Validation practices for satellite soil moisture products: What are (the) errors?

A. Gruber, G. De Lannoy, C. Albergel, A. Al-Yaari, L. Brocca, J.-C. Calvet, A. Colliander, M. Cosh, W. Crow, W. Dorigo, C. Draper, M. Hirschi, Y. Kerr, A. Konings, W. Lahoz, K. McColl, J. Muñoz-Sabater, J. Peng, R. Reichle, P. Richaume, C. Ruediger, T. Scanlon, R. van der Schalie, W. Wagner

Abstract

This paper presents a community effort to develop good practice guidelines for the validation of global coarse-scale satellite soil moisture products. We provide theoretical background, a review of state-of-the-art methodologies for estimating errors in soil moisture data sets, practical recommendations on data pre-processing and presentation of statistical results, and a recommended validation protocol that is supplemented with an example validation exercise focused on microwave-based surface soil moisture products. We conclude by identifying research gaps that should be addressed in the near future.

1 Introduction

The validation of soil moisture data sets aims to provide quantitative information about their quality by estimating systematic and random errors (*JCGM*, 2008). For satellite-derived products, this task is far from trivial because high-quality reference data are rarely available at the coarse spatial resolution of space borne microwave instruments that are predominantly used for soil moisture retrievals ($\sim 10^1 - 10^3$ km^2), and the retrieval quality is affected by numerous spatially and temporally variable factors (i.e. climatic, topographic and land cover conditions as well as instrument characteristics and the retrieval algorithm structure) (*Ochsner et al.*, 2013; *Crow et al.*, 2012; *Molero et al.*, 2018).

A host of methods exist to reconcile the distinct spatio-temporal characteristics of satellite and reference data sets (sampling and overpass times, penetration depths, representativeness errors, etc.; *Wang et al.*, 2012; *Albergel et al.*, 2008; *Gruber et al.*, 2013a; *Nicolai-Shaw et al.*, 2015; *Colliander et al.*, 2017), which is required before calculating various performance metrics (correlation coefficients, root-mean-square-differences, triple collocation analysis, etc.; *Entekhabi et al.*, 2010a; *Albergel et al.*, 2013; *Gruber et al.*, 2016a; *Loew et al.*, 2017). Given the complexity of the validation problem, however, ambiguous results for the quality and ranking of satellite soil moisture products can be found in the literature (e.g., *Wagner et al.*, 2014) depending on which pre-processing and evaluation strategies were followed and which reference data were used. This paper is a community effort that addresses this issue and aims towards standardizing good practices for the validation of satellite-based near-surface soil moisture retrievals.

Section 2 provides a review of on-going activities regarding the standardization of satellite soil moisture validation activities. Section 3 describes the most common reference data sources used for soil moisture validation. Section 4 discusses relevant theoretical aspects and the most common methods (including data pre-processing) for assessing soil moisture data quality. Section 5 presents a community-agreed validation guidance protocol with an example implementation of that protocol provided in Appendix A. Finally, Section 6 discusses research gaps that should be addressed in the near future.

2 Towards standardized validation practices

Many efforts have been made to assess and standardize validation practices across Earth observation (EO) communities (*Zeng et al.*, 2015; *Loew et al.*, 2017; *Su et al.*, 2018). In this section we review activities most relevant for satellite soil moisture products.

2.1 CEOS LPV

The main authority that guides validation activities for satellite-retrieved data of biogeophysical variables is the Committee on Earth Observation Satellites (CEOS) Working Group on Calibration and Validation (`http://ceos.org/ourwork/workinggroups/wgcv/`; last access: 1 July 2019). Activities related to soil moisture are coordinated by its Land Product Validation (LPV) subgroup (`https://lpvs.gsfc.nasa.gov/`; last access: 1 July 2019). The CEOS LPV defines four validation stages (see Table 1) that represent the level of sophistication of validation protocols employed for a particular data product. Relevant for the work presented here is that reaching validation stage 3 requires the implementation of a sophisticated validation framework, as illustrated in Figure 1. In such a framework, standardized community-agreed methods that are ideally described in a "Validation Good Practice Document" should be employed using fiducial reference data (see Sec. 3) to generate standardized validation reports. With this paper we aim at providing such a document. The last validation stage 4 is reached once these validation reports are updated on a regular (at least annual) basis.

2.2 Quality Assurance Frameworks

The CEOS endorses the Quality Assurance Framework for Earth Observation (QA4EO; `http://qa4eo.org/`; last access: 1 July 2019) as a framework to facilitate the provision of traceable quality indicators which "shall provide sufficient information to allow all users to readily evaluate the 'fitness for purpose' of the data or derived product" (*QA4EO*, 2010). The QA4EO provides top-level guidance documents and templates that encourage the use of metrological principles (see Sec. 2.3).

In 2014, the Quality Assurance for Essential Climate Variables (QA4ECV; `http://www.qa4ecv.eu/`; last access: 1 July 2019) project was initiated to developed a set of guidelines for the provision of traceble quality information taking in to account the key principles of QA4EO (*Scanlon et al.*, 2017). To demonstrate how reliable and traceable quality information can be provided, quality assurance frameworks were developed for selected ECVs (not including soil moisture; e.g., *Peng et al.*, 2017). The guidelines developed by QA4EO and QA4ECV are currently embraced by the Copernicus Climate Change Service (C3S; `https://climate.copernicus.eu/`; last access: 1 July 2019) in order to build quality assured, fully traceable Climate Data Records.

In 2018, the Quality Assurance for Soil Moisture project (QA4SM; `https://qa4sm.eodc.eu/`; last access: 1 July 2019) was launched, specifically to create an online validation tool that employs a community-agreed validation protocol (which is described in this paper) for automatically and regularly generating soil moisture product validation reports, thereby addressing the CEOS validation framework requirements (see Figure 1).

2.3 Metrology and traceability

The CEOS and the QA4EO encourage the use of metrological principles for validation purposes, which are described in the "Guide to the expression of uncertainty in measurement" (GUM; *JCGM*, 2008). The GUM is a reference document of the metrological community that provides strict guidelines on how quality estimates of measurements should be obtained and reported. In essence, it states that, since they never perfectly represent the true state of the physical quantity being measured, all measurements should be complemented by uncertainty estimates that summarize their probability density function (pdf). Furthermore, it states that these uncertainties should be obtained by propagating the uncertainties from all components that contribute to the measurement process in a way that is traceable back to the "International System of Units" (SI) standards, either through the standard method for the propagation of uncertainty (*Parinussa et al.*, 2011; *Merchant et al.*, 2017) or, if not possible analytically, through Monte Carlo simulations (*JCGM*, 2008).

However, while being relatively straightforward in a laboratory or numerical environment, the traceable propagation of uncertainties in space borne remote sensing measurements and retrievals thereof, in particular of soil moisture, faces two particular challenges. First, footprints of current microwave instruments used for retrieving soil moisture span over tens to thousands of square kilometers, thereby covering a large variety of climatic, topographic, and land cover conditions. Although certain large-scale homogeneous regions are used for calibrating instruments and determining Level 1 (L1) backscatter or brightness temperature uncertainties (e.g., rainforests or polar snow fields; *Figa-Saldaña et al.*, 2002; *Macelloni et al.*, 2006), it is virtually impossible to obtain global perfectly traceable uncertainty estimates representing all possible measurement conditions. Second, uncertainty propagation assumes that the models used to propagate uncertainties are themselves perfect (*Parinussa et al.*, 2011). For satellite soil moisture retrievals, this is particularly problematic because uncertainties resulting from simplifications and assumptions in both the L1 processing (i.e. geometric correction and radiometric calibration) and the Level 2 (L2) soil moisture retrieval algorithms cannot be accounted for. The soil moisture and other EO communities have established certain strategies to recover this broken traceability chain by validating the soil moisture estimates post retrieval against a range of reference data from various sources. Section 3 will discuss the requirements and current availability of such reference measurements suited for validation activities. Before entering those discussions, it is necessary to provide some relevant terminology.

2.4 Terminology

The CEOS and the QA4EO encourage the use of the terminology used within the metrological community as described in the "International Vocabulary of Metrology" (VIM; *JCGM*, 2012). However, there is a certain level of ambiguity in the existing EO literature, and even within the VIM and the GUM, regarding the usage of important terms such as errors, uncertainties, validation, and others. For a comprehensive summary of the most common definitions (from the VIM, the CEOS, and other sources) we refer the reader to *Loew et al.* (2017). For the purpose of this paper we stress that:

- the term *error* refers to the deviation of a single measurement from the true value of the quantity being measured (which is hence always unknown), whereas the term *uncertainty* refers to the probability distribution underlying an error. For validation purposes, this probability distribution is the actual quantity of interest;

- according to the GUM, the uncertainty of a measurement generally contains both systematic and random components. The laboratory environment of metrological practices typically allows for thorough measurement calibration, where it is assumed that systematic errors can be properly determined and corrected. Satellite soil moisture retrievals, however, usually contain considerable systematic errors which, especially for model calibration and refinement, provide better insight when estimated separately. Therefore, we use the term *bias* to refer to systematic errors only and the term *uncertainty* to refer to random errors only, specifically to their standard deviation (or variance);

- in the EO validation literature, bias is commonly defined as the temporal mean difference between two data sets. We follow the broader statistical definition of bias as auto-correlated error, or as a property of an estimator to systematically over- or underestimate some quantity (*Dee*, 2005). For better separability of its components, we use the terms *first-order bias* and *second-order bias* to refer more specifically to additive and multiplicative systematic errors, respectively (see Sec. 4.4.1);

- the terms *trueness*, *precision*, and *accuracy* are popular antonyms for systematic errors, random errors, and the combined systematic plus random errors, respectively (*JCGM*, 2012). However, trueness and precision are very rarely used in the soil moisture validation literature and the term accuracy is often ambiguously used to refer to either systematic or random errors alone;

- in Earth sciences, the term *validation* is often distinguished from the term *evaluation* such that validation is used to refer to bias or uncertainty assessment using highly accurate or at least well traceable in situ reference data (often misleadingly referred to as "ground truth"; see Sec. 4.2), whereas evaluation is used to refer to the comparison against other coarse-resolution satellite or modelled data with supposedly less well-defined uncertainties. However, technically, validation more specifically refers to quantitative data quality assessment (*Justice et al.*, 2000) whereas evaluation more broadly refers to "the process of judging something's quality" (*Loew et al.*, 2017). For simplicity, we use the term *validation* in this paper to refer to the process of estimating biases and uncertainties regardless of the reference data source used;

- the concept of uncertainty is closely related to the concept of confidence intervals. Both aim at describing the pdf underlying an estimate, although the term *uncertainty* is more commonly used for describing the pdf behind an estimate that results from measurement errors (see Sec. 4.1), whereas the term *confidence interval* is more commonly used for describing the pdf behind statistical parameters (such as statistical moments or validation metrics that derive from these moments) that results from finite sample sizes (see Sec. 4.5).

3 Reference data

The term *fiducial reference measurements* is often used to refer to a suite of independent, fully characterized, and traceable measurements that meet the requirements on *reference standards* as described by QA4EO (*Fox*, 2010), which should be used to assess the quality of EO products. However, although highly accurate in situ soil moisture measurement techniques exist and uncertainties of the measurement devices can be reliably determined through laboratory and field calibration (*Cosh et al.*, 2004, 2006; *Rüdiger et al.*, 2010), using such point-scale measurements for validating satellite soil moisture data sets over large areas is a very difficult task owing to the coarse resolution of space borne microwave instruments and vast heterogeneities across landscapes (*Famiglietti et al.*, 2008; *Brocca et al.*, 2010a; *Miralles et al.*, 2010; *Crow et al.*, 2012; *Nicolai-Shaw et al.*, 2015; *Molero et al.*, 2018).

For satellite validation purposes, numerous field and airborne campaigns have been carried out to obtain reliable satellite footprint scale reference data and to quantitatively assess the potential spatio-temporal representativeness (see Sec. 4.2) of single or small sets of in situ soil moisture stations (*De Rosnay et al.*, 2006; *Brocca et al.*, 2012; *McNairn et al.*, 2015). Additionally, validation activities are complemented with land surface model output and other satellite products for comparison to get as complete a picture as possible of a product's error characteristics (*Brocca et al.*, 2010b; *Draper et al.*, 2013; *Al-Yaari et al.*, 2014; *Dorigo et al.*, 2015; *Kerr et al.*, 2016; *Miyaoka et al.*, 2017). The various reference data sources and their limitations are discussed below. A list of publicly available reference data sources that are commonly used for satellite soil moisture validation is provided in Table 2.

3.1 Field campaigns

Field campaigns are labour-intensive studies that use highly accurate measurement techniques to obtain reliable and traceable representations of larger scale average soil moisture. Unfortunately, these campaigns provide only some snapshots in time, whereas the validation of satellite products requires long and consistent time series (see Sec. 4.4). Therefore, some field campaigns have identified and set up a limited number of permanent measurement stations (<15) at temporally stable locations (*Vachaud et al.*, 1985; *Starks et al.*, 2006) that sufficiently capture sub-pixel heterogeneities, allowing the continuous observation of satellite footprint-scale areas with sufficient and well characterized accuracy.

Ground measurements are often supplemented with airborne observations, which can be used to either directly validate the L1 satellite measurements or the derived soil moisture retrievals over a much larger area, allowing to evaluate spatial soil moisture variability across multiple satellite grid cells. Moreover, they can provide valuable information about soil moisture (or backscatter/brightness temperature) sub-pixel variability.

Early field campaigns were focused on understanding large-scale soil moisture dynamics with aircraft support such as HAPEX-MOBILHY (*Noilhan et al.*, 1991), BOREAS (*Cuenca et al.*, 1997), and the Washita'92 campaigns (*Jackson et al.*, 1995), assessing the potential of soil moisture monitoring as a part of hydrologic modeling. This evolved into satellite associated field campaigns such as the 1997 Southern Great Plains Hydrology Experiment (SGP97) and the Soil Moisture Experiments (SMEX) in 2002-2004 in the United States (*Jackson et al.*, 1999, 2005; *Bindlish et al.*, 2006, 2008), the National Airborne Field Experiments (NAFE) in Australia (*Panciera et al.*, 2008), the Australian Airborne Calibration/Validation Experiments for SMOS (AACES; *Peischl et al.*, 2012), the Canadian Experiment in Soil Moisture (CANEX-10; *Magagi et al.*, 2013), and the CAROLS airborne campaigns (*Albergel et al.*, 2011; *Zribi et al.*, 2011). These campaigns established a protocol for the synchronous collection of ground-based soil moisture measurements with airborne microwave instrumentation, which were supplemented with long-term in situ monitoring stations, thus providing long-term high density validation sites for satellites.

