is available to perform a SM retrieval) via a realistic, or `verified', K-member SM retrieval ensemble, which subsequently will allow us to perform an error variance decomposition and derive time-varying ensemble metrics without using reference data. The ensemble standard deviation is an estimate of the ensemble SM uncertainty, if the EPDF is reliably sampled, which is hard to guarantee. However, summary information on time series performances can be used to identify candidate ensembles, which possibly represent the SM retrieval uncertainty reliably. A specific ensemble of SM retrievals is assumed to reliably represent retrieval uncertainty when the corresponding reference *(in situ)* SM time series is not statistically distinct from the set of ensemble member time series (Anderson and Anderson 1999). To verify this condition, three ensemble verification tools, well-known in ensemble

forecasting, are adopted in this study: two spread-skill relationships (Brankovic et al. 1990, Talagrand, et al. 1997), and the rank histogram (Hamill 2001). The verification and selection of a reliable SM ensemble is discussed in Appendix B.

The ensemble selected for further analysis consists of K=85 member SM retrieval time series, created with 85 (=5×(16+1)) different RTM parameter sets (x). Since the ensemble variability caused by perturbing any single RTM parameter set was found insufficient, all 5 RTM parameter sets are used simultaneously in the creation of the candidate ensembles, and each of 4 RTM parameters (i.e. h_{min}, h_{max}, b_v&b_h simultaneously, and ω) is multiplied separately by ×0.50, ×0.75, ×1.25, or ×1.50, yielding 4×4=16 perturbed parameter values in addition to the default one for each parameter set. The selected SM ensemble uses all 14 TB species simultaneously in the CF and calculates τ as function of climatological LAI (option 1).

The selected ensemble is further processed to derive a zero-centered ensemble. The various ensemble members are strongly biased relative to each other, because each member's climatology is determined by its associated RTM parameters. The various ensemble members are therefore centered by removing their respective long-term means, yielding anomaly members. The aim of centering the SM retrieval members is to remove first-order bias and to investigate the remaining error (see Section 4.2.3, Figure 2b).

4.2.2 Ensembles mean and spread

We formulate the first two EPDF moments diagnosed from a K-member SM retrieval ensemble, both (i) as a function of time *t*, and (ii) temporally averaged over N time samples (marked by overbars, $\overline{\cdot}$). The ensemble mean enmn (Eq. 14) and the ensemble variance envar (Eq. 15) are defined as follows:

$$\mathrm{enmn}(t) = \frac{1}{\mathrm{K}}\sum_{m=1}^{\mathrm{K}} v_{\mathrm{m}}\,(t); \qquad \overline{\mathrm{enmn}} = \frac{1}{\mathrm{N}}\sum_{t=1}^{\mathrm{N}} \mathrm{enmn}(t); \tag{14}$$

$$\mathrm{envar}(t) = \frac{1}{\mathrm{K}}\sum_{m=1}^{\mathrm{K}}\left(v_{\mathrm{m}}(t) - \mathrm{enmn}(t)\right)^2; \qquad \overline{\mathrm{envar}} = \frac{1}{\mathrm{N}}\sum_{t=1}^{\mathrm{N}}\mathrm{envar}(t), \tag{15}$$

with $v_{\mathrm{m}}(t)$ the SM value associated with ensemble member m at instant *t*. The ensemble standard deviation $\sqrt{\mathrm{envar}(t)}$ or $\sqrt{\overline{\mathrm{envar}}}$ will be used as a metric for uncertainty and will be referred to as 'ensemble spread' in the following. These metrics are calculated for individual reference sites, averaged in time, and also averaged across all 11 sites to describe the spatiotemporally-lumped EPDF (defined as $\overline{\overline{\mathrm{EPDF}}}$, where the double overbar refers to space-time lumping). The higher order moments of $\overline{\overline{\mathrm{EPDF}}}$ will also be studied to assess whether simple SM retrieval uncertainty estimates in terms of standard deviations are justified by Gaussian distributions or not.

4.2.3 Ensemble SM retrieval uncertainty and error variance decomposition

Because SM retrieval ensembles are not dispersive and the individual members have their own climatology by design, the total ensemble uncertainty can be partitioned into a long-term-mean error and an anomaly error variance. The variability in the long-term-mean values of each ensemble member can be defined as the offset error. The variability in the remaining ensemble anomaly, or zero-centered, members (after removal of the long-term mean) is mainly due to the non-linear propagation of RTM parameter errors (in our case) and is thus closer to the random error uncertainty. However, the anomaly error component also includes systematic errors, e.g. differences in variances, between ensemble members.

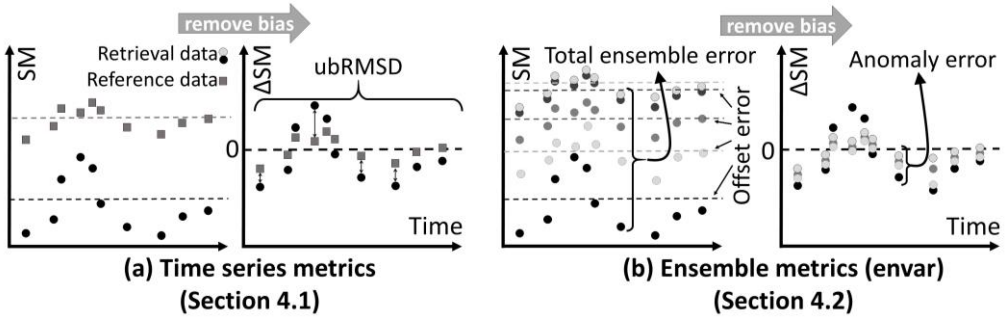


Figure 2: (a) Time series metrics relative to reference data and (b) ensemble error partitioning.

This partitioning into the offset error and the anomaly error components is also done in time series skill metrics (Section 4.1.2). For example, the target uncertainty of 0.04 m^3/m^3 for ~40 km L-band SM retrievals over areas with limited vegetation is typically defined as the ubRMSD between time series of *in situ* and retrieved SM (Entekhabi et al. 2014, Kerr et al. 2016). That is, the target uncertainty excludes the long-term mean bias. Consequently, both the ubRMSD and the (time-average) ensemble variance of zero-centered (anomaly) ensembles are by definition free of bias. Therefore the two metrics should be comparable provided the ensemble accurately represents the actual retrieval errors. This conceptual link between ensemble uncertainty metrics (without reference data) and time series skill metrics (ubRMSD relative to reference data) is illustrated in Figure 2 and provides the basis for using ensemble-based uncertainty metrics to estimate retrieval errors which are otherwise computed vs. *in situ* data across time or space. A comparison of ensemble statistics to temporal or spatial metrics will verify if the ergodicity principle holds, i.e. if the ensemble uncertainty metrics can approximate typical temporal or spatial error estimates.

5. Results

5.1 SM retrieval skill and sensitivity

The skill of three established retrieval SM products and CLSM SM estimates are shown for 10 of the 11 reference sites in Figure 3 (RC2 is not shown, for lack of enough data pairs). The ubRMSD, R, Ranom and bias are shown at the individual sites with >75 data pairs at 6:00 LT. The metrics are also averaged across all sites, only the sites with high heterogeneity and only the sites with low heterogeneity (hereafter referred to as “homogenous” sites; see Table 1). As can be expected, the overall performance is better for the homogenous sites. In terms of ubRMSD, SMOS IC performs best, followed by SMOS L2, CLSM, and SMOS CCI LPRM. However, in terms of R and Ranom, SMOS LPRM CCI performs best, after SMOS IC (second best), SMOS L2 and CLSM (worst). Bias is present in all SM products. However,

when averaged over all sites, bias is nearly absent in SMOS L2 and SMOS CCI LPRM, while SMOS IC SM is biased dry (about -0.025 m^3/m^3), and CLSM SM biased wet (about +0.060 m^3/m^3). Except for bias, SMOS IC outperforms SMOS L2, despite the simplicity of the SMOS IC algorithm. CLSM SM performs significantly worse than the established satellite-based SM retrievals in terms of R and Ranom. Note that values for Ranom are missing at SF and RC1, because of limited available data to compute a robust climatology.