In the process of developing standardized data collection protocols, these field campaigns specifically focused on the investigation of the spatial distribution of soil moisture and its evolution with drying or wetting, the soil moisture variability across scales, and the statistical relationship between spatial standard deviation and extent scale. These parameters drive the potential representativeness of in situ measurements for coarse soil moisture product validation and their knowledge hence allows the determination of the number of ground samples required to obtain sufficiently reliable validation reference data (*Famiglietti et al.*, 2008).

3.2 In situ networks

A large number of in situ soil moisture networks exist worldwide with different quality and spatial sampling densities as well as varying sensing depths (*Dorigo et al.*, 2011b; *Babaeian et al.*, 2019). For validation purposes, the soil moisture community distinguishes between dense networks, which have a large number of soil moisture stations located within single satellite footprints, and sparse networks, where footprint-scale areas usually contain only a single or very few soil moisture stations, although the quantitative cut-off between the two is not well-defined. The overall global coverage of in situ soil moisture networks (accessible and suited for satellite soil moisture validation) is unevenly distributed across the globe and particularly scarce in the tropical regions, the Southern Hemisphere and boreal regions (Fig. 2; *Ochsner et al.*, 2013).

3.2.1 Dense networks

To meet the requirements on fiducial reference data (*Fox*, 2010), the SMAP Calibration and Validation (Cal/Val) Team defined certain criteria for dense measuring networks, so-called core validation sites, ensuring that they provide a traceable representation of footprint-scale soil moisture and therefore allow for a reliable assessment of satellite soil moisture data quality. Currently, 18 densely stationed and thoroughly calibrated in situ measurement sites fulfill these requirements (*Jackson et al.*, 2012; *Colliander et al.*, 2017), operated by independent SMAP Cal/Val partners.

These SMAP Cal/Val partners have a diverse heritage. Some networks were deployed for Cal/Val of the AMSR-E product (*Jackson et al.*, 2010) or SMOS (*Djamai et al.*, 2015), while others evolved from hydrologic monitoring networks (*Bogena et al.*, 2018) or from some other purpose such as aircraft validation projects like AIRMOSS (*Moghaddam et al.*, 2010). During the SMAP project, several networks were selected as potential candidate sites for Cal/Val activities. The candidate networks whose accuracy versus physically collected volumetric soil moisture was already demonstrated and documented in a traceable manner, were promoted to core validation sites. To date, these sites are considered to provide the best possible ground reference data for satellite footprint-scale soil moisture dynamics (*Colliander et al.*, 2017).

3.2.2 Sparse networks

A host of other operational and experimental in situ sites exist worldwide, operating soil moisture measurement stations that are potentially suited for soil moisture validation yet with a considerably smaller station density and often lacking information on their coarse-scale representativeness and their own inherent error characteristics (*Gruber et al.*, 2013a; *Chen et al.*, 2017). Nonetheless, these sites are valuable to complement core validation sites due to their considerably larger spatial coverage across a variety of climatic regimes and biomes (see Sec. 4).

An important source for data from sparse networks is the International Soil Moisture Network (ISMN; *Dorigo et al.*, 2010, 2011b), which is a data hosting facility that harmonizes soil moisture measurements from in situ networks worldwide, applies automated and uniform quality control procedures to flag suspicious measurements (*Dorigo et al.*, 2013), and distributes them on their website (`http://ismn.geo.tuwien.ac.at/`; last access: 1 July 2019) on a cost-free basis in a common format. The ISMN was established by ESA in the framework of SMOS Cal/Val activities. Currently, it contains data from more than 2400 stations worldwide, operated across 59 different measurement networks (see Figure 2) including historical networks that are no longer operational.

3.3 Model simulations

Due to the limited coverage and representativeness of ground reference data, validation activities are complemented with soil moisture simulations from land surface models (LSMs) as an alternative reference data source (*Lahoz and De Lannoy*, 2014). Model simulations can provide spatially complete global soil moisture maps at a spatial (grid) resolution similar to that of satellite footprints, but they may still contain considerable representativeness errors (see Sec. 4.2) originating from simplifications of sub-grid heterogeneities, a scale-mismatch of the underlying atmospheric forcing data, errors in the model parameterization, or simply because the meaning of the modelled “soil moisture” is different. Moreover, biases and uncertainties in model simulations are highly variable and often also not well quantified (*Koster et al.*, 2009; *Albergel et al.*, 2013), making it difficult to separate satellite retrieval errors from modelling errors in a direct comparison (see Sec. 4).

Some examples of readily available global model-based data sets that have been used for satellite soil moisture validation activities (*Albergel et al.*, 2012; *Al-Yaari et al.*, 2014; *Kerr et al.*, 2016; *Dorigo et al.*, 2017; *Gruber et al.*, 2017; *Miyaoka et al.*, 2017) include simulations from NASA's Global Land Data Assimilation System (GLDAS; *Rodell et al.*, 2004), NASA's Modern-Era Retrospective analysis for Research and Applications (MERRA) land data products (*Reichle et al.*, 2011, 2017a), and the European Center for Medium-Range Weather Forecasts (ECMWF) Land Surface Reanalysis (ERA-Interim/Land) data sets (*Balsamo et al.*, 2015).

3.4 Satellite products

A multitude of soil moisture products from different satellite sensors (*Babaeian et al.*, 2019) are commonly used as additional coarse resolution reference data sets for validation purposes, either for consistency assessment through direct comparison (*Al-Yaari et al.*, 2014; *Burgin et al.*, 2017), or within triple collocation analysis (*Dorigo et al.*, 2010; *Draper et al.*, 2013, see Sec. 4). Like model simulations and sparse networks, they typically lack reliable and traceable bias and uncertainty characterization. Also, available satellite sensors observe at different wavelengths, polarizations, and incidence angles and have therefore a varying sensitivity to soil moisture (*Ulaby et al.*, 2014). Hence, the information gleaned from a direct comparison is limited (see Sec. 4.4.2). Furthermore, different satellite retrieval products (and model simulations) can use similar ancillary information such as temperature and/or vegetation information in a radiative transfer model, resulting in correlated errors (*Gruber et al.*, 2016b) which may complicate a fair data comparison (see Sec. 4.4.2). Comprehensive lists of commonly used and publicly available satellite soil moisture products, including some validation information where available, can be found at `https://lpvs.gsfc.nasa.gov/producers2.php?topic=SM` (last access: 1 July 2019) and in *Babaeian et al.* (2019).

4 Theory

This section provides the theoretical background for error characterization and how it relates to satellite soil moisture validation, including the assumptions, limitations and pre-processing steps involved. Although our main focus here is the validation of near-surface satellite soil moisture products, many of the principles discussed below can be equally applied to assess the quality of soil moisture products from other sources, as well as of other biogeophysical variables (*Loew et al.*, 2017).

4.1 Errors

A measurement error e_x is defined as the deviation of a measurement x, in our case a satellite soil moisture retrieval, from the true state t of the quantity under observation (*JCGM*, 2008):

$$e_x = x - t \tag{1}$$

Important for understanding errors is that the "truth" is a hypothetical concept. For the case of space borne microwave measurement instruments, actual satellite footprints are overlapping elliptical areas with strong signal intensity gradients from the footprint center outwards (depending on the antenna gain pattern) and varying surface property dependent vertical support (*Ulaby et al.*, 2014). Horizontal footprint boundaries are commonly defined as the -3 dB region (i.e. the antenna main beam region that covers 50% of the signal's power). Products derived thereof are typically sampled onto spatial grids with sharp boundaries between grid cells and a constant layer depth to facilitate further geospatial analysis (*Bartalis et al.*, 2006; *Brodzik et al.*, 2012; *Bauer-Marschallinger et al.*, 2014). The "true" soil moisture signal that drives the microwave measurement and the subsequent gridded soil moisture retrieval will therefore never be the real average soil moisture of the grid cell to which a measurement is assigned. Moreover, for validation purposes, the unknown "truth" is approximated by reference data, which themselves contain errors and may also be driven by a soil volume that is different from the satellite grid cell they are supposed to represent (see Sec. 3).

4.2 Representativeness

The difference between the true soil moisture that actually affects a (microwave) measurement associated with a particular grid cell and the true soil moisture within that grid cell is often referred to as representativeness error (*Gruber et al.*, 2016a). However, it is worth noting that representativeness errors have different definitions (*Van Leeuwen*, 2015). The remote sensing community mostly assigns them to the mismatch between the spatial support of a measurement and the spatial resolution of the defined sampling grid, sometimes also referred to as scaling error (*Miralles et al.*, 2010; *Crow et al.*, 2012; *Gruber et al.*, 2013a; *Molero et al.*, 2018). In the modelling community, representativeness errors mostly refer to a model's lacking ability to represent reality and, as such, to imperfections in the model structure and in parameterization (e.g., unresolved sub-grid scale processes). For the purpose of data validation, it is practical to use a definition that potentially allows us to separate representativeness errors from other error sources upon estimation. Therefore, recall that the general definition of error in Eq. (1) requires the choice of a "truth", which is the soil moisture state within a target volume (grid cell) that one aims to estimate as accurately as possible. We define representativeness errors as those deviations of a product from that chosen "true" state, which are related to real soil moisture variations. They can occur, for example, if the actual measurement footprint of a satellite extends beyond the grid cell boundaries associated with the "truth", if an inadequate soil parameterization in a radiative transfer model causes the soil moisture retrievals to represent deeper soil layers than the intended "truth", or if point-scale ground measurements are used as a reference for grid cell-scale soil moisture dynamics. As such, representativeness errors of different data sets may be correlated even if the products are otherwise independent.

In summary, representativeness errors have important implications for validation in that they limit the information one can glean from the comparison between products, even if a chosen reference product is itself highly accurate (see Sec. 4.4.1). Since the temporal and spatial resolution and sampling of satellite and available reference measurements hardly ever match, (relative) representativeness errors will often reach considerable magnitudes (*Miralles et al.*, 2010; *Crow et al.*, 2012). To minimize their influence, several pre-processing steps are typically applied, which are discussed in the following section together with other pre-processing steps that are necessary before validation metrics can or should be calculated.

4.3 Pre-processing

Pre-processing steps necessary for validation aim to find match-ups in space and time between measurements that have different spatial resolutions, are sampled on to different grids, and/or are acquired at different times. Additionally, depending on the reference data choice, statistical rescaling methods are often applied to minimize the impact of representativeness errors. Moreover, data pre-processing typically involves the masking of unreliable satellite retrievals and reference measurements. Lastly, data sets are sometimes decomposed into different frequency components in order to separately assess a product's ability of accurately representing short-term, seasonal, and inter-annual soil moisture variability (*Draper and Reichle*, 2015).

4.3.1 Data masking

Satellite-derived soil moisture products are typically accompanied by a set of quality flags. They can be indicators of suspected contamination of the microwave signals or problems during the retrieval. Typical examples are indicators for the probability of frozen soil, dense vegetation coverage, radio frequency interference (RFI), or urban or water contamination, to name a few (e.g., *Parinussa et al.*, 2011; *Naeimi et al.*, 2012; *Kerr et al.*, 2012; *de Nijs et al.*, 2015). The validation of a product should be based only on those retrievals that are considered "good" for later application. While masking data points using binary "use / do not use" flags is straightforward, some quality flags require the decision of a threshold below or above which individual retrievals are masked out (e.g., the probability of RFI occurrence or the water body fraction), which implies a trade-off between data quality and measurement density. Typically, data producers provide recommendations for these thresholds. In addition to the quality flags inherent in the soil moisture products, auxiliary static and/or dynamic data from land surface models or other sources are often used to mask out retrievals that can be considered unreliable, although it should be kept in mind that these sources themselves - and hence quality flags derived thereof - are subject to errors. The most commonly used masking criteria are based on surface and/or air temperature and snow height and/or snow water equivalent estimates obtained from land surface models, or vegetation estimates from satellite sensors or models (*Al-Yaari et al.*, 2014; *Dorigo et al.*, 2015; *Gruber et al.*, 2017). Note that reference data sets, in particular in situ measurements, also often undergo quality control procedures and provide quality flags, which should be used to mask out unreliable measurements before using them to validate satellite retrievals (as is the case for example for the ISMN; *Dorigo et al.*, 2013).

When comparing biases or uncertainties of different soil moisture products, the masking procedures applied to these data sets should be identical in order to compare the quality of retrievals from measurements that were taken under the same (or at least similar) conditions. However, if quality flags that are tailored to one data set are applied to another, some of the products may appear better or worse than they would when using only their own inherent quality control. This is especially true if the flags of one product are much more conservative than those of another. Most product comparison studies do not take this issue into account. One possible approach to address it would be to compare biases and uncertainties from collocated periods also with those in periods where only some products provide unflagged soil moisture retrievals (based on their own quality control) and to put this into perspective with the temporal measurement density before and after product collocation. However, this requires the availability of appropriate reference data in collocated and non-collocated periods as well as the ability to account for possibly varying accuracy and representativeness of the reference data in these periods. Also, depending on the overall data density, it may be difficult to assess biases and uncertainties in these periods due to the presence of large statistical sampling errors (see Sec. 4.5).

Finally, we stress that the choice of data masking criteria has a considerable impact on the overall validation results and should be carefully documented, especially for comparing different validation studies and when assessing long-term changes.

4.3.2 Collocation

Satellite sensors acquire measurements that are irregularly distributed in space and time owing to their orbiting nature and specific antenna patterns. In the soil moisture retrieval process, these measurements are typically sampled onto spatial grids (for noise reduction purposes these grids are often oversampled, i.e. the grid sampling - sometimes also referred to as grid posting - is typically higher than the antenna resolution) and sometimes also to regular time steps (e.g., 00:00 UTC) in order to generate, for example, daily global soil moisture maps and/or time series (*Kerr et al.*, 2012; *O'Neill et al.*, 2012; *H-SAF*, 2018; *Gruber et al.*, 2019a). However, neither the resolution nor the sampling of in situ reference measurements or model simulations ever perfectly match those of the satellite products being validated. Consequently, the process of finding match-ups between satellite and reference data points in space and time, commonly referred to as collocation, is essentially a resampling task (*Loew et al.*, 2017). Since the spatial resolution of the compared products can be very different (especially between in situ and satellite / modelled data), statistical rescaling methods are often additionally applied in the collocation process to minimize the impact of (especially spatial) representativeness errors on validation metrics.

Spatial resampling

In situ measurements are point-scale measurements that sample only a few cubic centimeters of the soil (with the exception of cosmic-ray neutron sensors, which sample areas in the order of hectares; *Zreda et al.*, 2012). When used for validating satellite products, stations from sparse networks are typically sampled onto the satellite grid using a nearest-neighbour (NN) search, i.e. by matching the stations to the satellite grid cells within which they are located (*Albergel et al.*, 2012; *Dorigo et al.*, 2015; *Chen et al.*, 2017). For dense networks, commonly all stations that lie within a particular satellite grid cell are (after quality control) averaged (*Jackson et al.*, 2010; *Gruber et al.*, 2015; *Colliander et al.*, 2017), either by calculating the arithmetic mean or by calculating a weighted average where higher weights are applied to stations that are expected to be more representative for the grid cell average soil moisture. Such stations can be identified, for example, via a temporal stability analysis (*Vachaud et al.*, 1985), through Voronoi diagrams (*Colliander et al.*, 2017), or by using landscape characteristics such as land cover or soil properties.

When comparing different gridded products (i.e. different satellite and/or land surface model products), one grid must be selected as the reference grid onto which the other products are resampled for collocation purposes. This is commonly done using either a NN search or inverse-distance-weighted (IDW) based approaches (*Al-Yaari et al.*, 2014; *Gruber et al.*, 2017, 2019a). However, the resampling provides mainly spatial match-ups of the data sets and can at best account for some of the spatial representativeness errors of the various data sets. How exactly these representativeness errors are affected and propagate into bias and uncertainty estimates will depend on the chosen reference grid and resampling method, and requires more research. The most common way to reduce spatial (systematic) representativeness errors is to apply statistical rescaling methods (see below).

Temporal resampling

In situ measurements and land model estimates are typically sampled more frequently than satellite soil moisture retrievals. Therefore, the reference measurements are matched in time to the irregular satellite observation times, typically by selecting the temporally closest (NN) reference measurement within a pre-defined search window (i.e. applying a maximum temporal distance threshold; *Chen et al.*, 2017). Depending on the sampling interval of the reference data sets (for in situ data typically hourly and for global land surface models typically one to six hourly) and on whether or not satellite observations have been a priori resampled already (see above), this can lead to considerable differences between the actual measurement times of collocated satellite and reference data points. The issue is typically limited when using in situ or model data as reference. However, if multiple satellite products are evaluated simultaneously, their different overpass times are usually accounted for by either picking one of them as (temporal) reference and matching the other ones against it, or by sampling all satellite products to regularized time steps (e.g., 00:00 UTC; *Gruber et al.*, 2017), which in any case favours the satellite data set whose actual measurement times are closest to the reference points. Note that the retrieval quality of satellite data sets may strongly depend on the time of observation. This is especially true for passive systems, where soil moisture retrievals are known to be strongly affected by temporal temperature fluctuations and temperature gradients in soil and vegetation cover (*Parinussa et al.*, 2015).

Taken together, the different measurement times of satellite and reference data sets that have been collocated will induce temporal representativeness errors, originating from the actual soil moisture changes that take place during these periods. Often these errors are assumed to be negligible or at least below the noise level of the products. In principle, one could employ more sophisticated resampling algorithms to minimize these representativeness errors, for example auto-regressive interpolation methods with or without auxiliary information such as precipitation, evapotranspiration, or soil texture. However, more research is needed to assess the impact of temporal interpolation approaches on validation metrics.

(Statistical) rescaling

The resampling procedures described above provide data set match-ups in space and time which are required for statistical comparison (see Sec. 4.4). As discussed in Sec. 4.1, the measurements of the collocated products are driven by the soil moisture state of different soil volumes at different times due to the different underlying actual spatio-temporal resolution of the data sets. The latter is related to the antenna and surface properties and cannot be corrected for by common resampling methods. Therefore, a direct comparison of these products will be subject to representativeness errors, which may dominate the total soil moisture retrieval errors (*Gruber et al.*, 2013a; *Chen et al.*, 2017; *Molero et al.*, 2018). However, owing to the large-scale and auto-correlated nature of processes that drive soil moisture changes (*Crow et al.*, 2012), parts of these errors are systematic and can hence be corrected for by removing *relative differences* between the considered data sets (see Sec. 4.4).

The two most common rescaling approaches are to match either the temporal mean and standard deviation of the data sets that are to be compared (*Scipal et al.*, 2008a; *Dorigo et al.*, 2010; *Albergel et al.*, 2012), or to match their complete cumulative distribution function (CDF), which additionally corrects for differences in higher statistical moments in case the products are expected not to be perfectly Gaussian distributed (*Reichle and Koster*, 2004; *Kumar et al.*, 2012). However, any rescaling approach that transforms one data set into the data space of another (without additional information) assumes the signal-to-noise ratios (SNRs) of the two involved data sets to be identical, which, since this is usually not the case, can lead to biased rescaling parameters that do not fully correct the systematic representativeness errors (see Sec. 4.4.2; *Stoffelen*, 1998; *Yilmaz and Crow*, 2013). Alternatively, triple collocation analysis (*Stoffelen*, 1998; *Su et al.*, 2014; *Gruber et al.*, 2016a) is often employed, using a third data set to take different SNRs into account when matching the standard deviation of the underlying soil moisture signals, thereby potentially providing consistent rescaling parameters (*Yilmaz and Crow*, 2013).

Note that rescaling soil moisture data sets can equally account for (systematic) representativeness errors that arise from different spatial resolution and spatial and temporal misalignment, as well as for those arising from different vertical measurement support, i.e. wavelength-dependent penetration depths of satellites, in situ sensor placement depths, and modelled soil layer thickness (*Gruber et al.*, 2013a). Also, in addition to correcting for systematic representativeness errors, rescaling can implicitly compensate for different units (provided that the used soil moisture representations are linearly related), most commonly volumetric soil moisture ($[m^3 m^{-3}]$) and the degree of soil saturation ([%]) which are linked through soil porosity as a multiplicative factor (*Walker et al.*, 2004). This avoids additional biases that are introduced through the use of inaccurate auxiliary data (such as soil maps) that would otherwise be needed for unit conversion.

After rescaling, long-term bias estimation is obviously no longer meaningful as systematic differences between the data sets, which would normally serve as proxy for biases, have been intentionally removed. However, shorter-term biases as well as random representativeness errors may remain and can considerably contribute to subsequent uncertainty estimates (see Sec. 4.4.1).

4.3.3 Signal decomposition

The quality of soil moisture products can vary considerably across time scales (*Su and Ryu*, 2015; *Draper and Reichle*, 2015; *Molero et al.*, 2018; *Gruber et al.*, 2019a). For example, some soil moisture products are better at accurately representing the seasonal cycle whereas other products more accurately capture short-term fluctuations. Therefore, products are often decomposed into different frequency components which are then validated separately (in addition to the bulk time series). In Earth sciences, such decomposition is often done using moving-average windows (*Narapusetty et al.*, 2009). For soil moisture, a moving window of several weeks, centered on the measurement time, is typically used to obtain intra-annual low-frequency soil moisture dynamics (*Albergel et al.*, 2012; *Chen et al.*, 2017), referred to as seasonalities. Residuals thereof are referred to as short-term anomalies which represent higher-frequency, sub-seasonal soil moisture variations. Additionally, so-called long-term anomalies are often calculated as residuals relative to a multi-year mean seasonal cycle, referred to as the soil moisture climatology, which is typically calculated by applying a moving-average window of similar size (a few weeks) to each day-of-the-year (DOY), i.e. averaging all measurements of all years that fall inside the specified time window around a particular DOY (*Miralles et al.*, 2010; *Draper et al.*, 2013).

While the validation of short-term soil moisture anomalies aims at assessing a data set's capability of capturing individual drying or wetting events, uncertainties of long-term anomalies represent its performance in capturing both short-term variability and inter-annual variations such as prolonged droughts or floods as well as climate trends. However, the latter rely on a climatology estimate that requires historical data records in the order of decades (*Dorigo et al.*, 2012), which are often not available, especially not at the beginning of a new mission (current microwave missions cover a time period of maximum 5-10 years). Therefore, one often has to rely on uncertainty estimates for seasonalities and short-term anomalies alone, which jointly drive uncertainties in long-term anomalies.

4.4 Metrics

After satellite and reference products have been masked, collocated, and optionally decomposed and/or rescaled, validation metrics can be calculated. In this section, we summarize commonly used bias and uncertainty estimators and their underlying assumptions. Other related metrics exist (e.g., the mean absolute error, Kendall's tau, and many others), but all are derived from the same statistical moments and have therefore similar information content. Our goal here is to present the metrics that are most commonly used for soil moisture validation and are considered to provide a comprehensive picture of a product's error characteristics. These metrics also largely coincide with those used in other EO communities (*Loew et al.*, 2017). We also stress that validation specifically aims at quantitatively assessing the errors of a data set, which is different from indirectly evaluating its quality for example by investigating its skill in a particular application, e.g., drought monitoring (*Bolten et al.*, 2010). Such indirect product evaluation is beyond the scope of this paper.

4.4.1 Assumptions

The fundamental assumption underlying almost all satellite soil moisture validation studies is that of additive zero-mean random errors (ε_x), and additive (first-order; α_x) and multiplicative (second-order; β_x) systematic errors (*Gruber et al.*, 2016a):

$$x = \alpha_x + \beta_x t + \varepsilon_x \tag{2}$$

This error model applies to both the data set one aims to validate and the reference data sets. Notice that the total error e_x in Eq. (1) has now been separated into its systematic (α_x and β_x) and random (ε_x) components. These components contain instrument errors (i.e. noise and mis-calibration), errors in the retrieval model and parameterization, and other representativeness errors with respect to the assumed grid cell average soil moisture t (although the boundaries between the latter two are somewhat fuzzy; see Sec. 4.1).

To disentangle errors from different data sets and from actual soil moisture variations, all common data comparison metrics require the errors to be homoscedastic (i.e. independent from the soil moisture state, in the literature often referred to as orthogonality with respect to the truth; *Yilmaz and Crow*, 2014) and mutually uncorrelated between products. Remember, however, that the *representativeness* error components of the different products may (by definition) be correlated both with the truth t and with each other, even if the products are otherwise independent (see Sec. 4.1).

All common validation metrics are derived from the first and second statistical moments of the data sets. This implies that soil moisture too is - even though in principle deterministic - assumed to behave as a random variable. Statistical moments are then typically estimated in the temporal domain (i.e. temporal means, variances, and covariances), assuming stationarity in soil moisture and the errors (i.e. means and variances are assumed to be constant over time), and relate to the various error components as follows:

$$\begin{aligned} \overline{x} &= \alpha_x + \beta_x \overline{t} \\ \sigma_x^2 &= \beta_x^2 \sigma_t^2 + \sigma_{\xi_x}^2 \\ \sigma_{xy} &= \beta_x \beta_y \sigma_t^2 + \sigma_{\xi_x, \xi_y} \end{aligned} \tag{3}$$

where the overline, σ_i^2 and σ_{ij} refer to the (temporal) mean, variance, and covariance, respectively; and y denotes a reference data set that follows the same error model as x (Eq. (2)). Because *representativeness* errors may contain an orthogonal, a non-orthogonal, and a mutually correlated component (see above), we combine it with all other random error in the individual data set's random error variability $\sigma_{\xi_x}^2 = \sigma_{\varepsilon_x}^2 + 2\beta_x\sigma_{t,\varepsilon_x}$ (containing representativeness and all other random errors) and the correlated error variability $\sigma_{\xi_x,\xi_y} = \beta_x\sigma_{t,\varepsilon_y} + \beta_y\sigma_{t,\varepsilon_x} + \sigma_{\varepsilon_x,\varepsilon_y}$ (driven by representativeness errors only), for clarity. Systematic representativeness errors are included in the α_x and β_x coefficients.

The goal of validation is now to estimate α_x and β_x, and the standard deviation of ε_x (σ_{ε_x}), i.e. biases and uncertainties in the satellite data set under validation. The properties of the different reference data sets available (see Sec. 3) determine which error components will be dominant in Eq. (3), and consequently, which ones can be estimated by the available validation metrics (see Sec. 4.4.3 and 4.4.4).

4.4.2 Relative and TCA-based metrics: opportunities and limitations

For discussing the various metrics we will follow the notation of fiducial reference data (see Sec. 3) to refer to data sets that provide a thoroughly calibrated soil moisture proxy at the satellite scale with traceable uncertainty characteristics (i.e. $\alpha_y \approx 0, \beta_y \approx 1$ in Eq. (2)). ε_y may be non-zero but $\sigma_{\varepsilon_y}^2$ has to be at least well determined from laboratory experiments and field campaigns and could hence be corrected for in the validation metrics. As mentioned, only the core validation sites are currently considered as fiducial reference data capable of providing a reliable representation of satellite footprint-scale soil moisture (see Sec. 3.2.1). They are therefore the only reliable proxy for bias and uncertainty estimation from direct comparison, but are limited to very few regions. Non-fiducial reference data refer to coarse-resolution products such as land surface model simulations or other satellite data sets which may have non-negligible or non-traceable biases and uncertainties as well as potentially considerable representativeness errors, or to in situ data from sparse networks or not properly calibrated and validated dense networks, both of which are expected to have larger representativeness errors than coarse-resolution reference data sets. Therefore, direct comparison against non-fiducial reference data can only provide information of which data set is systematically drier or wetter than the other but without relation to a true grid cell average, and only lumped estimates of the uncertainty of both compared products. Nonetheless, given their larger-scale and long-term availability, sparse networks and land surface models are of important complementary value for validating satellite products. In particular, one can obtain valuable information about the relative ranking of different products as well as about performance changes over time when comparing against the same reference product.

Introducing a second reference data set z that follows the same covariance properties (Eq. (3)) as y (commonly referred to as triple collocation analysis, TCA; *Stoffelen*, 1998; *Scipal et al.*, 2008b; *Gruber et al.*, 2016a) allows, under particular circumstances, to simultaneously estimate uncertainties of all three products and also to (partly) isolate random (relative) representativeness errors (*Miralles et al.*, 2010; *Gruber et al.*, 2013a; *Chen et al.*, 2017). Note, however, that the necessity of using two reference data sets instead of one may limit spatial and temporal data availability. Moreover, while non-orthogonal and mutually correlated errors are equally problematic for metrics that rely on one reference data set only (see below), it may be even more difficult to find a third data set that fulfills these requirements. Commonly, any combination of in situ measurements, land surface model estimates, active-microwave-based measurements, or passive-microwave-based measurements is expected to fulfil this requirement because their sources of errors are assumed to be mostly independent (*Gruber et al.*, 2016a), provided that neither of them has been used to generate another (e.g., by assimilating satellite data in to a land surface model; *Reichle et al.*, 2017b,c). However, several studies suggest that mutual error correlations may exist between commonly used data set combinations (*Yilmaz and Crow*, 2014; *Pan et al.*, 2015), resulting from unrecognized common data (e.g., similar vegetation or temperature input) or representativeness errors (e.g., if a land surface model used within TCA models a deeper layer than the sensing depth of two satellite data sets that are used in the triplet). It is therefore recommended to verify orthogonality and zero error correlation assumptions by using - where available - multiple data set triplets and checking for consistency between different TCA implementations (*Dorigo et al.*, 2010; *Draper et al.*, 2013), or by using the recently proposed TCA extension that utilizes four or more data sets to diagnose the existence, and estimate the magnitude of error correlations (*Gruber et al.*, 2016b; *Pierdicca et al.*, 2017).