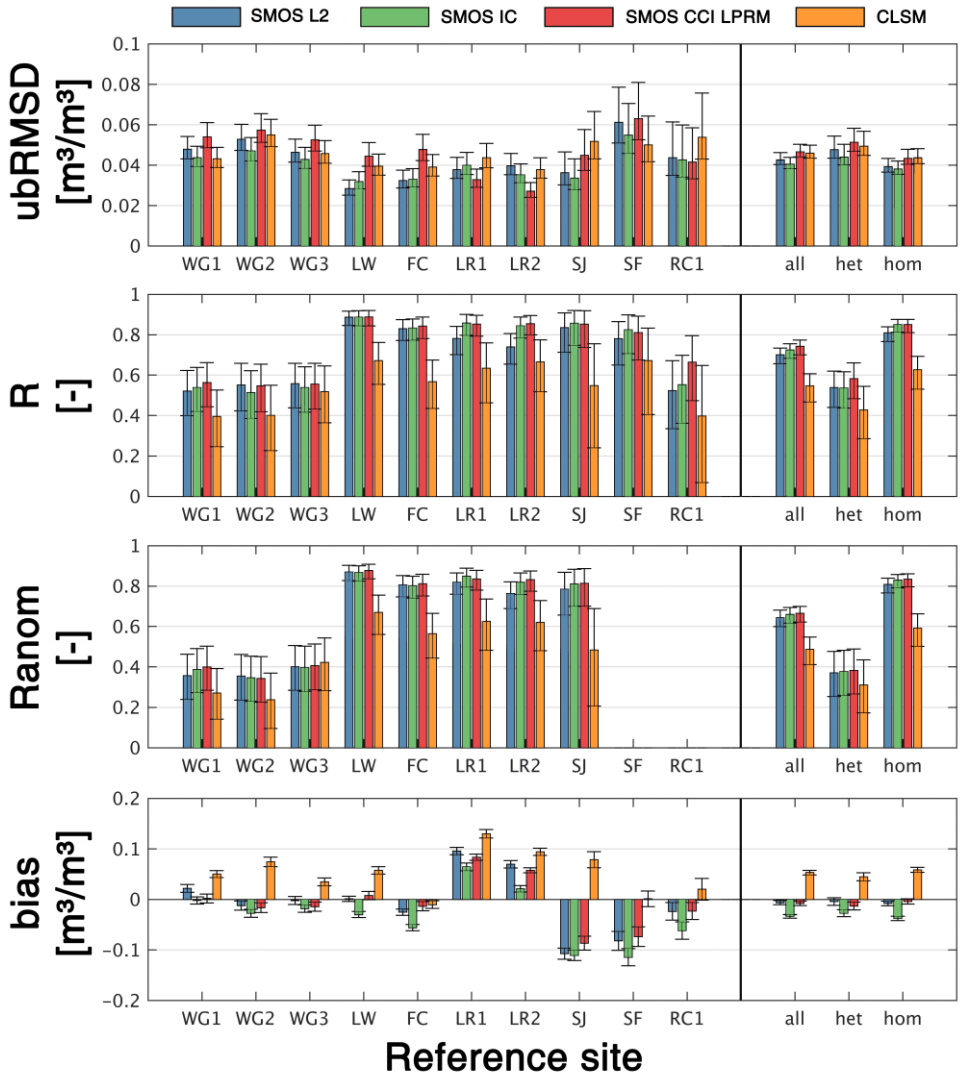


Figure 3. Skill comparison of three established SM retrieval products (i.e. SMOS L2 v620, SMOS IC v103, SMOS LPRM CCI) and CLSM SM during the period 1 June 2010-1 June 2015 (6:00 am LT only), in terms of ubRMSD, R, Ranom, and bias, with 95% confidence intervals. Skills are arranged per reference site and also averaged over all the sites (all), the high heterogeneity sites (het), and the low heterogeneity sites (hom); see Table 1 for details.

Next, the performance of various experimental retrievals is evaluated. First, the impact of TB configurations (incidence angle, polarization) and RTM parameterization is assessed in Figure 4. At each reference site, SM time series are retrieved for 5 RTM parameter maps. Retrievals are created for each single incidence angle and a single polarization, using (i) option 1 to calculate τ (both for H- and V-polarization), or (ii) option 2, retrieving τ (only for H-polarization). The results are presented for retrievals without the prior SM term in the CF.

For most RTM parameter sets, except Lit3, Figure 4 shows that calculating τ as input (option 1) is generally better than retrieving τ (option 2). Lower incidence angles perform better with option 1 (τ1_H, τ1_V), whereas the mid-range incidence angles yield better results with option 2 (τ2_H). The SM retrievals for τ1_H are slightly more skillful than for τ1_V. Surprisingly, the experimental τ1_H and τ1_V retrievals using single (and low) incidence angles can reach skill levels that are very close to that of the established retrievals; but note the sampling difference between the experimental (6:00 am & 6:00 pm LT) and established retrievals (6:00 am LT only) in Figure 4.

The various parameter sets cause large differences in bias (differences in climatology), which can be an order of magnitude larger than that of the ubRMSD. A simple regression analysis shows that the time-mean retrieved SM strongly correlates (0.83, $p<0.001$) with the value of the time-mean roughness parameter h, i.e. wetter SM is found for higher mean h values. The climatological mean SM also strongly correlates with τ values and weakly correlates with the scattering albedo ω. High R and Ranom, and low ubRMSD values are often associated with Lit1 and Lit5, which have low roughness values h in common. The Lit3 and Lit4 parameters were tuned for TB forward modeling. The retrievals using Lit3 parameters are performing worst, in terms of all metrics. In contrast, Lit4 parameters yield comparable SM results to the Lit1, Lit2 and Lit5 parameters that were tuned for SM retrievals.

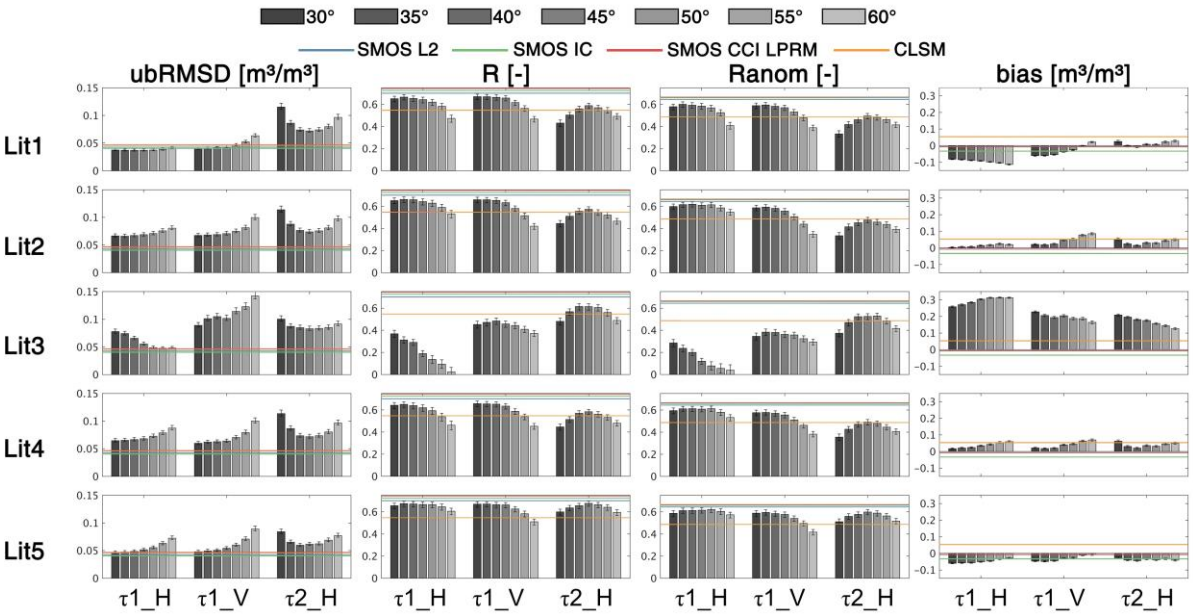


Figure 4. Sensitivity of retrieval skill to TB incidence angle, polarization, and RTM parameterization, assuming no prior SM constraint in the CF. All metrics are calculated against *in situ* SM data, averaged over the 11 reference sites (except for Ranom, not available for RC1, RC2, SJ) during 1 June 2010 – 1 June 2015 (6:00 am and 6:00 pm LT). The skill of the established retrievals and CLSM SM at 6:00 am is shown for reference (last bars of Fig. 3 here shown as lines).

Figure 5 compares SM retrievals as function of TB configuration and RTM parameter sets, while calculating τ as input (option 1 only), with and without the prior SM constraint in the CF. The bars in the figure present the average skill of multiple SM retrieval realizations resulting from perturbations of the individual RTM parameter sets (17 realizations for each of the 5 RTM parameter maps). Moreover, the symbols in Figure 5 show the average skill of an unperturbed SM retrievals, which could be an average of 14 realizations (case i) or just a single realization (all other cases). The TB configuration consists of either (i) a single TB species (one polarization, one angle; 17x2x7 retrievals), (ii) all 7 H-polarized TB species (17 retrievals), (iii) all 7 V-polarized TB species (17 retrievals), or (iv) all 14 H- and V-polarized TB species (17 retrievals). The multi-realization averages obtained by perturbing the RTM parameters perform slightly better in terms of ubRMSD than when not perturbing the RTM parameters. However, the ubRMSD for averages over multiple realizations is not reduced by $1/\sqrt{K}$, because the retrieval errors are correlated (Appendix A).

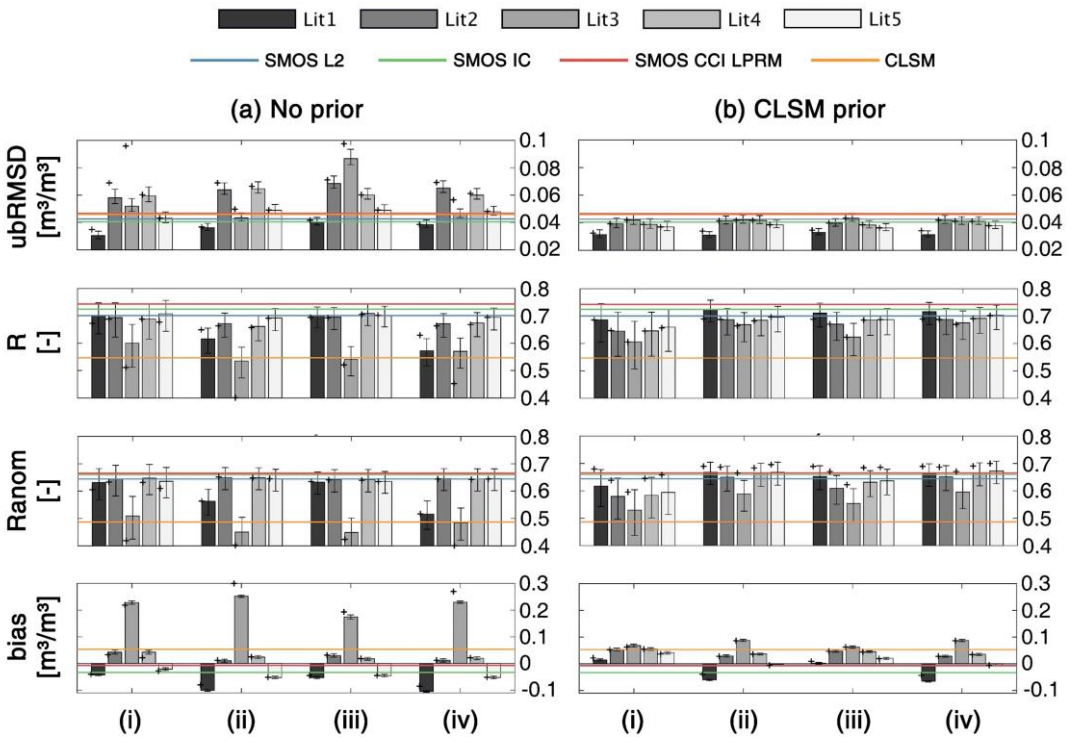
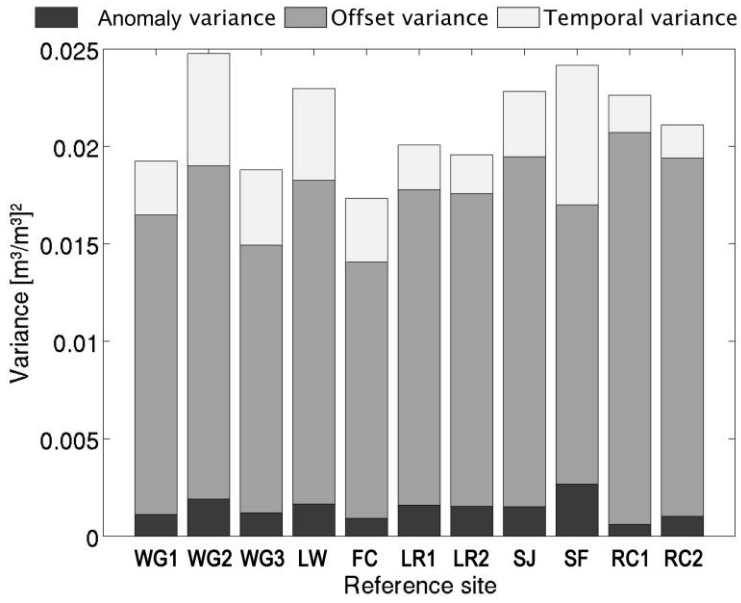


Figure 5. Skill of (bars) average retrievals obtained by perturbing RTM parameters and (+) retrievals without RTM parameter perturbation, for various TB configurations and RTM parameter sets calculating τ as input (option 1 only), (a) without prior and (b) with CLSM SM as prior constraint. The TB configurations include (i) 2x7 individual TB species, (ii) one set of 7 TbH species, (iii) one set of 7 TbV species and (iv) one set of 7 TbH and 7 TbV species jointly. All metrics are averaged over 11 reference sites during 1 June 2010 – 1 June 2015 (6:00 am and 6:00 pm LT), and the error bars are 95% confidence intervals. The skills of the established retrievals and CLSM SM at 6:00 am are shown for reference (last bars of Fig. 3 here shown as lines).

Overall, the TB configuration has only a small impact on the skill metrics. The multi-angular dual-polarization retrievals (case iv) do not outperform the other retrievals. Regardless of the TB configuration, the ubRMSD of the experimental retrievals is typically worse than the established products, except for Lit1, which results in the lowest ubRMSD. In general, the ubRMSD is better for parameter sets with low h values, i.e. Lit1 and Lit5. As shown before, the results with Lit3 (high roughness) are worst in all metrics. However, when including a prior constraint of CLSM SM (Figure 5b), all retrievals can be greatly improved. The best results are obtained using dynamically varying CLSM SM, but a time-mean CLSM SM as prior constraint also effectively improves the retrieval results over what is obtained without a prior constraint (not shown).

576 *5.2 Ensemble uncertainty*



577

578 **Figure 6.** Anomaly variance, offset variance and temporal variance in SM retrievals
579 at 11 reference sites (1 June 2010 - 31 May 2015). The anomaly and offset variances
580 are calculated across the ensemble at each time step and then averaged over time,
581 the temporal variance is calculated for the ensemble mean over the same time period.

582 Figure 6 partitions the total ensemble variance of the 85-member ensemble without zero-
583 centering (Section 4.2.1 & Appendix B) into the variance in the long-term means of the
584 members (named 'offset variance' here, medium gray bars) and the variance in the anomaly
585 members (dark gray bars). The latter is identical to the variance in the zero-centered
586 ensemble. The offset variance is much larger than the anomaly ensemble variance at all sites.
587 Figure 6 also shows that the variation in long-term mean values greatly exceeds the temporal
588 SM variability at each location (light gray bars). The time-averaged ensemble anomaly
589 spread (standard deviation) ranges between 0.025 m^3/m^3 at RC1 and 0.051 m^3/m^3 at SF, with
590 a mean of 0.037 m^3/m^3 and a standard deviation of 0.007 m^3/m^3 across sites. Time series of
591 the zero-centered ensemble underlying these time-averaged statistics are visualized in Figure
592 7 (grey dots), together with the ensemble mean (blue dots) and the centered *in situ* SM time
593 series (red dots), for four reference sites. The *in situ* data are generally well embedded inside
594 the ensemble envelope, indicating that retrieval uncertainty is well represented by this
595 ensemble (confirmed by Figure B1c). No spatial properties, such as soil porosity, wilting

point, RTM roughness h or vegetation parameter b, could directly explain the limited spatial variability of the time-averaged anomaly ensemble spreads (not shown) and hence it is justified to combine all sites in the following analyses.

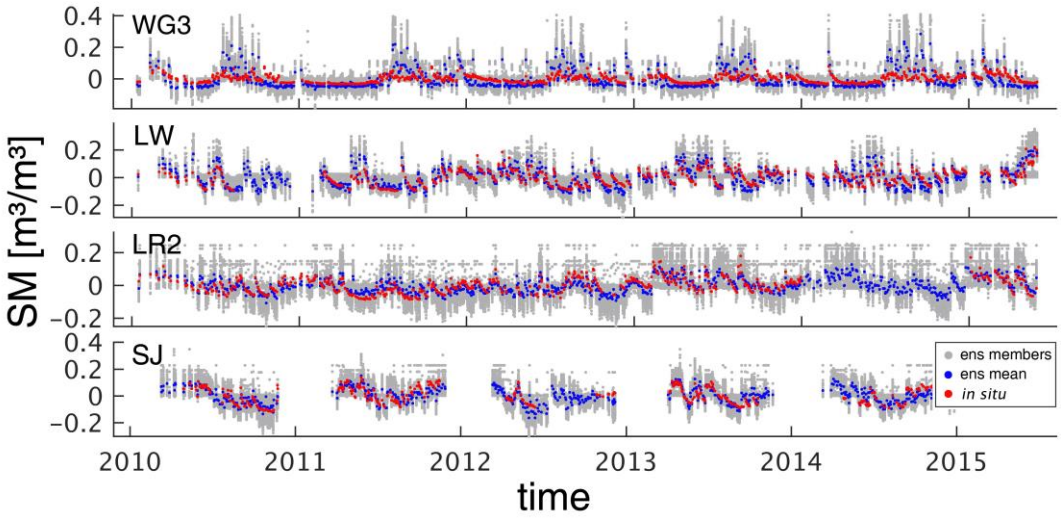


Figure 7. Zero-centered SM retrieval time series at four reference sites: (gray dots) 85-member SM ensemble without prior SM constraint, (blue) associated ensemble mean time series, and (red) *in situ* SM time series, for the period 1 January 2010 - 1 June 2015. Gaps in the ensemble time series are caused by a lack of good SMOS TBs, gaps in *in situ* SM time series are mainly caused by failing sensors.

Figure 8 shows that the spatiotemporally-averaged or lumped ensemble distribution of the zero-centered SM ensemble set is nearly Gaussian, not skewed, and only slightly leptokurtic (i.e. having a positive kurtosis; more concentrated about the mean). The space-time mean anomaly ensemble standard deviation is 0.037 m^3/m^3.

Instead of lumping the ensemble information in space and time, Figure 9 illustrates how the ensemble spread varies with the ensemble mean SM in space and time. The total ensemble spread without zero-centering (Figure 9a) is around 0.07 m^3/m^3 for dry SM, increases to about 0.17 m^3/m^3 for moderate SM and decreases again for wet SM. By contrast, the anomaly ensemble spread (i.e. with zero-centering) is lowest around the long-term mean SM and increases for both lower and higher SM values (Figure 9b). The anomaly spread varies between 0.01 and 0.12 m^3/m^3, with an average of 0.037 m^3/m^3 (Figure 8) and a temporal standard deviation of 0.015 m^3/m^3, averaged across sites. A high proportion (i.e. 62%; based on R=0.79) of this variation can be explained by the absolute value of the zero-centered

enmn(t). Put differently, the anomaly spread increases with the deviation of SM from its long-term mean value.

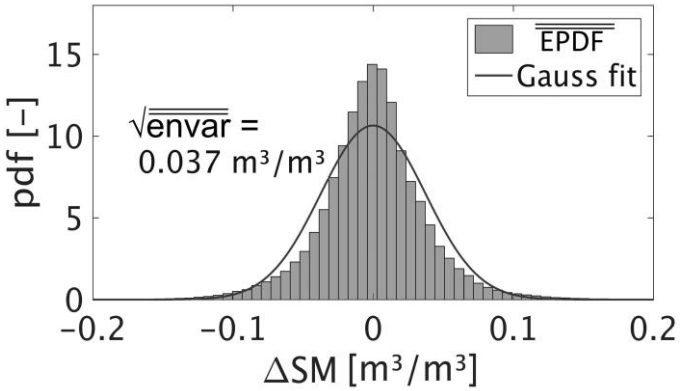


Figure 8. Temporally and spatially lumped centralized ensemble SM probability density distribution. The x-axis shows the ensemble member deviations ΔSM from the instantaneous ensemble mean. Data points are sampled over 85 ensemble members, 1 June 2010 - 1 June 2015, and over 11 reference sites. A Gaussian fit is shown as a line.