The following sections discuss the most common bias and uncertainty metrics, either (i) based on direct comparison between two data sets, which will be referred to as relative metrics, or (ii) based on the simultaneous comparison of three products, which will be referred to as TCA-based metrics. All metrics can be equally applied to soil moisture anomaly estimates or the raw time series, except for first-order bias estimators (see below) as the anomaly calculation per definition removes differences in the mean (see Sec. 4.3.3).

Note that none of the metrics presented below require assumptions about the shape of the pdf of the random errors or the true signal (*McColl et al.*, 2016). However, the bounded nature of soil moisture may cause violations in the orthonality assumption if cut-off values (e.g., zero and the soil porosity as lower and upper physical limit, respectively) are applied to the soil moisture estimates of a particular data sets. Especially in very dry or very wet regimes, where random errors would often cause these thresholds to be exceeded, this can result in considerable biases in all (both relative and TCA-based) uncertainty metrics.

4.4.3 Bias estimation

Bias estimation is only meaningful against reference data at the satellite footprint scale, i.e. without considerable representativeness errors and if no rescaling has been applied (see Sec. 4.3.2).

Temporal mean bias

The term bias commonly refers to the (temporal) mean difference between two data sets (*Entekhabi et al.*, 2010a):

$$b_{xy} = \overline{x} - \overline{y} = \alpha_x - \alpha_y + (\beta_x - \beta_y)\bar{t} \tag{4}$$

Typically, b_{xy} is considered to represent first-order (additive) biases only. However, as can be seen in Eq. (4), the mean difference is also sensitive to second-order (multiplicative) biases, amplified by the actual mean soil moisture content ($\bar{t}$). When using non-fiducial reference data, b_{xy} provides an indication of which data set is systematically drier or wetter than the other, but without relation to the assumed true grid cell average. Moreover, a positive difference in the mean ($\alpha_x > \alpha_y$) and a negative difference in variability ($\beta_x < \beta_y$) can cause the same sign in b_{xy} as a negative mean difference and a positive variability difference. When calculated against fiducial reference data, b_{xy} collapses to $\alpha_x + (\beta_x - 1)\bar{t}$. That is, it is a direct estimate for biases in the satellite retrieval, yet it is still susceptible to both first and second-order biases, and influenced by the average soil moisture conditions.

Second-order bias

Most validation studies do not attempt to estimate second-order biases and neglect their impact on b_{xy} and other validation metrics such as the (unbiased) Root-Mean-Square-Difference (see *Gupta et al.* (2009) and Sec. 4.4.4). TCA potentially allows for the direct estimation of second-order biases (*Gruber et al.*, 2016a) as:

$$\beta_x^y = \frac{\sigma_{xz}}{\sigma_{yz}} = \frac{\beta_x \beta_z \sigma_t^2 + \sigma_{\xi_x,\xi_z}}{\beta_y \beta_z \sigma_t^2 + \sigma_{\xi_y,\xi_z}} \approx \frac{\beta_x}{\beta_y} \tag{5}$$

where β_x^y denotes the TCA-based second-order bias estimate of x relative to y which, if y is a fiducial reference data set and if no non-orthogonal or correlated random representativeness errors exist ($\beta_y \approx 1, \sigma_{\xi_x,\xi_z} \approx 0, \sigma_{\xi_y,\xi_z} \approx 0$), provides a direct estimate of the second-order bias β_x. Notice that neither first nor second-order biases in z influence β_x^y. Alternatively, Eq. (5) can also be used for rescaling purposes (*Yilmaz and Crow*, 2013; *Su et al.*, 2014; *Gruber et al.*, 2016a, see Sec. 4.3.2).

4.4.4 Uncertainty estimation

As discussed, uncertainty estimates aim at representing the pdf of the random errors (see Sec. 2), which is typically done by means of their standard deviation (or variance).

(Unbiased) Root-Mean-Square-Difference

The most common relative metric for estimating uncertainty is the Root-Mean-Square-Difference (RMSD; *Entekhabi et al.*, 2010a):

$$\begin{aligned} RMSD_{xy} &= \sqrt{\overline{(x-y)^2}} = \sqrt{(\overline{x}-\overline{y})^2 + \sigma_x^2 + \sigma_y^2 - 2\sigma_{xy}} \\ &= \sqrt{(\alpha_x - \alpha_y + (\beta_x - \beta_y)\overline{t})^2 + (\beta_x - \beta_y)^2 \sigma_t^2 + \sigma_{\xi_x}^2 + \sigma_{\xi_y}^2 - 2\sigma_{\xi_x,\xi_y}} \end{aligned} \tag{6}$$

Since the RMSD is sensitive to both systematic and random errors, the bias component is - for uncertainty estimation purposes - typically removed, resulting in the unbiased RMSD (ubRMSD):

$$\begin{aligned} ubRMSD_{xy} &= \sqrt{RMSD^2 - b_{xy}^2} = \sqrt{\sigma_x^2 + \sigma_y^2 - 2\sigma_{xy}} \\ &= \sqrt{(\beta_x - \beta_y)^2 \sigma_t^2 + \sigma_{\xi_x}^2 + \sigma_{\xi_y}^2 - 2\sigma_{\xi_x,\xi_y}} \end{aligned} \tag{7}$$

The common definition of the ubRMSD specifically corrects for differences between the mean of the data sets (*Entekhabi et al.*, 2010a). However, as can be seen in Eq. (7), it remains susceptible to second-order biases, which are amplified by the actual soil moisture variability (σ_t^2). Moreover, as was the case for b_{xy}, this second-order bias dependency in $ubRMSD_{xy}$ persists even when calculated against fiducial reference data, in which case Eq. (7) collapses to $\sqrt{(\beta_x - 1)^2\sigma_t^2 + \sigma_{\xi_x}^2}$. As discussed in Sec. 4.3.2, data sets are often rescaled before calculating validation metrics to account for systematic representativeness errors, especially when validating against data from sparse networks. This is most commonly done by matching the temporal mean and the standard deviation of the data sets, or their entire cdf (i.e. also higher statistical moments). However, as can be seen from Eq. (3), this only properly corrects for relative differences in β if the SNRs (including random representativeness errors) of the data sets are equal, which is very unlikely. Consequently, Eq. (7) will still contain the remaining difference between β_x and the rescaled β_y, multiplied with the actual soil moisture variability, and also random representativeness errors.

(Unbiased) Root-Mean-Square-Error

As mentioned in the previous section, TCA potentially allows for the estimation of relative rescaling coefficients that are independent from the SNRs of the data sets (see Eq. (5)), which would allow to fully correct for the second-order bias component in Eq. (7). Moreover, TCA allows to more directly estimate the satellite uncertainty (i.e. its error standard deviation σ_{ξ_x}, commonly referred to as unbiased Root-Mean-Square-Error; ubRMSE) as:

$$\begin{aligned} ubRMSE_x &= \sqrt{\left|\overline{(x-y)(x-z)}\right|} = \sqrt{\left|\sigma_x^2 - \frac{\sigma_{xy}\sigma_{xz}}{\sigma_{yz}}\right|} \\ &= \sqrt{\left|\beta_x^2\sigma_t^2 + \sigma_{\xi_x}^2 - \frac{(\beta_x\beta_y\sigma_t^2 + \sigma_{\xi_x,\xi_y})(\beta_x\beta_z\sigma_t^2 + \sigma_{\xi_x,\xi_z})}{\beta_y\beta_z\sigma_t^2 + \sigma_{\xi_y,\xi_z}}\right|} \approx \sigma_{\xi_x} \end{aligned} \tag{8}$$

Note that when calculating the ubRMSE using the cross-multiplied differences instead of the statistical moments, the data sets y and z do have to be bias-corrected with respect to x a priori using Eqs. (4) and (5). The absolute value is taken to prevent negative signs in uncertainty estimates that could occur due to sampling errors (*Gruber et al.*, 2018, see Sec. 4.5). As one can see, $ubRMSE_x$ is (as opposed to $ubRMSD_{xy}$ in Eq. (7)) fully unbiased in that it contains neither first nor second-order biases from both the satellite and the validation data sets, and it also no longer contains the uncertainties inherent in the reference data products (*Gruber et al.*, 2016a). However, estimates that are unbiased *with respect to the assumed true grid cell average* can only be obtained if at least one fiducial reference data set is available (*Chen et al.*, 2017). Moreover, $ubRMSE_x$ is not affected by random representativeness errors in y and z as long as they are orthogonal and not correlated. Such representativeness error correlations could occur for example when applying TCA to in situ measurements together with two coarse resolution products. This case, however, provides an opportunity to estimate the representativeness of in situ stations while uncertainty estimates for the coarse resolution products remain unaffected (*Miralles et al.*, 2010; *Gruber et al.*, 2013a; *Chen et al.*, 2017). For a more detailed derivation of how representativeness errors affect the TCA-based uncertainty estimates we refer the reader to *Vogelzang and Stoffelen* (2012) and *Gruber et al.* (2016a).

The above described metrics are direct estimators for data set uncertainty. However, for many applications, how "good" a data set is depends on how large its uncertainties are relative to the variability of the actual soil moisture signal. Simply put, the larger the soil moisture variations one strives to observe, the more easily they can be distinguished from noise in the measurements. Therefore, some metrics aim at estimating the SNR rather than the uncertainty alone, the most important ones for soil moisture validation being discussed below.

Pearson correlation coefficient

The most common SNR-related relative metric is the linear (Pearson) correlation coefficient, which is typically described as a measure for statistical dependency between two data sets. From the error model in Eq. (3) one can see that it is also a direct, normalized (between -1 and 1) representation of the SNRs of the two data sets for which it is calculated (*Gruber et al.*, 2016a):

$$
\begin{aligned}
R_{xy} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} &= \frac{\beta_x \beta_y \sigma_t^2 + \sigma_{\xi_x,\xi_y}}{\sqrt{(\beta_x^2 \sigma_t^2 + \sigma_{\xi_x}^2)(\beta_y^2 \sigma_t^2 + \sigma_{\xi_y}^2)}} \\
&\approx \operatorname{sgn}(\sigma_{xy}) \frac{1}{\sqrt{(1 + SNR_x^{-1})(1 + SNR_y^{-1})}}
\end{aligned}
\tag{9}
$$

with $SNR_x = \frac{\beta_x^2 \sigma_t^2}{\sigma_{\xi_x}^2}$ and $SNR_y = \frac{\beta_y^2 \sigma_t^2}{\sigma_{\xi_y}^2}$. $sgn(\cdot)$ denotes the signum function. When calculated against fiducial reference data, R_{xy} is a direct representation of the SNR of the satellite under validation (i.e. SNR_x). Notice that the "signal" to which the "noise" in the SNR estimator is related is the true soil moisture variability scaled with the second-order satellite bias (i.e. $\beta_x^2 \sigma_t^2$). Even if β_x could be estimated reliably, for example from Eq. (5), rescaling does not change the SNR as the uncertainty would be scaled as well. However, the ratio $\frac{\beta_x^2 \sigma_t^2}{\sigma_{\xi_x}^2}$ is in fact the quantity of interest that determines how well signal variations can be distinguished from noise, regardless of whether systematic errors have been corrected for (*Gruber et al.*, 2016a), which can be also interpreted as the (linear) correlation with the true soil moisture signal (*McColl et al.*, 2014). When R_{xy} is calculated against non-fiducial reference data, it is additionally influenced by second-order systematic and random representativeness errors as well as the uncertainties of that reference data set.

TCA-based correlation coefficient

Influences of the reference data set can be again isolated using TCA (*McColl et al.*, 2014) by directly estimating R_x as:

$$\begin{aligned} R_x &= \sqrt{\left|\frac{\sigma_{xy}\sigma_{xz}}{\sigma_x^2\sigma_{yz}}\right|} = \sqrt{\left|\frac{(\beta_x\beta_y\sigma_t^2 + \sigma_{\xi_x,\xi_y})(\beta_x\beta_z\sigma_t^2 + \sigma_{\xi_x,\xi_z})}{(\beta_x^2\sigma_t^2 + \sigma_{\xi_x}^2)(\beta_y\beta_z\sigma_t^2 + \sigma_{\xi_y,\xi_z})}\right|} \\ &\approx \sqrt{\left|\frac{\beta_x^2\sigma_t^2}{\beta_x^2\sigma_t^2 + \sigma_{\xi_x}^2}\right|} = \frac{1}{\sqrt{1 + SNR_x^{-1}}} \end{aligned} \tag{10}$$

As was the case for the ubRMSE, the validity of Eq. (10) requires that there is no correlation or non-orthogonality between random representativeness errors, but their individual variance may well be non-zero. If these assumptions are respected, then R_x will be an unbiased representation of the correlation between x and the (unknown) hypothetical truth. Consequently, R_x will always be larger than R_{xy} although this difference decreases as the quality of the reference y increases. Note, however, that R_x only ranges between 0 and 1, as an anti-correlation (with respect to the true signal) cannot be unambiguously inferred from the three covariances in Eq. (10).

(Logarithmic) Signal-to-Noise Ratio

Instead of expressing the SNR normalized between 0 and 1, it is often estimated directly and linearized by converting it into decibel (dB) units (*Gruber et al.*, 2016a):

$$SNR_x[dB] = -10\log\left(\left|\left|\frac{\sigma_x^2\sigma_{yz}}{\sigma_{xy}\sigma_{xz}}\right| - 1\right|\right) \approx 10\log\left(\frac{\beta_x^2\sigma_t^2}{\sigma_{\xi_x}^2}\right) \tag{11}$$

This provides a more direct, linear representation of the ratio between soil moisture and uncertainty magnitude than R_x, yet the information content in both metrics is identical; it is simply a different way of presentation. Note that the SNR_x is already being used as a more coherent (than RMSD or RMSE based metrics) satellite data quality indicator for defining target accuracy requirements (see Sec. 4.8.2).

4.5 Statistical significance testing

All the above described (and also most other less common) validation metrics are based on statistical moments, sampled in time. Since these estimates are based on finite samples (i.e. the discrete soil moisture time series), they are subject to sampling errors. The most common way to deal with statistical uncertainty (i.e. sampling errors) across science communities is Null Hypothesis Significance Testing (NHST) using p-values and/or confidence intervals (*Wilks*, 2011). In a validation context, typical hypotheses to be nullified are, for example, that a soil moisture product does not meet a target accuracy threshold or that one product does not exhibit higher correlation with a reference product than another. For testing such hypotheses, the sampling distribution of the statistical estimate under consideration (such as a validation metric) is constructed based on the magnitude of the estimate and the size of the sample used to draw this estimate (see below). Then, either the p-value is calculated, which is the probability of values of the sampling distribution to be equal to or below (or above, depending on which tail is considered) the pre-defined Null-value (representing the Null hypothesis), or the $(1-\alpha)\cdot 100\%$ confidence interval is considered. A rejection of the Null-hypothesis is considered statistically significant, if the p-value is below a pre-defined significance level α (typically 0.05) or if the $(1-\alpha)\cdot 100\%$ confidence interval does not contain the Null-value. When comparing estimates of different samples (e.g., the performance of different soil moisture products), it is common to consider their relative difference as statistically significant if their confidence intervals do not overlap. Note that the term "Null-value" refers to the Null hypothesis and not to a value of zero of the test statistic (i.e. the validation metric). A common Null-value for testing soil moisture accuracy requirements, for instance, is 0.04 $\mathrm{m^3m^{-3}}$ ubRMSD (see Sec. 4.8.2). Hence, if the p-value for 0.04 $\mathrm{m^3m^{-3}}$ of the sampling distribution around an estimated ubRMSD is below the defined α level, the product is said to meet accuracy requirements with statistical significance.

However, the American Statistical Association (ASA) has recently issued a statement on statistical significance and p-values (*Wasserstein and Lazar*, 2016) warning about the science-wide misuse and abuse of NHST through the replacement of scientific reasoning with a dichotomous and arbitrary classification of results into "significant" or "non-significant". In this statement, the ASA is advocating the abandonment of statistical significance testing altogether for two main reasons. The first one is that an alarming fraction of articles in the scientific literature present unjustified inferences based on misinterpreted p-values and confidence intervals (*Greenland et al.*, 2016; *Gelman and Stern*, 2006; *Wasserstein and Lazar*, 2016). For example, overlapping confidence intervals are often wrongly considered to imply a non-significant difference between two estimates. The second and more important argument against significance testing is that p-values alone provide no grounds for meaningful decision making. While the magnitude of p itself may be informative about how consistent the data at hand are with an assumed stochastic model, "[...] a label of statistical significance does not mean or imply that an association or effect is highly probable, real, true, or important. Nor does a label of statistical nonsignificance lead to the association or effect being improbable, absent, false, or unimportant." (*Wasserstein et al.*, 2019). Therefore, no practical conclusion or decision should be based on whether p-values do or do not meet an arbitrarily defined threshold.

In a recent special issue of The American Statistician (*Wasserstein et al.*, 2019), the statistical community is aiming to propose more appropriate alternatives. Their key message is that, naturally, there should not and cannot be a one-fits-all approach or threshold for statistical/scientific inference. Instead of strictly yet arbitrarily categorizing study results based on dichotomous significance tests, one should strive for more careful study design and more rigorous understanding, interpretation and reporting of the stochastic properties of the data at hand (*Greenland et al.*, 2016; *Tong*, 2019).

In conclusion, for soil moisture validation purposes, we follow the above guidance and recommend to avoid any statement or interpretation about statistical "significance" or "non-significance" and to instead always provide and interpret a statistical summary of calculated validation metrics in the form of confidence intervals alongside the metrics themselves. How confidence intervals can be calculated and recommendations of how they can be presented are described in the following sections.

4.6 Confidence intervals

In general, confidence intervals represent the pdf of the sampling errors of an estimate and are defined at a certain confidence level. A confidence level of, say, 95% means that if one would repeatedly calculate 95% confidence intervals in a series of similar experiments, then 95% of them would - on average - contain the true value, provided that all assumptions made for the stochastic model are met. Note that this is *not* the probability that the true value that is approximated by the estimate lies within the confidence interval (*Neyman*, 1937; *Greenland et al.*, 2016). In theory, this probability - which would indeed be more informative - could be represented by a Bayesian credible interval, but calculating it would require a priori knowledge about the pdf of the parameter that is being estimated (i.e. the so-called "prior") and this is typically not available.

Estimating confidence intervals for validation metrics is not always straightforward because the sampling error pdfs of the various estimators are often not well understood or contain parameters that are typically unknown (*Zwieback et al.*, 2012). The only validation metrics (presented here) for which analytical solutions for confidence intervals exist are the temporal mean bias (b_{xy}), the unbiased RMSD ($ubRMSD_{xy}$), and the Pearson correlation coefficient (R_{xy}). For TCA-based metrics, one has to rely on bootstrapping (*Efron and Tibshirani*, 1986) to approximate the sampling error pdf.

4.6.1 Analytical calculation

The sampling errors in b_{xy} and $ubRMSD_{xy}$ are equivalent to the sampling errors of the population mean and the population standard deviation of the difference series $u = x - y$, which are known to follow a t-distribution and a χ-distribution, respectively (*Gilleland*, 2010; *De Lannoy and Reichle*, 2016):

$$\frac{\overline{u} - \mu_u}{\frac{s_u}{\sqrt{n}}} \sim t_{n-1} \tag{12}$$

and

$$\frac{\sqrt{n-1}\, s_u}{\sigma_u} \sim \chi_{n-1} \tag{13}$$

where n is the sample size; $\overline{u}$ and s_u represent the sample mean and standard deviation of the difference series $(x - y)$; and μ_u and σ_u are their corresponding true population parameters. The population moments of u are estimated within the $(1 - \alpha) \cdot 100\%$ confidence intervals as a function of the sample moments of u. Specifically, the confidence intervals (CI) for b_{xy} and $ubRMSD_{xy}$ can be inferred from Eqs. (12) and (13) as:

$$CI_{b_{xy}} = \left[b_{xy} + t_{n-1}^{\alpha/2} \frac{ubRMSD_{xy}}{\sqrt{n}} \;,\; b_{xy} + t_{n-1}^{1-\alpha/2} \frac{ubRMSD_{xy}}{\sqrt{n}} \right] \tag{14}$$

and

$$CI_{ubRMSD_{xy}} = \left[ubRMSD_{xy} \, \frac{\sqrt{n-1}}{\chi_{n-1}^{1-\alpha/2}} \;,\; ubRMSD_{xy} \, \frac{\sqrt{n-1}}{\chi_{n-1}^{\alpha/2}} \right] \tag{15}$$

No such simple direct relationships between the sampled and true values have yet been found for the other validation metrics presented here. For the Pearson correlation coefficient, it can be indirectly obtained through Fischer's z-transformation, which transforms R_{xy} into a variable that approximately follows a normal distribution with mean z_{xy} and standard deviation $(n-3)^{-0.5}$ (*Bonett and Wright*, 2000):

$$z_{xy} = 0.5 \ln \left(\frac{1 + R_{xy}}{1 - R_{xy}} \right) \sim \mathcal{N}_{z_{xy},(n-3)^{-0.5}} \tag{16}$$

The confidence interval for R_{xy} can be obtained by back-transforming z as:

$$CI_{R_{xy}} = \left[\frac{e^{2z^{1-\alpha}} - 1}{e^{2z^{1-\alpha}} + 1}, \frac{e^{2z^{\alpha}} - 1}{e^{2z^{\alpha}} + 1} \right] \tag{17}$$

One major issue for calculating confidence intervals from the analytical expressions described above is the inherent assumption of independence between samples. For soil moisture time series, this assumption is often not met due to the auto-correlated nature of soil moisture governing processes. Since such auto-correlation in the data essentially causes a widening of the confidence intervals, one popular way to account for it is to reduce the degrees of freedom (sample size) of the used distribution. This is typically done by assuming a first-order auto-regressive AR(1) behaviour in the time series and using the lag-1 auto-correlation (ρ) to calculate a correction factor for the sample size n (*Dawdy and Matalas*, 1964; *Draper et al.*, 2012):

$$n_e = n \cdot \frac{1 - \rho}{1 + \rho} \tag{18}$$

where n_e is the effective sample size that is used to estimate auto-correlation corrected confidence intervals according to Eqs. (14)-(17). A combined effective value for ρ, which summarizes the possibly different lag-1 auto-correlation of the two considered time series for which the respective validation metric is calculated, can be obtained as their geometric average:

$$\rho = \sqrt{\rho_x \cdot \rho_y} \tag{19}$$

with ρ_x and ρ_y obtained from a fitted AR(1) model as:

$$\rho_i = e^{-\frac{d_m}{\tau_i}} \tag{20}$$

where $i \in [x, y]$, τ_i is the fitted persistence time of the individual time series x and y, and d_m is the the median distance between consecutive valid, collocated observations, i.e. the lag-1 distance accounting for the typically irregular spacing between satellite measurements. Note that averaging correlation coefficients is generally not recommended (see Sec. 4.7), but required here to determine a single effective proxy of the auto-correlation of collocated data pairs with possibly deviating individual memory. Using the geometric average avoids the dominance of data sets with large auto-correlation (e.g., land surface models often have a different memory than satellite observations), which may cause excessively large confidence intervals.

Note that the necessity of relying on a possibly crude approximation of a lumped effective auto-correlation correction parameter for calculating confidence intervals is but one factor undermining their ability to serve as decision basis for declaring results as significant or non-significant (see the previous section). One should always bear in mind that confidence intervals inevitably are - just as the estimates they are meant to describe - uncertain.

4.6.2 Bootstrapping

No exact solvable analytical expressions or transformations for confidence intervals around TCA-based metrics have yet been derived. *Zwieback et al.* (2012) presented a formulation of confidence intervals for TCA-based RMSE estimates in a synthetic study which, however, required the knowledge of the true RMSE states and is therefore of limited practical use. Alternatively, several studies (e.g., *Caires and Sterl*, 2003; *Zwieback et al.*, 2012; *Draper et al.*, 2013) have suggested the use of bootstrapping as a potential non-parametric method for obtaining confidence intervals of estimators with unknown sampling distribution (*Efron and Tibshirani*, 1986).

Bootstrapping is a special case of Monte Carlo simulation, which uses the sample itself as approximation of the population. More specifically, it constructs an empirical probability distribution of the test statistic (in our case the validation metric) by resampling the original sample multiple times, with replacement to preserve the sample size, and repeated calculation of the test statistic from those resamples. This bootstrapped distribution then allows for the direct derivation of confidence intervals as well as other parameters of the sampling error pdf. The advantages of this method lie in its algorithmic simplicity and that it can be applied to any metric without the need to assume a particular sampling distribution (such as t or χ). However, bootstrapping confidence intervals requires a considerable number of resamples, which may lead to large computational costs, and relies on the assumption that the sample is indeed a reliable representation of the population, which requires a large sample size. A general recommendation for bootstrapping confidence intervals is to use a minimum of 1000 resamples (*Efron and Tibshirani*, 1986). However, the number of required resamples may be chosen more specifically for a given study by testing for convergence of the results with increasing sample size. For example, *Draper et al.* (2013) used 1000 resamples for estimating confidence intervals for TCA-based $ubRMSE$ estimates, although their testing found that 500 would have been sufficient.

As was the case for the analytical expressions, bootstrapped confidence intervals are also susceptible to auto-correlation in the data. This can be accounted for by resampling blocks of data instead of single data points, referred to as block-bootstrapping (*Ólafsdóttir and Mudelsee*, 2014), which preserves the auto-correlation properties of the original sample. An estimate of the optimal block length (l_{opt}) for bootstrapping CIs around TCA-based estimates can be obtained following *Chen et al.* (2018) as:

$$l_{opt} = \mathrm{NINT}\left\{ \sqrt[3]{\left(\frac{\sqrt{6} \cdot n \cdot \rho}{1-\rho^2} \right)^2} \right\} \tag{21}$$

where NINT$\{\cdot\}$ denotes rounding to the nearest integer. As before, a single effective value for ρ can be obtained as the geometric average of the lag-1 auto-correlations of the three data sets used to obtain the respective TCA estimate ($\rho = \sqrt[3]{\rho_x \cdot \rho_y \cdot \rho_z}$). The lag-1 is the median time interval between consecutive valid, collocated data triplets. To prevent data gaps from causing an auto-correlation degradation during the resampling, we recommend to discard data blocks from the resamples if they contain less than 50% of valid data.

4.7 Summary statistics

Validation metrics and their confidence intervals should be calculated and assessed over a wide range of spatial locations to understand error characteristics of a soil moisture product under different climatic, topographic and land cover conditions. However, it may be practical to summarize spatially distributed skill estimates into a single combined metric (for example to obtain an overall ranking of different products or to track the performance evolution of a product over time), which requires also the aggregation of their associated confidence intervals.

4.7.1 Averaging metrics

The most common way of obtaining a combined skill estimate is arithmetic averaging:

$$\overline{\nu} = \mathbf{w}^\intercal \mathbf{v} \tag{22}$$

where $\overline{\nu}$ is the average of k spatially distributed skill metrics that are summarized in the skill vector $\mathbf{v} = [\nu_1 \cdots \nu_k]^\intercal$; and $\mathbf{w} = [w_1 \cdots w_k]^\intercal$ contains the weights that are attributed to the individual skill estimates with $\sum w_i = 1$. Averaging skill metrics in a weighted fashion to minimize the impact of sampling errors is in principle possible by deriving weights from the sampling error magnitudes (*Aitkin*, 1936), but in most cases, an unweighted average is preferred because validation points are typically selected to represent a wide range of varying conditions, and areas with lower sampling errors (i.e. regions with better temporal coverage, for instance because less data are masked out) could dominate a weighted averaged skill estimate. For such unweighted average, the weight vector takes the form $\mathbf{w} = [k^{-1} \cdots k^{-1}]^\intercal$.