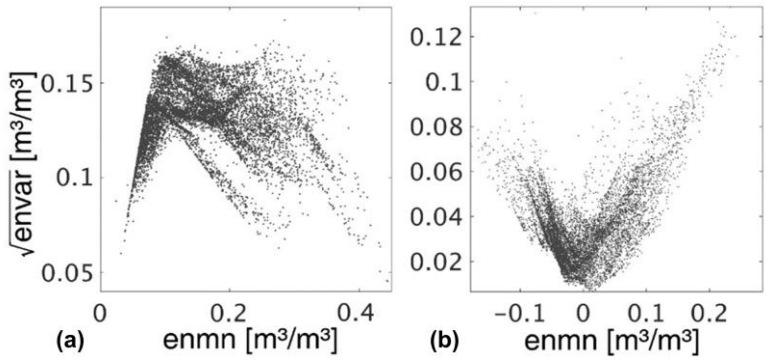


Figure 9. Ensemble standard deviation √envar(t) in function of ensemble mean enmn(t) (a) without and (b) with zero-centering the SM members, for all 11 reference sites, and all retrieval time steps in the period 1 June 2010 - 1 June 2015.

6. Discussion

6.1 SM retrieval skill and sensitivity

When comparing various established SMOS SM retrievals, better performance is generally found over homogenous sites, and the SMOS IC product performs surprisingly well when considering the simplicity of the algorithm compared to that of the SMOS L2 v602 product. The CLSM SM simulations are always worse than the satellite retrievals in terms of R and Ranom, highlighting the relevance of satellite-based L-band SM retrievals

637 The experimental SM retrievals show a clear sensitivity to the RTM parameterization
638 (Figure 4, 5). The climatological SM level (and bias) is closely related to the soil roughness
639 h and the vegetation opacity τ. Higher roughness parameters h typically yield wetter SM
640 retrievals, as higher roughness decreases soil reflectivity, and therefore increases soil
641 emissivity (Kirchoff, 1860). To compensate for such a rise in emissivity during the
642 optimization process, high roughness often leads to wetter SM retrievals. The τ-values
643 similarly impact the SM retrieval climatology (Wigneron et al. 2017): larger b- and LEWT-
644 parameters lead to higher τ-values, which often lead to higher TB predictions and thus wetter
645 SM retrievals. The high roughness parameter h in Lit3 was used to produce sufficiently high
646 mean TB values for SMOS monitoring at ECMWF but creates SM retrievals which hover
647 around saturation to compensate for the high h values. Lit4 parameters were also tuned for
648 TB forward modeling (De Lannoy and Reichle, 2016a, b, Reichle et al. 2017), but in such a
649 way that both the mean and standard deviation in TB were optimized. As a result, these Lit4
650 parameters yield comparable SM results to the Lit1, Lit2 and Lit5 parameters that were tuned
651 for SM retrievals.

652 When including the TB configuration (Figure 4) in the sensitivity analysis, it is found
653 that lower incidence angles do not perform well when retrieving τ (option 2) as part of the
654 SM retrieval, because the MPDI relies on TB_V-TB_H differences, which are typically small
655 for low incidence angles. Consequently, calculating τ (option 1) as input results in better
656 performance when SM is retrieved from TB at lower incidence angles.

657 Finally, a sensitivity to the prior constraint in the CF (Figure 5) shows an improvement
658 in all skill metrics when either a time variable or time-mean CLSM SM estimate is included
659 as prior. This prior constraint avoids that SM retrievals hit boundary values and helps to
660 confine the SM range. Most importantly, the retrievals using CLSM output as prior are better
661 than either SM retrievals without using any prior constraint or CLSM SM output alone. This

662 highlights the power of Bayesian merging of two data sources into a more accurate SM
663 product.

664 The validation here is based on a handful of locations with very high-quality soil
665 moisture observations from locally dense networks and with a favorable landscape for
666 satellite-based SM retrieval. It is left for future work to explore the use of continental-scale
667 networks that provide only one point-scale sensor per satellite footprint. In any case, the
668 premise of this study, like that of most other validation studies, is that the performance and
669 sensitivity of the (ensemble) SM retrievals is similar in comparable locations where high-
670 quality in situ measurements are not available.

671 *6.2 Ensemble uncertainty*

672 Ensemble retrievals provide an option to quantify spatio-temporally varying SM
673 retrieval uncertainty, without using *in situ* reference data. However, it is not straightforward
674 to define a reliable ensemble that is statistically undistinguishable from the (unknown) truth.
675 A candidate ensemble with 85 members obtained by perturbing multiple literature-based
676 RTM parameter sets shows a large variability in long-term mean of the members (offset
677 variance) which greatly exceeds the anomaly variance and the temporal SM variability
678 (Figure 6). This large offset variance is consistent with the large time series skill differences
679 between various established and experimental retrieval products (Section 6.1) in terms of
680 bias. The ensemble offset variance underlines again the importance of potential large biases
681 in a retrieval product associated with a chosen RTM parameter set.

682 With inclusion of these individual member biases in the candidate ensemble, the
683 ensemble spread is too wide to be reliable (as defined in Section 4.2.1). When zero-centering
684 the ensemble members, the ensemble becomes reliable for further analysis (see Appendix B).
685 The space-time lumped EPDF is near Gaussian with an ensemble spread of 0.037 m^3/m^3,
686 mainly due to non-linear propagation of RTM parameter errors. This value is in line with the

temporal ubRMSD of 0.042±0.012 m^3/m^3, which was found as the mean (±spatial standard deviation) time series skill metric of ensemble mean retrievals versus *in situ* data across the 11 reference sites, and which is also approximately the target uncertainty for L-band SM retrievals. Indeed, a reliable ensemble was identified (not tuned) in such a way that the time-averaged anomaly spread would match the ubRMSD with *in situ* data (VI_A, Eq. B3), across all sites.

The ensemble approach for retrieving SM further facilitates the investigation of retrieval uncertainty as a function of varying SM. When the individual members are not zero-centered, a typical n-shape relationship between the mean (enmn(t)) and spread (√envar(t)) of the SM ensemble emerges (Figure 9a), with a lower spread as SM becomes drier or wetter in time, as expected for bounded variables. This shape is also found in earlier studies investigating the SM mean-variance relationship in space: spatial variance increases as the SM dries down and decreases after a threshold is met (Pan and Peters-Lidard, 2008). This mean-variance relationship in space is assumed to be dependent on soil hydraulic parameters (Vereecken et al. 2007; Zarlenga et al. 2018). The relationship between the SM mean and variance is thus similar regardless of whether it is derived by sampling using *in situ* data across time or space (per the literature), or using the ensemble (per this study), which suggests that ergodic sampling of SM statistics is appropriate.

In contrast, when zero-centering the ensemble members, high ensemble spreads are found at times when the ensemble-mean SM is either drier or wetter than its long-term mean (i.e. u-shape, Figure 9b). This is a reflection of remaining higher-order bias in the retrieval members. More specifically, each member is zero-centered and all members are highly correlated. When further deviating from the long-term mean SM (or from the 0-value after centering), the different ensemble members differ more, because the individual members may have different long-term variances (higher order bias) e.g. due to non-linear propagation of (time-invariant) RTM parameter errors. Consequently, the u-shape is not indicative of how

absolute retrieval error varies with SM. However, it does suggest that after a mean-only rescaling (Yilmaz et al. 2013) of various SM retrieval products (or other SM estimates), the *relative* error between SM products (or other estimates) will increase with the deviation from the long-term SM mean, if higher-order biases are present. Note that a related issue applies to the calculation of the time series ubRMSD, which only removes a stationary long-term bias from retrievals and reference data, while keeping higher-order biases. Consequently, the ubRMSD over a long time series is dominated by larger deviations during wet and dry periods (away from the long-term mean), if higher-order biases are present in either retrieval or reference data.

The SM retrieval error characterization via a time series comparison with *in situ* data and that obtained via ensembles reveal similar features. Persistent errors (or bias) are dominant in both the time series skills of various SM products (Figure 3) and in the large offset variance of SM retrievals (Figure 6). After removing a long-term mean bias, the remaining errors amount on average to about 0.04 m^3/m^3, both in terms of time series ubRMSD and anomaly ensemble spread (Figure 3 and 7). The latter two metrics suffer from the assumption that the removed bias is stationary or constant in time, whereas in reality this bias is not stationary and higher-order biases between various products and reference (*in situ*) data are present. In short, the retrieval errors increase when deviating further from the long-term mean SM.

The temporal and spatial behavior of the SM retrieval errors is related to the RTM structure and parameters. The counterpart of ensemble SM uncertainties estimated by land surface model (LSM) simulations (De Lannoy et al. 2006) may show different spatial and temporal patterns than those observed in retrievals. Uncertainty in SM from LSMs responds to uncertainties in the imposed meteorological forcings, initial conditions and soil hydraulic parameters, which are mostly distinct from the uncertainties in the RTM inputs that dominate the SM retrieval uncertainty. The different spatial and temporal nature of errors in SM

retrievals and errors in SM model simulations is often overlooked, but should be accounted for, when merging both data sources.

7. Conclusions

Three established SMOS SM retrieval products, SMOS L2 v620, SMOS IC v103 and SMOS CCI LPRM, as well as CLSM SM simulations and a range of experimental SM retrievals are compared to *in situ* observations at 11 reference sites across the USA. The various surface SM products perform similarly in terms of ubRMSD, R and Ranom, but differ considerably in terms of bias. Despite their inferior performance compared to the established SMOS SM retrieval products, CLSM SM simulations are always beneficial as prior constraint to improve experimental SM retrievals using Bayesian inversion. Besides the prior SM information, other aspects strongly impact the quality of experimental retrievals, such as the TB configuration and the RTM parameterization. The largest differences in retrieval skill are due to choices in RTM parameter values, which introduce both first- and higher-order biases. Wetter SM retrievals are typically found for higher vegetation τ and/or soil roughness h values. Retrievals for five different literature-based RTM parameter sets have been compared, covering parameters that were optimized for SM retrievals and parameters that were optimized for TB forward modeling. The latter parameters do not necessarily yield good results. However, the RTM parameters calibrated for TB forward modeling in the SMAP L4_SM product were optimized for both the mean and standard deviation in TB and are shown to perform well for SM retrievals too. Finally, retrievals that average multiple realizations typically perform slightly better than a single realization, but the desired reduction in time series ubRMSD is limited by the strong error correlations between different SM retrieval realizations: unlike model forecasts which propagate and randomize errors in time via the prognostic model, retrievals are obtained from a diagnostic system that does not disperse errors in time.

763 The uncertainty of SM retrievals is typically expressed with time-integrated skill metrics
764 relative to reference data. To obtain spatiotemporally varying estimates of error estimates
765 without involving reference data, time series of ensemble SM retrievals are studied. The SM
766 members show a large variability in the long-term mean values associated with the choice of
767 the RTM parameters. After zero-centering the members, the time-space average anomaly
768 standard deviation is 0.037 m^3/m^3 (without tuning), closely matching both the site-averaged
769 actual time series error between ensemble mean retrievals and *in situ* data of 0.042 m^3/m^3 as
770 well as the target uncertainty of ubRMSD=0.04 m^3/m^3 for L-band retrievals. The near-
771 Gaussian shape of this anomaly ensemble SM retrieval distribution supports the common
772 approach to represent the SM uncertainty by the ensemble standard deviation.

773 A closer analysis shows that the total ensemble spread in SM retrievals is small for dry
774 and wet SM, as expected for a bounded variable. This is consistent with earlier reported
775 mean-variance relationship based on spatial and temporal *in situ* sampling and supports
776 ergodic sampling of SM uncertainty. In contrast, the zero-centered ensembles show high
777 spreads for SM values that are wetter or drier than the long-term mean SM. This u-shape in
778 the ensemble mean-spread relationship is the result of only removing the long-term mean
779 bias, while leaving the higher-order bias (e.g. differences in variances) in the individual
780 ensemble members. This indicates that, whereas the absolute SM retrieval errors are smaller
781 for wet and dry SM, the *relative* error between various SM products after mean-only rescaling
782 is expected to increase when SM values deviates more from the long-term mean SM. The
783 removal of bias (first- and higher-order) in retrievals and an accurate identification of the
784 remaining spatio-temporally varying error, or in other words a good error variance
785 decomposition, is crucial to efficiently merge various data sources.

786

Author Contributions: Conceptualization, J.Q. and G.D.L..; validation, M.C., R.v.d.S., J.P.W.; formal analysis, all authors; writing, all authors.

Funding: The authors acknowledge the support from KU Leuven C1 internal fund C14/16/045 and the support from the Research Foundation Flanders FWO-1512817N. Rolf Reichle was supported by the NASA SMAP mission.

Acknowledgments: The authors thank the reviewers and editorial board.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A: Ensemble mean ubRMSD

This section is based on the work of Bishop (2006). Consider a model that attempts to estimate a variable, e.g. SM. Model ensembles are then built on the theory that the ensemble mean of the model ensemble is a better estimator for this variable than any of the individual ensemble member estimators. Indeed, consider:

$$v_{\mathrm{m}}(t) = h(t) + \epsilon_{\mathrm{m}}(t) \tag{A1}$$

where $v_m(t)$ signifies ensemble member m (or estimator m), $h(t)$ the true data generating process (e.g. true SM), and $\epsilon_m(t)$ the estimation error of ensemble member m, all in function of time t. The ensemble mean estimator $v_{\mathrm{EM}}(t)$ is then described by:

$$v_{\mathrm{EM}}(t) = \frac{1}{\mathrm{K}}\sum_{\mathrm{m}=1}^{\mathrm{K}} v_{\mathrm{m}}(t) \tag{A2}$$

with K the number of ensemble members in the ensemble model. Then the mean squared difference between ensemble member m and truth is given by:

$$\mathrm{MSD}_{\mathrm{m}} = \frac{1}{\mathrm{N}}\left(\sum_{t=1}^{\mathrm{N}}(v_{\mathrm{m}}(t) - h(t))^2\right) = \frac{1}{\mathrm{N}}\left(\sum_{t=1}^{\mathrm{N}} \epsilon_{\mathrm{m}}(t)^2\right) \tag{A3}$$

The mean squared difference between the ensemble mean and truth is then:

$$\mathrm{MSD}_{\mathrm{EM}} = \frac{1}{\mathrm{N}}\left(\sum_{t=1}^{\mathrm{N}}\left(\frac{1}{\mathrm{K}}\sum_{\mathrm{m}=1}^{\mathrm{K}} v_{\mathrm{m}}(t) - h(t)\right)^2\right) = \frac{1}{\mathrm{N}}\left(\sum_{t=1}^{\mathrm{N}}\left(\frac{1}{\mathrm{K}}\sum_{\mathrm{m}=1}^{\mathrm{K}} \epsilon_{\mathrm{m}}(t)\right)^2\right) =$$

$$\frac{1}{\mathrm{N}}\left(\sum_{t=1}^{\mathrm{N}}\frac{1}{\mathrm{K}^2}\sum_{\mathrm{m}=1}^{\mathrm{K}} \epsilon_{\mathrm{m}}(t)\sum_{\mathrm{m}=1}^{\mathrm{K}} \epsilon_{\mathrm{m}}(t)\right) \tag{A4}$$

When the assumption is made that every ensemble member is an unbiased estimator, i.e. the time-mean estimation error $\mu_{\epsilon_{\mathrm{m}}} = 0, \forall \mathrm{m} \in \{1, \ldots, \mathrm{m}\}$, and that ensemble errors of distinct ensemble members are statistically independent, and thus uncorrelated, Eq.A4 can be written as:

$$\mathrm{ubMSD}_{\mathrm{EM}} = \frac{1}{\mathrm{K}^2\mathrm{N}}\sum_{t=1}^{\mathrm{N}}\left(\left(\sum_{\mathrm{m}=1}^{\mathrm{K}} \epsilon_{\mathrm{m}}(t) - \mu_{\sum_{\mathrm{m}=1}^{\mathrm{M}} \epsilon_{\mathrm{m}}(t)}\right)\left(\sum_{\mathrm{m}=1}^{\mathrm{K}} \epsilon_{\mathrm{m}}(t) - \mu_{\sum_{\mathrm{m}=1}^{\mathrm{M}} \epsilon_{\mathrm{m}}(t)}\right)\right) \tag{A5}$$

with $\mathrm{ubRMSD_{EM}}$ the unbiased $\mathrm{MSD_{EM}}$. Indeed, $\mu_{\sum_{m=1}^{K}\epsilon_m(t)} = \mu_{\epsilon_1}(t) + \mu_{\epsilon_2}(t) + \cdots + \mu_{\epsilon_M}(t) = 0$ under the assumption of unbiased ensemble members, thereby leading to the equality of the right hand sides of Eq.A4 and Eq.A5. Eq.A5 can further be written as the product of $\frac{1}{K^2}$ and the variance of $\sum_{m=1}^{K}\epsilon_m(t)$:

$$\mathrm{ubMSD_{EM}} = \frac{1}{K^2}\mathrm{var}\left(\sum_{m=1}^{K}\epsilon_m(t)\right) \quad \text{(A6)}$$

The assumption of uncorrelated errors between any pair of ensemble members implies that:

$$\mathrm{ubMSD_{EM}} = \frac{1}{K^2}\sum_{m=1}^{K}\mathrm{var}(\epsilon_m(t)) = \frac{1}{K^2}\sum_{m=1}^{K}\frac{1}{N}\sum_{t=1}^{N}\left(\epsilon_m(t) - \mu_{\epsilon_m}\right)^2 = \frac{1}{K^2}\sum_{m=1}^{K}\frac{1}{N}\sum_{t=1}^{N}\epsilon_m(t)^2 \quad \text{(A7)}$$

According to EqA1, EqA6 can be rewritten as:

$$\mathrm{ubMSD_{EM}} = \frac{1}{K^2}\sum_{m=1}^{K}\mathrm{ubMSD_m} = \frac{K}{K^2}\mathrm{E_{ens}}[\mathrm{ubMSD_m}] = \frac{1}{K}\mathrm{E_{ens}}[\mathrm{ubMSD_m}] \quad \text{(A8)}$$

$$\mathrm{ubRMSD_{EM}} = \frac{1}{\sqrt{K}}\mathrm{E_{ens}}[\mathrm{ubRMSD_m}] \quad \text{(A9)}$$

with $\mathrm{E_{ens}}[.]$ the ensemble mean operator. Eq.A9 shows that the expected ubRMSD of the ensemble mean improves $\sqrt{K}$ times (lower values are better) as compared to the ensemble averaged ensemble member ubRMSDs. Note that Eq. A8-A9 only hold true under the assumption of totally uncorrelated ensemble member errors.

Appendix B: Ensemble verification

An ensemble of SM retrievals is assumed to reliably represent retrieval uncertainty when the corresponding reference *(in situ)* SM time series is not statistically distinct from the set of ensemble member time series (Anderson and Anderson 1999). To verify this condition, three ensemble verification tools are used in this study: two spread-skill relationships (Brankovic et al. 1990, Talagrand, et al. 1997), and the rank histogram (Hamill 2001).