While many metrics can be averaged safely, it is - against common practice - not recommended to average correlation coefficients (neither Pearson nor TCA-based) because they are calculated as ratios using standard deviations (variances) and covariances or SNRs (see Eqs. (9) and (10)). Therefore, they behave highly non-linearly and neither an average of these ratios nor a ratio of averaged nominators / denominators would allow for a meaningful inference about statistical properties. For example, averaging correlation coefficients of 0.1 and 0.9, which correspond to a SNR of 0.01 and 4.26, respectively (in the case of Pearson correlation assuming a random error-free reference data set), would lead to an average correlation of 0.5 with an associated SNR of 0.33. This is far from their average SNR of 2.14 (ignoring for the moment that this too is an average of ratios) which would correspond to a correlation coefficient of 0.83. In contrast, correlation coefficients of 0.3 and 0.7, representing SNRs of 0.1 and 0.96, respectively, would have the same average correlation yet the average of their associated SNR is 0.53, corresponding to a correlation of 0.59. Moreover, the skewed probability distribution of the Pearson correlation coefficient causes the arithmetic average to be systematically biased. Some studies suggest to average Fisher-transformed z-values instead (*Corey et al.*, 1998), which have a Gaussian sampling distribution, but a back-transformed z-average is just as difficult to interpret. Following the above example, averaging correlation coefficients of 0.1 and 0.9 in z-space would lead to an average correlation (or more precisely, an inverse average-z) of 0.66 (SNR =

0.76), whereas when averaging z-transformed correlations of 0.3 and 0.7, it would be 0.53 (SNR = 0.39).

In other words, the choice of whether to average correlation coefficients, Fisher-transformed z-values, or SNRs - albeit representing the exact same uncertainty properties - will lead to different values / interpretations of the resulting average and this difference also depends on the degree of variability across the estimates that are being averaged. Moreover, the resulting average number (regardless of the approach) no longer represents an actually meaningful statistical property. Alternatively, instead of averaging pre-calculated correlation coefficients, one may be tempted to calculate the correlation coefficient directly over the concatenated measurements of all available locations to obtain an overall skill estimate. However, this is not meaningful as the effects of different populations are lumped together. As a consequence, for example, two data sets that individually exhibit strong positive correlation in a wet and in a dry soil moisture regime, respectively, may appear to have an overall weak anti-correlation when put together, an effect also known as Simpson's paradox (*Blyth*, 1972). Therefore, such an approach should be strictly avoided.

4.7.2 Averaging confidence intervals

The uncertainty in the spatially averaged skill metric in Eq. (22) associated with the *sampling* errors of the individual skill estimates can be calculated through the standard method for the propagation of uncertainty as:

$$s_{\overline{\nu}}^2 = \mathbf{w}^\intercal \mathbf{\Sigma} \mathbf{w} \tag{23}$$

where $s_{\overline{\nu}}^2$ is the sampling uncertainty in the averaged skill $\overline{\nu}$ (i.e. its sampling error variance); and $\mathbf{\Sigma}$ is the sampling error covariance matrix for the k individual skill estimates. The corresponding aggregated confidence intervals can be derived from a Gaussian distribution (which will generally be assured by the Central Limit Theorem for reasonably large samples) with mean $\overline{\nu}$ and standard deviation $s_{\overline{\nu}}$.

Diagonal elements in $\mathbf{\Sigma}$ are the sampling error variances of the individual skill estimates, i.e. $diag(\mathbf{\Sigma}) = \mathbf{s^2}$ with $\mathbf{s^2} = [s_{\nu_1}^2 \cdots s_{\nu_k}^2]^\intercal$. For b_{xy} and $ubRMSE_{xy}$ estimates, they are the squared standard errors of the sample mean and sample variance (of the difference series $u = x - y$ at each individual location), respectively:

$$
\begin{aligned}
s^2_{b_{xy}} &= \frac{ubRMSD^2_{xy}}{n} \\
s^2_{ubRMSD_{xy}} &= \frac{ubRMSD^2_{xy}}{2(n-1)}
\end{aligned}
\tag{24}
$$

For TCA-based metrics, the sampling error variance can be directly calculated from the bootstrapped sampling distribution.

In an ideal case, the reference data used for calculating skill metrics span a wide range of varying conditions with samples that are independent of each other in time and space. In this case, off-diagonal elements in $\mathbf{\Sigma}$ would be zero. However, in many cases, differences between soil moisture time series contain sample auto-correlation due to the large-scale auto-correlated nature of soil moisture (*Vachaud et al.*, 1985; *Crow et al.*, 2012) and because estimation errors (those of satellite retrievals as well as those of in situ measurements or of model predictions) may be correlated over large distances (*Gruber et al.*, 2015, 2018). Ignoring such sampling error auto-correlation can lead to considerably underestimated confidence intervals of spatially averaged skill estimates. Hence, calculating off-diagonal elements in $\mathbf{\Sigma}$, which represent sampling error covariances between the skill estimates of different locations, is critical. Although these covariances cannot be estimated directly, they can be derived from the sample auto-correlation matrix $\mathbf{R}$ and the sampling error *standard deviations* $\mathbf{s}$ (see above) as:

$$
\mathbf{\Sigma} = \mathbf{R} \circ \mathbf{s}\mathbf{s}^{\mathsf{T}} \tag{25}
$$

where $\circ$ denotes the Hadamard product, i.e. element-wise matrix multiplication. $\mathbf{R}$ differs for the various skill metrics. For b_{xy} and $ubRMSD_{xy}$, it is the *spatial* auto-correlation matrix of the difference series u, and of the squared, bias-corrected difference series $(u - \overline{u})^2$, respectively, at the different locations u where skill metrics are calculated. For TCA-based metrics, the sampling error covariance can be calculated as the covariance between the bootstrapped samples (*Gruber et al.*, 2019b), provided that the order in which bootstrap-resamples are drawn is the same at all different locations, which may be difficult when using block-bootstraps with different block-length.

Earlier research (*De Lannoy and Reichle*, 2016) has proposed a clustering approach to take possible sampling error correlations into account. This approach first calculates mean metrics and confidence intervals per spatial cluster, assuming that the sampling errors of the spatially close data sets within each cluster are perfectly correlated. Next, averaged skill metrics and confidence intervals from within the clusters are averaged, assuming that all clusters are completely independent. However, this approach is expected to overestimate confidence intervals because: (i) sampling errors will never be perfectly correlated unless validation metrics are calculated multiple times from the exact same data, and (ii) clusters are formed based on the expected auto-correlation length of the soil moisture data sets, which will be much larger than that of the difference series between data sets, as required in Eq. (25).

Finally, although averaging of some metrics and confidence intervals is possible, we generally recommend to retain detailed information about their spatial variability, and to leverage this information to obtain a better understanding of product performance and its relation to land cover, topography, climate, and other possibly important factors. If point-wise assessments are not feasible or if simple product summaries are desired, percentile statistics such as medians and inter-quartile-ranges (of both calculated skill estimates and their confidence intervals) are generally more informative than spatial averages and their increasingly inaccurate averaged confidence intervals. More specific recommendations of how validation metrics and confidence intervals can be presented are provided in Sec. 5 and Appendix A.

4.8 Practical remarks

4.8.1 Validating downscaled products

Currently, most space-borne microwave sensors available for soil moisture retrieval operate at spatial resolutions of about 25^2 - 50^2 km^2 (*Gruber et al.*, 2019a). Some higher-resolution Synthetic Aperture Radar (SAR) sensors exist that allow for reasonable soil moisture retrieval at scales up to approximately 1 km^2 (*Pathe et al.*, 2009; *Gruber et al.*, 2013b), yet with considerably lower temporal resolution and accuracy. In addition, many downscaling approaches have been developed to improve the spatial resolution of coarse-resolution soil moisture products, e.g., by fusing coarse-resolution radiometer or scatterometer measurements with high-resolution SAR data (*Das et al.*, 2017; *Bauer-Marschallinger et al.*, 2018), by fusing microwave observations with optical/thermal measurements (*Chauhan et al.*, 2003), or through data assimilation (*Reichle et al.*, 2017c). For a comprehensive review of downscaling methods see *Peng et al.* (2017).

The validation of downscaled products is mostly done as for coarse-resolution products, i.e. through time series analysis with a focus on temporal dynamics at individual locations (see

Sec. 4). In doing so, it has been shown that the downscaling process often actually decreases the temporal performance of the products, that is, the original coarse-resolution products often correlate better with local soil moisture dynamics, even at a point scale, than their downscaled counterparts (*Peng et al.*, 2015). While downscaled soil moisture images provide more visual level-of-detail, only few studies have quantitatively assessed whether the obtained spatial patterns actually represent real soil moisture variations (e.g., *Bauer-Marschallinger et al.*, 2018; *Sabaghy et al.*, in review) or whether they are just mimicking spatial patterns of ancillary data such as soil texture maps (for a comprehensive review of validation studies for downscaled products see *Peng et al.*, 2017).

Therefore, we highly recommend that future validation studies for downscaled products put a strong emphasis on assessing also the spatial soil moisture variations obtained from the downscaling, e.g., by estimating spatial correlation coefficients (*Sahoo et al.*, 2013; *Kolassa et al.*, 2017; *Sabaghy et al.*, in review), in addition to time series analyses. To that end, we further encourage the setup of field campaigns / validation sites dedicated to support such high-resolution validation activities, especially in regions where soil moisture variations are very heterogeneous.

4.8.2 Target accuracy requirements

Satellite soil moisture validation studies most commonly evaluate products against a target accuracy threshold of 0.04 m^3m^{-3} ubRMSD across the globe, excluding regions of snow and ice, frozen ground, complex topography, open water, urban areas, and vegetation with water content greater than 5 kg/m^2. This requirement was defined by the Soil Moisture and Ocean Salinity (SMOS; *Kerr et al.*, 2001) and the Soil Moisture Active Passive (SMAP; *Entekhabi et al.*, 2010a) missions, and by the Terrestrial Observation Panel for Climate (TOPC; *WMO*, 2016). Alternatively, the Satellite Application Facility in Support to Operational Hydrology and Water Management (H SAF) of the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) has defined (TCA-based) SNR product requirements (*H-SAF*, 2017) for the operational soil moisture products that are retrieved from measurements of the Advanced Scatterometer (ASCAT) onboard the MetOp satellites (*Naeimi et al.*, 2009). In particular, the EUMETSAT H SAF defines 0, 3 and 6 dB SNR as threshold, target and optimal SNR requirements to make product assessment possible on a larger scale and spatially better comparable (see Sec. 4.4).

Both of these requirements are based on relatively practical, easy-to-estimate single numbers that represent a rough average of what is currently achievable rather than being an indication of "good" or "bad" product quality. While they provide easy means to monitor product performance evolution over time and to compare products, they are not necessarily related to the suitability of a product for specific applications. However, the actual specification of bias and uncertainty requirements for the fitness-for-purpose for a particular application (including the specification of the appropriate metrics) is a task of the respective user community and needs further research (*Entekhabi et al.*, 2010b).

4.8.3 Reproducibility

The research community generally suffers from a lack of reproducibility in scientific studies (*Baker*, 2016). Also in soil moisture validation studies, contradictory results for the performance and relative ranking between different satellite products have been reported (e.g., *Wagner et al.*, 2014). These ambiguities originate from: (i) the choice of reference data and product versions; (ii) the use of different spatial regions and time periods; (iii) different approaches used for data preparation and pre-processing; (iv) statistical sampling errors; and (v) software implementation errors. Note, however, that contradicting results are not necessarily caused by bad study design but often originate from stochastic uncertainties, which are inevitably dominant in space borne Earth observation measurements (*Greenland et al.*, 2016).

Embracing statistical uncertainty and developing an in-depth understanding of soil moisture product quality requires more comprehensive descriptions of data sets, software, and methodology than are usually provided as well as the mandatory, additional estimation and presentation of sampling errors. To that end, we recommend that:

- all validation results should be accompanied by confidence intervals as measure for sampling errors;
- all methodological steps should be described with sufficient detail to be reproducible;
- all data sets used for the study should be made publicly available and unambiguously identifiable by providing their exact product version information and, where available, their Digital Object Identifier (DOI);
- all used software packages that are relevant for the exact reproduction of validation results should be referenced with their complete version number and, where available, their

DOI. If not accessible via open repositories (in particular software specifically designed for that study), we recommend to make source code publicly available, preferably on GitHub (`https://github.com/`; last access: 1 July 2019).

A list of some current publicly available software that is specifically aimed at, or closely related to soil moisture validation is provided in Table 3. An online validation tool that is built around these software packages and follows the good practice guidelines presented in this paper is provided by the Quality Assurance Framework for Soil Moisture (QA4SM; `https://qa4sm.eodc.eu/`; last access: 1 July 2019).

Note that the re-distribution of in situ measurements (see the third point above) may be particularly problematic as many networks do not operate for free. Requiring networks to freely distribute their data will likely decrease the number of datasets available for validation activities, which may ultimately hamper the evolution of satellite soil moisture products and downstream products derived thereof. We therefore emphasize the tremendous value of ground reference measurements and encourage the community to support, by any means possible, the development and continuation of operational Cal/Val sites.

5 Validation Good Practice Protocol

This section provides a compilation of the theoretical considerations presented above in the form of a validation good practice protocol for satellite soil moisture products, i.e. guidelines for:

- the selection of reference data;
- data pre-processing steps;
- the selection and implementation of appropriate metrics;
- the presentation of validation results.

Figure 3 illustrates the process and Appendix A provides an example that follows these recommendations. We stress that there is no one-size-fits-all approach for validating Earth observation data. Depending on the application in question, several analyses may not be necessary. Also, recommended thresholds may need to be adjusted depending on data quality requirements (e.g., more strict data masking procedures may be employed) or data availability (e.g., the allowed in situ measurement depth may be increased if only retrievals from long wavelengths in dry and sandy regions are used).

5.1 Data selection

As discussed in Sec. 3, no reference data source provides a sufficiently accurate and traceable soil moisture proxy for reliable error assessment on a global scale. A complete and comprehensive product validation therefore requires comparisons against each of the following: (i) dense networks, in particular core validation sites; (ii) sparse networks; (iii) land surface model output; and (iv) other satellite products, always making sure that the latest or most recommended product versions are used. However, given the large number of satellite and reference products available, a complete analysis that considers all these data sources is typically beyond the capacity of a single validation study. Therefore, separate studies may be conducted for dense network validation (*Colliander et al.*, 2017), sparse network validation (*Dorigo et al.*, 2015), or coarse-resolution product inter-comparison (*Al-Yaari et al.*, 2014) and their results compiled together in a meta-analysis.

Since satellite soil moisture retrievals represent only the top few centimeters of the soil, in situ sensors and modelled soil layers used for validation should reach no deeper than 5-10 cm, which is considered as the maximum sensing depth for currently available microwave wavelengths (X-band to L-band). Information where currently publicly available reference data sets can be accessed is provided in Table 2.