First, consider the ensemble mean squared difference between the ensemble member values $v_{\mathrm{m}}(t)$ and the *in situ* observation $w(t)$, at one site, either calculated (Eq. B1, left) at each time *t*, i.e. $\mathrm{enMSD}(t)$, or (Eq. B1, right) averaged over the time dimension, i.e. $\overline{\mathrm{enMSD}}$:

$$\mathrm{enMSD}(t) = \frac{1}{\mathrm{K}}\sum_{m=1}^{\mathrm{K}}\left(v_{\mathrm{m}}(t) - w(t)\right)^2; \qquad \overline{\mathrm{enMSD}} = \frac{1}{\mathrm{N}}\sum_{t=1}^{\mathrm{N}}\mathrm{enMSD}(t). \tag{B1}$$

Also consider the temporal mean squared error between the ensemble mean $\mathrm{enmn}(t)$ and the reference *in situ* observations $w(t)$:

$$\overline{\mathrm{SE}_{\mathrm{enmn}}} = \frac{1}{\mathrm{N}}\sum_{t=1}^{\mathrm{N}}\left(\mathrm{enmn}(t) - w(t)\right)^2 \ . \tag{B2}$$

The relationship between the ensemble variance and the above time-averaged differences between *in situ* and ensemble mean or ensemble member retrieval SM, yields two ensemble Verification Indices (VI) to evaluate the 'reliability' of an ensemble:

$$\mathrm{VI}_{\mathrm{A}} = \frac{\overline{\mathrm{SE}_{\mathrm{enmn}}}}{\overline{\mathrm{envar}}} \ , \tag{B3}$$

$$\mathrm{VI}_{\mathrm{B}} = \sqrt{\frac{\overline{\mathrm{SE}_{\mathrm{enmn}}}}{\overline{\mathrm{enMSD}}}}. \tag{B4}$$

The VI_{A} compares the actual error of the ensemble mean (relative to reference data) to the estimated ensemble variance, whereas VI_{B} is the ratio of the time mean error between the ensemble mean and reference data to the ensemble mean of the time series differences between single members and reference data (note how ensemble and time averaging are interchanged). Both VI_{A} and VI_{B} range between 0 and +infinity. If the members of a given

SM retrieval ensemble are statistically indistinguishable from the *in situ* SM time series, that is, if the SM ensemble reliably represents the SM retrieval uncertainty, VI_A approaches 1 and VI_B approaches $\sqrt{\frac{(K+1)}{2K}}$ (Brankovic et al. 1990, Moradkhani et al. 2005).

The rank histogram is a third verification tool. Rank histograms reveal where *in situ* SM values rank relative to a SM distribution defined by a K-member ensemble of SM retrievals. For each time *t*, the K ensemble SM values are ranked and the ensemble SM range is divided into k histogram bins (k=10 in this study). The first and last bin respectively cover any SM values lower than the lowest ranked ensemble SM, and higher than the highest ranked ensemble SM. The remaining k-2 bin borders are defined so that each bin contains an equal number of ensemble members. The number of *in situ* SM values in each ensemble histogram bin, summed up over the entire time series period, yields a rank histogram. Flat rank histograms indicate reliable ensembles. U- or n-shaped rank histograms respectively indicate that a large proportion of the *in situ* observations respectively fall outside or only within a small subset of the ensemble range, indicating a too narrow or too wide ensemble. L- or J-shaped rank histograms signify a wet or dry bias in the ensembles (Hamill 2001).

A range of different SM retrieval ensembles are verified with VI_A, VI_B and a 10-bin rank histogram, using reference *in situ* SM observations at 11 locations (Section 2.1). For each of the candidate SM ensembles, the three above-mentioned verification tools are first calculated separately for each site, and then spatially averaged. The SM ensemble deemed to most reliably represent SM retrieval uncertainty, according to the above verification tools is then retained to analyze SM retrieval uncertainty in time and space.

Figure B1 shows the rank or Talagrand diagrams for two candidate ensembles without and with zero-centering (Section 4.2.1). These ensembles are created for a range of RTM parameters (Section 4.2.1) and differ only in whether they use a prior constraint or not : one ensemble uses no prior SM (case α=0) in the CF, the second ensemble uses prior CLSM SM

(case α=1) in the CF. The candidate ensemble without the prior SM constraint yields a n-shaped histogram without zero-centering (Figure B1a) and a near uniform histogram after zero-centering (Figure B1c). The candidate ensemble with the strong prior SM constraint yields an n-shaped histogram without zero-centering (Figure B1b), similar to the ensemble without the prior constraint (Figure B1a) . Zero-centering of the ensemble with the prior SM constraint, however, yields a u-shaped rank histogram (Figure B1d), suggesting a lack of ensemble spread. Even though retrievals with a strong prior constraint have a better skill relative to reference data (Figure 5), the spread in our ensemble with a strong prior constraint is not realistic and therefore not further considered. The zero-centered ensemble without prior can be deemed the most reliable, based on the nearly flat Talagrand diagram and verification indices VI_A (0.88) and VI_B (1.10) relatively close to their respective ideal values of 1 and $\frac{\sqrt{85+1}}{2*85} \cong 0.71$.

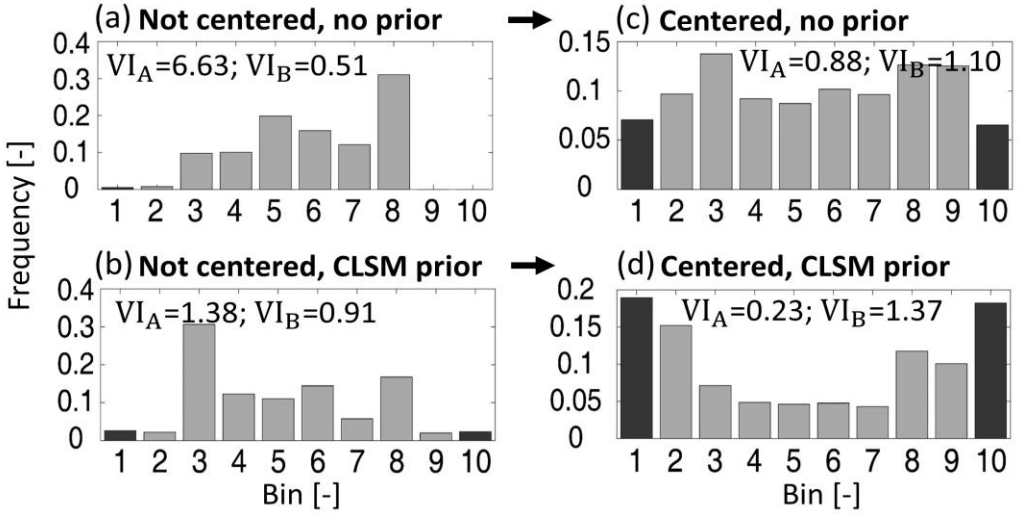


Figure B1. Rank histograms for two 85-member candidate SM ensembles (5 RTM parameters and 17 perturbations). The SM retrievals in (a,c) do not use prior SM constraints in the CF, whereas (b,d) include a strong prior constraint. (c,d) are derived from (a,b) after zero-centering each SM retrieval member (and also the *in situ* time series) by subtracting the time-mean values, for the period 1 June 2010 – 1 June 2015, and all 11 reference sites. Each ensemble set is verified by verification indices VI_A and VI_B (Eq.B3 and Eq.B4, respectively).

References

1. Albergel, C., P. de Rosnay, C. Gruhier, J. Munoz-Sabater, S. Hasenauer, L. Isaksen, Y. Kerr & W. Wagner (2012). Evaluation of remotely sensed and modelled soil moisture products using global ground-based in situ observations. Remote Sensing of Environment, 118, 215-226.
2. Alvarez-Garreton, C., Ryu, R., Western, A. W., Su, C.-H., Crow, W. T., Robertson, D.E., Leahy, C. (2015). Improving operational flood ensemble prediction by the assimilation of satellite soil moisture: comparison between lumped and semi-distributed schemes, Hydrol. Earth Syst. Sci., 19, 1659-1676, https://doi.org/10.5194/hess-19-1659-2015.
3. Al Yaari, A., Wigneron, J.-P., Ducharne, A., Kerr, Y., Wagner, W., De Lannoy, G.J.M., Reichle, R.H., Al Bitar, A., Dorigo, W., Richaume, P., Mialon, A. (2014). Global-scale comparison of passive (SMOS) and active (ASCAT) satellite-based microwave soil moisture retrievals with soil moisture simulations (MERRA-Land). Remote Sensing of Environment, 152, 614-626, doi:10.1016/j.rse.2014.07.01.
4. Al Yaari, A., Wigneron, J-P, Dorigo, W., Colliander, A., Pellarin, T., Hahn, S., Mialon, A., Richaume, P., Fernandez, R., Fan, L., Kerr, Y., De Lannoy, G. (2019). Assessment and inter-comparison of recently developed/reprocessed microwave satellite soil moisture products. Remote Sensing of Environment, in press.
5. Anderson, J. L. & S. L. Anderson (1999). A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. Monthly Weather Review, 127, 2741-2758.
6. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Cambridge: Springer.
7. Bindlish, R., M. Cosh, and co-authors (2018). GCOM-W AMSR2 Soil Moisture Product Validation Using Core Validation Sites, IEEE JSTAR, 11(2), 209-219.
8. Brankovic, C., T. N. Palmer, F. Molteni, S. Tibaldi & U. Cubasch (1990). Extended-range predictions with ECMWF models – time-lagged ensemble forecasting. Quarterly Journal of the Royal Meteorological Society, 116, 867-912.
9. Brodzik, M. J., B. Billingsley, T. Haran, B. Raup & M. H. Savoie (2012). EASE-Grid 2.0: Incremental but Significant Improvements for Earth-Gridded Data Sets. Isprs International Journal of Geo-Information, 1, 32-45.
10. --- (2014). EASE-Grid 2.0: Incremental but Significant Improvements for Earth-Gridded Data Sets (vol 1, pg 32, 2012). ISPRS International Journal of Geo-Information, 3, 1154-1156.
11. Burgin, M., A. Colliander, and co-authors (2017). A Comparative Study of the SMAP Passive Soil Moisture Product With Existing Satellite-Based Soil Moisture Products (2017). IEEE Transactions on Geoscience and Remote Sensing, 55(5), 2959--2971.
12. Calvet, J. C., J. P. Wigneron, J. Walker, F. Karbou, A. Chanzy & C. Albergel (2011). Sensitivity of Passive Microwave Observations to Soil Moisture and Vegetation Water Content: L-Band to W-Band. IEEE Transactions on Geoscience and Remote Sensing, 49, 1190-1199.
13. Chan, S., R. Bindlish, and co-authors (2016). Assessment of the SMAP passive soil moisture product. IEEE Transactions on Geoscience and Remote Sensing, 54(8), 4994-5007.
14. Chen, F., Crow, W.T., Bindlish, R., Colliander, A., Burgin, M.S., Asanuma, J., Aida, K. (2018). Global-scale evaluation of SMAP, SMOS and ASCAT soil moisture products using triple collocation. Remote Sensing of Environment, 214, pp. 1-13. DOI: 10.1016/j.rse.2018.05.008.
15. Chen, F., Crow, W. T., Colliander, A., Cosh, M. H., Jackson, T. J., Bindlish, R., ... & Goodrich, D. C. (2017). Application of triple collocation in ground-based validation of

950 Soil Moisture Active/Passive (SMAP) level 2 data products. IEEE Journal of Selected
951 Topics in Applied Earth Observations and Remote Sensing, 10(2), 489-502.
952 16. Cheung, K. K. W. (2001). A review of ensemble forecasting techniques with a focus on
953 tropical cyclone forecasting. Meteorological Applications, 8, 315-332.
954 17. Colliander, A., T. J. Jackson, R. Bindlish, S. Chan, N. Das, S. B. Kim, M. H. Cosh, R.
955 S. Dunbar, L. Dang, L. Pashaian, J. Asanuma, K. Aida, A. Berg, T. Rowlandson, D.
956 Bosch, T. Caldwell, K. Caylor, D. Goodrich, H. al Jassar, E. Lopez-Baeza, J. Martinez-
957 Fernandez, A. Gonzalez-Zamora, S. Livingston, H. McNairn, A. Pacheco, M.
958 Moghaddam, C. Montzka, C. Notarnicola, G. Niedrist, T. Pellarin, J. Prueger, J.
959 Pulliainen, K. Rautiainen, J. Ramos, M. Seyfried, P. Starks, Z. Su, Y. Zeng, R. van der
960 Velde, M. Thibeault, W. Dorigo, M. Vreugdenhil, J. P. Walker, X. Wu, A. Monerris, P.
961 E. O'Neill, D. Entekhabi, E. G. Njoku & S. Yueh (2017). Validation of SMAP surface
962 soil moisture products with core validation sites. Remote Sensing of Environment, 191,
963 215-231.
964 18. Crow, W. T., S. T. K. Chan, D. Entekhabi, P. R. Houser, A. Y. Hsu, T. J. Jackson, E. G.
965 Njoku, P. E. O'Neill, J. C. Shi & X. W. Zhan (2005). An observing system simulation
966 experiment for Hydros radiometer-only soil moisture products. Ieee Transactions on
967 Geoscience and Remote Sensing, 43, 1289-1303.
968 19. Crow, W., Su, C.-H., Ryu, D., Yilmaz, T. (2015). Optimal averaging of soil moisture
969 predictions from ensemble land surface model simulations. Water Resources Research,
970 51, 9273-9289.
971 20. De Lannoy, G.J.M., Reichle, R.H., Peng, J., Kerr, Y., Castro, R., Kim, E.J., Liu, Q.
972 (2015). Converting Between SMOS and SMAP Level-1 Brightness Temperature
973 Observations Over Nonfrozen Land. IEEE Geoscience and Remote Sensing Letters,
974 12(9), 1-5, doi:10.1109/LGRS.2015.2437612.
975 21. De Lannoy, G. J. M. & R. H. Reichle (2016a). Assimilation of SMOS brightness
976 temperatures or soil moisture retrievals into a land surface model. Hydrology and Earth
977 System Sciences, 20, 4895-4911.
978 22. De Lannoy, G.J.M., Reichle, R.H. (2016b). Global Assimilation of Multi-Angle and
979 Multi-Polarization SMOS Brightness Temperature Observations into the GEOS-5
980 Catchment Land Surface Model for Soil Moisture Estimation. Journal of
981 Hydrometeorology, 17(2), 669-691, doi: 10.1175/JHM-D-15-0037.1.
982 23. De Lannoy, G. J. M., R. H. Reichle & V. R. N. Pauwels (2013). Global Calibration of
983 the GEOS-5 L-Band Microwave Radiative Transfer Model over Nonfrozen Land Using
984 SMOS Observations. Journal of Hydrometeorology, 14, 765-785.
985 24. De Lannoy, G.J.M., Koster, R.D., Reichle, R.H., Mahanama, S.P., Liu, Q. (2014). An
986 Updated Treatment of Soil Texture and Associated Hydraulic Properties in a Global
987 Land Modeling System. Journal of Advances in Modeling Earth Systems, 6, 23p.,
988 doi:10.1002/2014MS000330.
989 25. De Lannoy, G.J.M., Houser, P.R., Pauwels, V.R.N., Verhoest, N.E.C. (2006).
990 Assessment of model uncertainty for soil moisture through ensemble verification.
991 Journal of Geophysical Research, 111, D10101, doi:10.1029/2005JD006367
992 26. de Rosnay, P., M. Drusch, A. Boone, G. Balsamo, B. Decharme, P. Harris, Y. Kerr, T.
993 Pellarin, J. Polcher & J. P. Wigneron (2009). AMMA Land Surface Model
994 Intercomparison Experiment coupled to the Community Microwave Emission Model:
995 ALMIP-MEM. Journal of Geophysical Research-Atmospheres, 114, 18.
996 27. Draper, C., R. Reichle, R. de Jeu, V. Naeimi, R. Parinussa & W. Wagner (2013).
997 Estimating root mean square errors in remotely sensed soil moisture over continental
998 scale domains. Remote Sensing of Environment, 137, 288-298.
999 28. Drusch, M., T. Holmes, P. de Rosnay & G. Balsamo (2009). Comparing ERA-40-Based
1000 L-Band Brightness Temperatures with Skylab Observations: A Calibration/Validation

Study Using the Community Microwave Emission Model. Journal of Hydrometeorology, 10, 213-226.
29. Entekhabi, D., 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 & J. Van Zyl (2010a). The Soil Moisture Active Passive (SMAP) Mission. Proceedings of the Ieee, 98, 704-716.
30. Entekhabi, D., R. H. Reichle, R. D. Koster & W. T. Crow (2010b). Performance Metrics for Soil Moisture Retrievals and Application Requirements. Journal of Hydrometeorology, 11, 832-840.
31. Entekhabi, D., and co-authors (2014). The SMAP Handbook. JPL.
32. Fernandez-Moran, R., A. Al-Yaari, A. Mialon, A. Mahmoodi, A. Al Bitar, G. De Lannoy, N. Rodriguez-Fernandez, E. Lopez-Baeza, Y. Kerr & J. P. Wigneron (2017). SMOS-IC: An Alternative SMOS Soil Moisture and Vegetation Optical Depth Product. Remote Sensing, 9, 21.
33. Fortin, V., M. Abaza, F. Anctil & R. Turcolle (2014). Why Should Ensemble Spread Match the RMSE of the Ensemble Mean? Journal of Hydrometeorology, 15, 1708-1713.
34. Grant, J. P., K. Saleh-Contell, J. P. Wigneron, M. Guglielmetti, Y. H. Kerr, M. Schwank, N. Skou & A. A. V. de Griend (2008) Calibration of the L-MEB model over a coniferous and a deciduous forest. Ieee Transactions on Geoscience and Remote Sensing, 46, 808-818.
35. Gruber, A., Su, C. H., Zwieback, S., Crow, W., Dorigo, W., & Wagner, W. (2016). Recent advances in (soil moisture) triple collocation analysis. International Journal of Applied Earth Observation and Geoinformation, 45, 200-211.
36. Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. Monthly Weather Review, 129, 550-560.
37. Imaoka, K., T. Maeda, M. Kachi, M. Kasahara, N. Ito & K. Nakagawa (2012). Status of AMSR2 instrument on GCOM-W1. In Conference on Earth Observing Missions and Sensors - Development, Implementation, and Characterization II. Kyoto, JAPAN: Spie-Int Soc Optical Engineering.
38. Jackson, T. J., M. H. Cosh, R. Bindlish, P. J. Starks, D. D. Bosch, M. Seyfried, D. C. Goodrich, M. S. Moran & J. Y. Du (2010). Validation of Advanced Microwave Scanning Radiometer Soil Moisture Products. Ieee Transactions on Geoscience and Remote Sensing, 48, 4256-4272.
39. Kerr, Y. H., A. Al-Yaari, N. Rodriguez-Fernandez, M. Parrens, B. Molero, D. Leroux, S. Bircher, A. Mahmoodi, A. Mialon, P. Richaume, S. Delwart, A. Al Bitar, T. Pellarin, R. Bindlish, T. J. Jackson, C. Rudiger, P. Waldteufel, S. Mecklenburg & J. P. Wigneron (2016). Overview of SMOS performance in terms of global soil moisture monitoring after six years in operation. Remote Sensing of Environment, 180, 40-63.
40. Kerr, Y. H., P. Waldteufel, P. Richaume, J. P. Wigneron, P. Ferrazzoli, A. Mahmoodi, A. Al Bitar, F. Cabot, C. Gruhier, S. E. Juglea, D. Leroux, A. Mialon & S. Delwart (2012) The SMOS Soil Moisture Retrieval Algorithm. IEEE Transactions on Geoscience and Remote Sensing, 50, 1384-1403.
41. ——, P. Waldteufel, P. Richaume, P. Ferrazzoli, and J.- P. Wigneron (2011). SMOS level 2 soil moisture processor development continuation project. Algorithm Theoretical Basis Doc., v3.4, Tech.Note SO-TN-ARR-L2PP-0037, 121 pp
42. Kirchoff, G. (1860) Ueber das Verhältniss zwischen dem Emissionsvermögen und dem Absorptionsvermögen der Körper für Wärme und Licht. annalen der physik, 185, 26.
43. Konings, A., Entekhabi, D., Chan, S., Njoku, E. (2011). Effect of radiative transfer uncertainty on L-band radiometric soil moisture retrieval. IEEE Transactions on Geoscience and Remote Sensing, 49(7), 2686-2678.