5.2 Pre-processing

5.2.1 Masking

In situ measurements and satellite retrievals are typically accompanied by quality flags, which must be used to mask out all measurements that are considered unreliable. If this masking requires the decision of a threshold (for example the probability of RFI occurrence), recommendations from data providers should be followed and the employed thresholds carefully documented. Ancillary information on dynamic geophysical variables, such as snow, temperature or vegetation, should be used to screen microwave-based satellite measurements since no reliable soil moisture retrieval is possible under frozen or snow-covered conditions and the quality of soil moisture retrievals depends on the vegetation density. Such ancillary data can be supplied by land surface models or complementary satellite data. Specifically, we recommend masking out pixels classified as tropical forests, water bodies, wetlands, and inundation areas as well as all measurements on days with non-zero snow indicators (e.g., snow height or snow-water-equivalent), or surface or soil temperature below 4°C. When biases or uncertainties of multiple products are compared, they should be calculated from the exact same, collocated data points. However, care should be taken that single products with poor data coverage do not distort the overall assessment (see Sec. 6).

To avoid excessively large confidence intervals that can hamper meaningful data comparison, grid cells with less than 50-500 collocated data points may be masked out depending on data availability (*Zwieback et al.*, 2012). Also, many studies mask out correlation coefficients based on Student's t-test (i.e. applying p-value thresholds for correlation coefficients), and/or bias and uncertainty estimates based on vegetation density (e.g., vegetation water content > 5 kg/m^2) or other thresholds (e.g., open-water fraction > 0.05) (*Dorigo et al.*, 2010; *Brocca et al.*, 2011; *Al-Yaari et al.*, 2014). However, carefully reporting and interpreting confidence intervals and sample sizes at locations with low data coverage could indeed provide valuable additional insight and may be more informative than masking out estimates completely (*Wasserstein et al.*, 2019). Also, complete reporting of results prevents generating publication biases due to "cherry-picking" which is sometimes found in the scientific literature (*Greenland et al.*, 2016).

5.2.2 Collocation

Spatial collocation requires the selection of a spatial comparison grid, which is often the grid of the satellite product under validation. In situ measurements should be assigned to the grid cell in which they are located. For dense networks, all stations that lie within a particular grid cell should be averaged, if possible taking their respective spatial representativeness for that grid cell into account. To avoid artificial jumps due to sensor drop-outs, only time steps where all stations provide valid measurements should be considered. For the SMAP core validation sites (see Sec. 3.2.1), a validation grid that minimizes upscaling errors has been developed as described in *Colliander et al.* (2017).

Gridded reference products (i.e. other satellite and land surface model products) should be resampled onto the chosen comparison grid, e.g., using a Nearest Neighbor (NN) search. If the grid resolution of the reference product is coarser than that of the comparison grid, individual grid cells of that product may be assigned to multiple comparison grid cells. If the grid resolution is much finer, all NNs of single comparison grid cells (in case more than one exist) should be averaged, if possible taking spatial representativeness into account.

Temporal collocation at comparison time steps should minimize the time difference between data match-ups and be based on a NN-search with a maximum time difference threshold of 1-12 hours, depending on data availability. Note that the choice of the comparison grid and time steps may affect the presence and distribution of (spatial and temporal) representativeness errors among the considered data sets (see Sec. 6).

5.2.3 Decomposition

All validation metrics should be calculated for the raw soil moisture time series (of collocated retrievals and reference data) as well as for short-term and long-term anomalies, except for temporal mean biases whose calculation is trivial for anomalies. Short-term anomalies should be estimated as residuals from a seasonality that is computed by applying a 4-8 week moving average window to the time series. Long-term anomalies should be estimated as residuals from a climatology that is computed by averaging the measurements of all years within a 4-8 week moving window around each DOY, but only if at least 5-10 years of data are available. To avoid data-density related artefacts, especially in the transition periods from frozen to non-frozen periods, moving averages should only be calculated if at least 25-50% of the maximal data pair coverage is available within a particular time window.

5.2.4 Rescaling

When using fiducial reference data, units (e.g., m^3m^{-3} and degree of saturation) should be unified for the purpose of bias estimation using soil texture information, keeping in mind that inaccuracy in soil information directly propagates into the bias estimates. To account for (horizontal and vertical) systematic representativeness errors and different soil moisture units, the data set under validation should be rescaled (before decomposition for validating raw time series and after decomposition for validating anomalies) towards the reference data when estimating absolute uncertainties (i.e. ubRMSDs or ubRMSEs). When calculating relative metrics, data sets should be rescaled by matching their temporal mean and standard deviation. When calculating TCA-based metrics, data sets should be rescaled using also TCA-based rescaling coefficients. Note that no rescaling or unit conversion is necessary for Pearson correlation coefficients or TCA-based correlation and SNR estimates, since these metrics are not affected by linear data transformation.

5.3 Metric calculation

Remember that all covariance-based metrics require zero error correlation. Any combination of in situ measurements, land surface model estimates, active-microwave-based measurements, or passive-microwave-based measurements is expected to mostly fulfil this requirement (see Sec. 4.4.2; *Gruber et al.*, 2016a). Different products from within any of these categories (except for in situ data), on the other hand, are expected to have correlated errors (*Gruber et al.*, 2016b). Therefore, the metrics described below should not be applied to such product combinations. Moreover, since non-zero error correlations may exist even when using products from different categories (see Sec. 4.4.2; *Yilmaz and Crow*, 2014; *Pan et al.*, 2015), it is strongly recommended to verify if assumptions are met (see Sec. 5.3.2).

5.3.1 Relative metrics

Temporal mean biases (Eq. (4)) should be calculated between all data sets that are expected to be properly collocated and have comparable spatial resolution, and are hence not dominated by spatial representativeness errors. These data sets may include dense networks, land surface models, and other satellite data sets. It should be kept in mind, however, that the underlying measurement resolution often considerably differs from the sampling grid resolution, which potentially causes representativeness errors that are not directly apparent as such. Correlation coefficients and unbiased Root-Mean-Square-Differences (Eqs. (9) and (7), respectively) should be calculated between all data sets whose errors are not expected to be correlated (see above).

5.3.2 TCA-based metrics

Second-order biases (Eq. (5)) of the validation data set should be calculated using fiducial reference data (i.e. at the core validation sites). Unbiased Root-Mean-Square-Errors and SNRs (Eqs. (8) and (11), respectively) should be calculated for all data sets. If more than one triplet with independent errors is available to estimate the bias or uncertainty of a particular product, TCA should be applied to all possible triplets and redundant estimates should be averaged (*Gruber et al.*, 2016b). The spread between redundant estimates should be used as a diagnostic to verify if orthogonality and zero error correlation assumptions are met (*Dorigo et al.*, 2010; *Draper et al.*, 2013; *Chen et al.*, 2017).

5.3.3 Confidence intervals

For each metric, 80-95% confidence intervals should be calculated using their analytical estimators (Eqs. (14)-(17)) or, if not available, block-bootstrapping. The latter should be based on at least 1000 bootstrap samples (*Efron and Tibshirani*, 1986) or possibly less if tested for convergence, and all confidence intervals should be corrected for sample auto-correlation.

5.4 Presentation

Validation metrics together with sample size and upper and lower confidence intervals/limits should be presented for each location where they are calculated, either by means of spatial maps or, if not meaningful (for example for core validation sites), in tabular form. Additionally, summary statistics (representing average conditions and spatial variability) of both validation metrics and their confidence intervals/limits should be provided, e.g., in the form of boxplots (i.e. median, inter-quartile-range and 5th/95th percentiles). The presentation can be further customized, for example by stratifying the summary statistics for climatological or land surface conditions.

Ratio-based metrics (i.e. Pearson and TCA-based correlation coefficients as well as SNRs) must not be averaged. Differences between these metrics must always be related to their absolute values and be interpreted with care (see Sec. 4.7). SNR-related properties of different products may be compared in terms of SNR ratios or SNR differences in decibel space (Eq. (11)).

Examples of how validation metrics and associated confidence intervals can be presented are provided in Appendix A.

6 Towards best practices: discussion and conclusions

In this paper we have reviewed state-of-the-art validation methods, including reference data sources and data pre-processing procedures, and provided community-agreed good practice guidelines for the validation of satellite soil moisture products. Moreover, we have identified several weak links that require careful attention to increase the reliability of soil moisture data quality assessments. Specifically, the following research gaps should be addressed in the near future:

- On assumptions: the majority of studies assume that estimated biases and uncertainties are stationary (i.e. constant over time) or at least that they represent the average data quality of a product. However, given the strong link between soil moisture data quality and vegetation (*van der Schalie et al.*, 2018; *Zwieback et al.*, 2018; *Gruber et al.*, 2019a), retrieval accuracy can be expected to vary strongly between seasons and many applications could greatly benefit from temporally varying quality information. Given the rapidly growing temporal coverage of soil moisture products, efforts should be made to provide bias and uncertainty estimates at different time scales, which also requires the use of seasonally varying bias correction (i.e. rescaling) parameters.

- On pre-processing: very little is known about how spatial and temporal collocation mismatches contribute to bias and uncertainty estimates. Using simple NN or IDW approaches to find match-ups between measurements that sample very different soil volumes or were taken at different times will give rise to representativeness errors that may considerably affect the overall picture of the quality of a product. More research is needed to quantify these representativeness errors and to develop resampling methods that more rigorously take actual measurement resolution into account.

- On metric calculation: most current studies neglect the impact of second-order biases on various validation metrics such as the temporal mean difference or the ubRMSD. Several attempts are made to mitigate their impact using rescaling methods that match the statistical moments of the data sets, yet most of these methods do not account for random errors and therefore match the moments in an insufficient manner. More research is needed to quantify the impact of suboptimal rescaling on second-order biases, on the impact of uncorrected second-order biases on validation metrics, and on how such uncorrected biases can be accounted for.

- On reference data: validation targets are typically defined against an unknown truth. Comparing metrics against error-prone estimates of this truth (i.e. reference data) will be inflated by some unknown amount. Efforts should be made to obtain proper bias and uncertainty estimates for reference data sets, which should be further used to correct over- or underestimated validation metrics (*Miralles et al.*, 2010; *Chen et al.*, 2017).

- On statistical uncertainty: most validation studies do not report confidence intervals, even though they are critical for a reliable interpretation of validation results. Although an accurate analytical calculation of confidence intervals for large-scale validation is not trivial for all metrics, bootstrapping provides an easy and robust alternative. However, care must be taken to properly account for spatial and temporal auto-correlation in the data.

- On continuity: given the perpetual changes in the land surface character and climate as well as progressively increasing data record lengths, sensor drifts, changing reference data availability, and improving soil moisture retrieval algorithms, validation should be a continuous process and validation reports frequently (at least annually) updated throughout and beyond the lifetime of the various satellite missions.

- On accuracy requirements: the well-known soil moisture mission target accuracy requirement of 0.04 m^3m^{-3} (as specified by the Global Climate Observing System as well as for individual products and missions), against which soil moisture products are typically evaluated, does not relate to the fitness-for-purpose for a specific application. We therefore strongly encourage a closer collaboration between satellite data providers and the soil moisture user community to determine application specific accuracy requirements that provide deeper insight into what constitutes "good" or "bad" soil moisture data quality, thereby fostering the development of improved satellite products.

Finally, many of the discussed principles and methods are not exclusively restricted to soil moisture. By setting this example, we hope to also nurture the development and evolution of validation good practice guidelines in other Earth observation communities.

7 Acknowledgements

We acknowledge the support from the International Space Science Institute (ISSI; `http://www.issibern.ch/`; last access: 1 July 2019). This publication is an outcome of the ISSI's Team on "Adding value to soil moisture information for climate studies" and has received funding from the eartH2Observe project (European Union's Seventh Framework Programme, Grant Agreement No. 603608), from the KU Leuven C1 internal fund C14/16/045 and from the Research Foundation Flanders (FWO-1530019N).

Appendix

A Validation example

Sec. 5 compiles community-agreed validation good practices into a recommended validation protocol. In this appendix, we provide an example that follows this protocol, not to actually assess the quality of certain products, but to show an illustrative scenario that can be easily extrapolated to more specific validation tasks that readers may face. This includes a comprehensive description of the validation setup, demonstrative examples of how validation results may be presented, and a discussion on where the currently available satellite soil moisture validation literature often fails to comply with the good practice recommendations presented here. Results shown in this appendix have been generated using the python programming language. All source code is available at `https://github.com/alexgruber/validation_good_practice/` (last access: 1 July 2019). Metric calculation routines have been additionally translated into MATLAB.

A.1 Data sets and study area

Select validation examples are shown for soil moisture retrievals from the Advanced SCATterometer (ASCAT; *Naeimi et al.*, 2009), the Soil Moisture and Ocean Salinity (SMOS) mission (*Kerr et al.*, 2010), and the Soil Moisture Active Passive (SMAP) mission (*Entekhabi et al.*, 2010a). Reference data used are coarse-resolution model estimates from the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2; *Gelaro et al.*, 2017). This analysis is performed over the Contiguous United States (CONUS) using data from the beginning of 2015 through the end of 2018.

ASCAT data used are the EUMETSAT H SAF H113 data record and its extension H114, which are Level 2 (L2) soil moisture products that have been retrieved from inter-calibrated backscatter measurements from identical ASCAT instruments onboard the MetOp-A and MetOp-B satellites using the TU Wien WAter Retrieval Package (WARP) algorithm (*Wagner et al.*, 1999; *Naeimi et al.*, 2009). ASCAT is an active C-band radar with a spatial resolution of 25 km. Soil moisture is retrieved as the degree of saturation and sampled onto a 12.5 km discrete global grid. Data can be obtained upon registration from `http://hsaf.meteoam.it/soil-moisture.php` (last access: 1 July 2019).

SMOS data are the reprocessed L2 soil moisture retrievals version V650, which can be obtained upon registration from `https://smos-diss.eo.esa.int/` (last access: 1 July 2019; *Kerr* *et al.*, 2012). SMOS is a passive L-band interferometric radiometer with an average spatial resolution of 43 km. Soil moisture is retrieved in volumetric units and sampled on a 15 km discrete global grid.

SMAP data used are the 36 km L2 radiometer-only soil moisture retrievals (SPL2SMP), algorithm version 5 (R16010) (*O'Neill et al.*, 2018, DOI: 10.5067/SODMLCE6LGLL). The passive SMAP radiometer operates at L-band at a spatial resolution of 40 km. Soil moisture is retrieved in volumetric units and sampled on the 36 km EASE grid version 2 (*Brodzik et al.*, 2012).