44. Koster, R. D., M. J. Suarez, A. Ducharne, M. Stieglitz, and P. Kumar, 2000: A catchment-based approach to modeling land surface processes in a general circulation model 1. Model structure. J. Geophys. Res., 105 (D20), 24 809–24 822.
45. Mahanama, S. P., R. D. Koster, G. K. Walker, L. L. R. Takacs, R.H., G. De Lannoy, Q. Liu, B. Zhao & M. J. Suarez (2015). Land Boundary Conditions for the Goddard Earth Observing System Model Version 5 (GEOS-5) Climate Modeling System - Recent Updates and Data File Descriptions. In Technical Report Series on Global Modeling and Data Assimilation, ed. R. D. Koster, 43. Greenbelt: National Aeronautics and Space Administration.
46. Meesters, A., R. A. M. De Jeu & M. Owe (2005) Analytical derivation of the vegetation optical depth from the microwave polarization difference index. Ieee Geoscience and Remote Sensing Letters, 2, 121-123.
47. Mironov, V., Kerr, Y., Wigneron, J.P., Kosolapova, L., Demontoux, F., 2013. Temperature and texture dependent dielectric model for moist soils at 1.4 GHz. IEEE Geosci. Remote Sens. Lett. 10, 419–423.
48. Mo, T., Choudhury, B. J., Schmugge, T. J., Wang, J. R., & Jackson, T. J. (1982). A model for microwave emission from vegetation-covered fields. Journal of Geophysical Research: Oceans, 87(C13), 11229-11237.
49. Moradkhani, H., Sorooshian, S., Gupta, H., Houser, P. (2005). Dual state-parameter estimation of hydrological models using ensemble Kalman filter. Advances in Water Resources, 28(2), 135-147.
50. O'Neill, P. (2012). Soil Moisture Active Passive (SMAP) Algorithm Theoretical Basis Document Level 2 & 3 Soil Moisture (Passive) Data Products. 88. Jet Propulsion Laboratory.
51. Pan, F. and Peters-Lidard, C. (2008). On the Relationship Between Mean and Variance of Soil Moisture Fields, JAWRA, 44(1), 235-242.
52. Parinussa, R. M., A. Meesters, Y. Y. Liu, W. Dorigo, W. Wagner & R. A. M. de Jeu (2011) Error Estimates for Near-Real-Time Satellite Soil Moisture as Derived From the Land Parameter Retrieval Model. Ieee Geoscience and Remote Sensing Letters, 8, 779-783.
53. Parinussa, R.M.; De Jeu, R.A.M.; Van der Schalie, R.; Crow, W.T.; Lei, F.; Holmes, T.R.H. A Quasi-Global Approach to Improve Day-Time Satellite Surface Soil Moisture Anomalies through the Land Surface Temperature Input. Climate 2016, 4, 50.
54. Reichle, R. H., R. D. Koster, J. R. Dong & A. A. Berg (2004) Global soil moisture from satellite observations, land surface models, and ground data: Implications for data assimilation. Journal of Hydrometeorology, 5, 430-442.
55. Reichle, R. H., and Coauthors (2017). Assessment of the SMAP Level-4 Surface and Root-Zone Soil Moisture Product Using In Situ Measurements. Journal of Hydrometeorology, 18, 2621-2645, doi:10.1175/JHM-D-17-0063.1.
56. Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., Bosilovich, M. G., Schubert, S. D., Takacs, L., Kim, G.-K., Bloom, S., Chen, J., Collins, D., Conaty, A., da Silva, A., Gu, W., Joiner, J., Koster, R. D., Lucchesi, R., Molod, A., Owens, T., Pawson, S., Pegion, P., Redder, C. R., Reichle, R., Robertson, F. R., Ruddick, A. G., Sienkiewicz, M., and Woollen, J.: MERRA – NASA's Modern-Era Retrospective Analysis for Research and Applications (2011). J. Climate, 24, 3624–3648, doi:10.1175/JCLI-D-11-00015.1.
57. Rodriguez-Fernandez, N., Sabater, J., Richaume, P., de Rosnay, P., Kerr, Y., Drusch, M., Mecklenburg, S. (2017). SMOS near-real-time soil moisture product: processor overview and first validation results. HESS, 21, 5201-5216.
58. Sabater, J. M., P. De Rosnay & G. Balsamo (2011). Sensitivity of L-band NWP forward modelling to soil roughness. International Journal of Remote Sensing, 32, 5607-5620.

59. Saleh, K., J. P. Wigneron, P. Waldteufel, P. de Rosnay, M. Schwank, J. C. Calvet & Y. H. Kerr (2007) Estimates of surface soil moisture under grass covers using L-band radiometry. Remote Sensing of Environment, 109, 42-53.
60. Scipal, K., T. Holmes, R. de Jeu, V. Naeimi & W. Wagner (2008) A possible solution for the problem of estimating the error structure of global soil moisture data sets. Geophysical Research Letters, 35.
61. Talagrand, O., R. Vautard & B. Strauss (1997). Evaluation of probabilistic prediction systems. Reading, UK: European Center for Medium-Range Weather Forecast, Workshop on Predictability, 20-22 October.
62. Toth, Z, Kalnay, E. (1993). Ensemble Forecasting at NMC: the Generation of Perturbations, BAMS, 74(12), 2317-2330.
63. Van der Schalie, R., R. A. M. de Jeu, Y. H. Kerr, J. P. Wigneron, N. J. Rodriguez-Fernandez, A. Al-Yaari, R. M. Parinussa, S. Mecklenburg & M. Drusch (2017). The merging of radiative transfer based surface soil moisture data from SMOS and AMSR-E. Remote Sensing of Environment, 189, 180-193.
64. Van der Schalie, R., Kerr, Y., Wigneron, J.-P., Rodriguez-Fernandez, N., Al-Yaari, A., de Jeu, R. (2016). Global SMOS soil moisture retrievals from the Land Parameter Retrieval Model. International Journal of Applied Earth Observation and Geoinformation, 2016, 45 (B), p. 125-134.
65. van Leeuwen, P. J. (2015) Representation errors and retrievals in linear and nonlinear data assimilation. Quarterly Journal of the Royal Meteorological Society, 141, 1612-1623.
66. Vereecken, H., Kamai, T., Harter, T., Kasteel, R., Hopmans, J., Vanderborght, J. (2007). Explaining soil moisture variability as a function of mean soil moisture: A stochastic unsaturated flow perspective. Geophysical Research Letters, 34(22), 10.1029/2007GL031813
67. Wang, J. R. & T. J. Schmugge (1980) An empirical model for the complex dielectric permittivity of soils as function of water content. IEEE Transactions on Geoscience and Remote Sensing, 18, 288-295.
68. Wigneron, J. P., T. J. Jackson, P. O'Neill, G. De Lannoy, P. de Rosnay, J. P. Walker, P. Ferrazzoli, V. Mironov, S. Bircher, J. P. Grant, M. Kurum, M. Schwank, J. Munoz-Sabater, N. Das, A. Royer, A. Al-Yaari, A. Al Bitar, R. Fernandez-Moran, H. Lawrence, A. Mialon, M. Parrens, P. Richaume, S. Delwart & Y. Kerr (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, 238-262.
69. Wigneron, J. P., Y. Kerr, P. Waldteufel, K. Saleh, M. J. Escorihuela, P. Richaume, P. Ferrazzoli, P. de Rosnay, R. Gurney, J. C. Calvet, J. P. Grant, M. Guglielmetti, B. Hornbuckle, C. Matzler, T. Pellarin & M. Schwank (2007) L-band Microwave Emission of the Biosphere (L-MEB) Model: Description and calibration against experimental data sets over crop fields. Remote Sensing of Environment, 107, 639-655.
70. Yilmaz, M. T., & Crow, W. T. (2013). The optimality of potential rescaling approaches in land data assimilation. Journal of Hydrometeorology, 14(2), 650-660.
71. Yilmaz, M. T., E. R. Hunt, and T. J. Jackson (2008). Remote sensing of vegetation water content from equivalent water thickness using satellite imagery. Remote Sens. Environ., 112, 2514–2522.
72. Zarlenga, A., Fiori, A., and Russo, D. (2018). Spatial Variability of Soil Moisture and the Scale Issue: A Geostatistical Approach. Water Res Research, 54, 1765-1780.