MERRA-2 (*Gelaro et al.*, 2017) is the latest atmospheric reanalysis produced by NASA's Global Modelling and Assimilation Office. Soil moisture is estimated on a 0.5° × 0.625° grid in volumetric units as internal state variable of its land surface component, the Catchment Land Surface Model (*Koster et al.*, 2000). Here we use soil moisture estimates of the surface layer, which refers to the top 5 cm of the soil (*GMAO*, 2015). MERRA-2 data can be downloaded from `https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/data_access/` (last access: 1 July 2019).

A.2 Pre-processing

Unreliable soil moisture retrievals of the individual satellite products are masked out following the recommendations of the data providers. ASCAT soil moisture retrievals are masked out if the correction flag has a value other than 0 or 4, if the confidence flag and the processing flag have values other than 0, or if the surface state flag (*Naeimi et al.*, 2012) has a value other than 1. SMOS retrievals are masked out if the RFI probability exceeds 0.1 or if the Chi-2 probability drops below 0.05. SMAP data are masked out if the retrieval quality flag has a value other than 0 or 8. In addition, soil moisture retrievals of all satellite products are masked out at time steps where MERRA-2 estimates a soil temperature below 4°C or non-zero snow mass.

ASCAT, SMOS and MERRA-2 are resampled to the 36 km EASE v2 grid that is used for SMAP retrievals using a nearest-neighbor approach. Note that ASCAT data is, although sampled on a 12.5 km grid, not aggregated as the actual measurement resolution (25 km) is already close to the EASE v2 grid resolution. Data sets are collocated in time by resampling them to fixed reference time steps with 24 hour intervals using a nearest-neighbor search. Reference time steps are selected for each grid cell separately such that they maximize the number of collocated time steps where all data sets provide valid soil moisture estimates. Note that the choice of this reference time step can increase or decrease the sample size - depending on the spatial location of the grid cell - by up to a factor of two.

After spatial and temporal collocation, short-term anomalies are calculated for each data set using a 35-day moving average window. Long-term anomalies are not considered here because the study period of four years (2015-2018) is too short to calculate reliable long-term climatologies. The term "raw time series" is used to refer to the non-decomposed data, i.e. before anomalies have been calculated. For the estimation of unbiased RMSDs, data sets (both raw and anomaly time series) are rescaled by matching their temporal mean and standard deviation using MERRA-2 as scaling reference for comparability.

A.3 Skill metrics and presentation

A.3.1 Sample size

All metrics are calculated from the same collocated data points, i.e. days where all four data sets provide valid soil moisture estimates. The number of temporal matches at each grid cell within our study domain is shown in Figure A.1. As discussed in Sec. 4, sample size directly translates into statistical power, i.e. reliability (in terms of confidence intervals) of the calculated skill metrics. Sample sizes obtained here, which range from 150 in the more mountainous areas to up to about 300-500 in the rest of the CONUS, are typically considered high and associated with reasonably low confidence intervals for validation purposes.

However, as discussed in Sec. 4.6, confidence intervals are affected by temporal auto-correlation. "Effective" sample sizes, corrected for auto-correlation using Eq. (18), are additionally shown in Figure A.1 considering all data sets (for TCA metrics), and in Figure A.2 for raw soil moisture time series and Figure A.3 for soil moisture anomalies considering different data set pairs. Effective sample sizes are considerably smaller than actual sample sizes, especially for raw time series due to the strong auto-correlation of the seasonal soil moisture cycle. Since auto-correlation levels vary between data sets, effective sample sizes vary when calculated for different data set pairs (albeit only slightly), which in turn leads to differences in the confidence intervals of relative skill metrics that are calculated between these data pairs.

In the following, all analytical confidence intervals (Eqs. (14), (15), and (17)) are calculated using these auto-correlation corrected effective sample sizes. For bootstrapped confidence intervals, temporal auto-correlation is accounted for using block-bootstrapping (see Sec. 4.6.2) where block-lengths are estimated from the same auto-correlation levels that are underlying the calculation of effective sample sizes (see Eq. (21)).

A.3.2 Relative metrics

Figures A.4, A.5 and A.6 show spatial plots of relative (mean) bias, ubRMSD and R^2 (coefficient of determination or squared Pearson correlation) estimates for raw soil moisture values, respectively, and Figures A.7 and A.8 show ubRMSD and R^2 estimates for soil moisture anomalies, respectively.

Biases are only calculated for raw soil moisture time series and between soil moisture estimates that are expressed in the same unit, i.e. for SMOS, SMAP, and MERRA-2 which provide estimates of volumetric soil moisture. ASCAT estimates of the degree of saturation could be converted into volumetric units using porosity information, but since the quality of soil texture maps on these scales is questionable, this is not recommended for bias estimation purposes. Note also, that the biases between the remaining three data sets also include collocation and (vertical and horizontal) scale mismatches and should therefore be interpreted with care.

Along with the skill estimates, maps of confidence intervals are shown as the difference between the upper and lower confidence limits, chosen to be the 90th and the 10th percentile of the sampling distribution, respectively. Important to note is that confidence intervals for R^2 and ubRMSD estimates depend on the magnitude of the respective skill estimate, and are for R^2 not centered around the skill estimate. Misinterpretations may be avoided by directly presenting the actual confidence limits (see Sec. 4.7).

We choose a confidence level of 80% because confidence intervals at the more common (yet completely arbitrary) 95% confidence level typically become excessively large for the sample sizes available from collocated satellite products (*Gruber et al.*, 2019a), especially when taking temporal auto-correlation into account.

Figure A.9 shows spatial summary statistics of the relative skill metrics as well as of their upper and lower confidence limits. Hardly any skill differences would be considered significant when tested in the common way of checking for overlap between upper and lower confidence limits, even though Figures A.4 - A.8 show clear differences in spatial patterns.

A.3.3 Triple collocation metrics

As discussed in Sec. 4, TCA requires three data sets with independent random errors. Since errors of SMAP and SMOS are expected to be correlated (see Sec. 5.3), two independent data set triplets can be formed, i.e. ASCAT - SMOS - MERRA-2 and ASCAT - SMAP - MERRA-2. This results in unambiguous skill estimates for SMAP and SMOS, and in two skill estimates for

ASCAT, which are averaged for increased precision.

Figures A.10 and A.11 show spatial plots of TCA-based ubRMSE and R^2 (coefficient of determination w.r.t. the unknown truth) estimates, respectively, and Figures A.12 and A.13 show ubRMSE and R^2 estimates for short-term soil moisture anomalies, respectively. The skill estimates represent the median of the bootstrapped sampling distribution, which are more robust than the direct estimates, and 80 % confidence intervals (i.e. the range between the 90th and the 10th percentile of the bootstrapped sampling distribution) are provided. Spatial summary statistics of the TCA estimates (sampling distribution median) as well as of the upper and lower confidence limits are shown in Figure A.14.

The two degrees of freedom in TCA-based ASCAT skill estimates can not only be used for increasing the precision of the estimates by averaging them, but also to verify if TCA assumptions (i.e. zero error cross-correlation and error orthogonality) are met because if so, skill estimates should be identical. To this end, Figure A.15 shows the differences between R^2 and ubRMSE estimates for ASCAT when calculated once using SMOS as third data set and once using SMAP as third data set.

On average, differences are close to zero and especially R^2 estimates do not exhibit spatial patterns of notable magnitude, which suggests that differences are mainly caused by sampling errors and hence that the TCA assumptions are generally respected. Some positive skill biases for raw soil moisture estimation for ASCAT are apparent in some northern and western parts of the CONUS, with skill estimates being slightly higher when using SMOS rather than SMAP in the triplet. These areas strongly coincide with regions of generally poor ASCAT performance (see Figure A.11), which is more pronounced in the ubRMSD because SNR biases of a given magnitude are associated with larger biases in error variance at low SNR levels than at high SNR levels. (see Sec. 4.7). Poor ASCAT performance in the northern CONUS is associated with issues in the vegetation correction of the WARP retrieval algorithm (see Sec. A.1). These uncorrected vegetation signals are removed when using soil moisture anomalies, which results in a considerable increase in skill metrics (see Figure A.13) and also removes the non-zero difference in ASCAT skill estimates when using SMOS versus SMAP for TCA, i.e. spurious error cross-correlations (see Figure A.15).

A.4 Final remarks

In this appendix, we provide an illustrative validation example that follows the good practice guidelines presented in this paper. For brevity, we omit the presentation of ground data comparisons, which can be calculated and presented in the exact same way as the area-wide coarse-scale comparisons shown above. For simplicity, results are presented in spatial maps and boxplots that cover all of CONUS without further stratification. For summary information or if metrics are only computed at a few locations using ground reference data, results could be further presented in tabular format. Some examples of comprehensive ground reference data comparison including both sparse networks and core validation sites can be found in *Dorigo et al.* (2015); *Chen et al.* (2017); *Colliander et al.* (2017).

References

Aitkin, A. (1936), On least squares and linear combination of observations, *Proceedings of the Royal Society of Edinburgh*, **55**, p. 42–48, doi:10.1017/S0370164600014346.

Al-Yaari, A., J.-P. Wigneron, A. Ducharne, Y. Kerr, W. Wagner, G. De Lannoy, R. Reichle, A. Al Bitar, W. Dorigo, P. Richaume, et al. (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**, p. 614–626, doi:10.1016/j.rse.2014.07.013.

Albergel, C., C. Ruediger, T. Pellarin, J. Calvet, N. Fritz, F. Froissard, D. Suquia, A. Petitpa, B. Piguet, and E. Martin (2008), From near-surface to root-zone soil moisture using an exponential filter: an assessment of the method based on in-situ observations and model simulations., *Hydrology and earth system sciences.*, **12**(6), p. 1323–1337, doi:10.5194/hess-12-1323-2008.

Albergel, C., E. Zakharova, J.-C. Calvet, M. Zribi, M. Pardé, J.-P. Wigneron, N. Novello, Y. Kerr, A. Mialon, and N. ed Dine Fritz (2011), A first assessment of the smos data in southwestern france using in situ and airborne soil moisture estimates: The carols airborne campaign, *Remote Sensing of Environment*, **115**(10), p. 2718 – 2728, doi:10.1016/j.rse.2011.06.012.

Albergel, C., P. de Rosnay, C. Gruhier, J. Munoz-Sabater, S. Hasenauer, L. Isaksen, Y. Kerr, and

W. Wagner (2012), Evaluation of remotely sensed and modelled soil moisture products using global ground-based in situ observations, *Remote Sensing of Environment*, **118**, p. 215–226, doi:10.1016/j.rse.2011.11.017.

Albergel, C., W. Dorigo, R. Reichle, G. Balsamo, P. De Rosnay, J. Muñoz-Sabater, L. Isaksen, R. De Jeu, and W. Wagner (2013), Skill and global trend analysis of soil moisture from reanalyses and microwave remote sensing, *Journal of Hydrometeorology*, **14**(4), p. 1259–1277, doi:10.1175/JHM-D-12-0161.1.

Babaeian, E., M. Sadeghi, S. B. Jones, C. Montzka, H. Vereecken, and M. Tuller (2019), Ground, proximal, and satellite remote sensing of soil moisture, *Reviews of Geophysics*, **57**, doi:10.1029/2018RG000618.

Baker, M. (2016), 1,500 scientists lift the lid on reproducibility, *Nature News*, **533**(7604), p. 452, doi:10.1038/533452a.

Balsamo, G., C. Albergel, A. Beljaars, S. Boussetta, E. Brun, H. Cloke, D. Dee, E. Dutra, J. Muñoz-Sabater, F. Pappenberger, et al. (2015), ERA-Interim/Land: a global land surface reanalysis data set, *Hydrology and Earth System Sciences*, **19**(1), p. 389–407, doi:10.5194/hess-19-389-2015.

Bartalis, Z., R. Kidd, and K. Scipal (2006), Development and implementation of a discrete global grid system for soil moisture retrieval using the MetOp ASCAT scatterometer, in *1st EPS/MetOp RAO Workshop*, vol. ESA SP-618, ESRIN, Frascati, Italy.

Bauer-Marschallinger, B., D. Sabel, and W. Wagner (2014), Optimisation of global grids for high-resolution remote sensing data, *Computers & Geosciences*, **72**, p. 84–93, doi:10.1016/j.cageo.2014.07.005.

Bauer-Marschallinger, B., C. Paulik, S. Hochstöger, T. Mistelbauer, S. Modanesi, L. Ciabatta, C. Massari, L. Brocca, and W. Wagner (2018), Soil moisture from fusion of scatterometer and sar: Closing the scale gap with temporal filtering, *Remote Sensing*, **10**(7), p. 1030, doi:10.3390/rs10071030.

Bindlish, R., T. J. Jackson, A. J. Gasiewski, M. Klein, and E. G. Njoku (2006), Soil moisture mapping and AMSR-E validation using the PSR in SMEX02, *Remote Sensing of Environment*, **103**(2), p. 127–139, doi:10.1016/j.rse.2005.02.003.

Bindlish, R., T. Jackson, A. Gasiewski, B. Stankov, M. Klein, M. Cosh, I. Mladenova, C. Watts, E. Vivoni, V. Lakshmi, et al. (2008), Aircraft based soil moisture retrievals under mixed vegetation and topographic conditions, *Remote Sensing of Environment*, **112**(2), p. 375–390, doi:10.1016/j.rse.2007.01.024.

Blyth, C. R. (1972), On Simpson's paradox and the sure-thing principle, *Journal of the American Statistical Association*, **67**(338), p. 364–366, doi:10.1080/01621459.1972.10482387.

Bogena, H., C. Montzka, J. Huisman, A. Graf, M. Schmidt, M. Stockinger, C. von Hebel, H. Hendricks-Franssen, J. van der Kruk, W. Tappe, et al. (2018), The TERENO-Rur hydrological observatory: A multiscale multi-compartment research platform for the advancement of hydrological science, *Vadose Zone Journal*, **17**(1), doi:10.2136/vzj2018.03.0055.

Bolten, J. D., W. T. Crow, X. Zhan, T. J. Jackson, and C. A. Reynolds (2010), Evaluating the utility of remotely sensed soil moisture retrievals for operational agricultural drought monitoring, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **3**(1), p. 57–66, doi:10.1109/JSTARS.2009.2037163.

Bonett, D. G., and T. A. Wright (2000), Sample size requirements for estimating pearson, kendall and spearman correlations, *Psychometrika*, **65**(1), p. 23–28, doi:10.1007/BF02294183.

Brocca, L., F. Melone, T. Moramarco, and R. Morbidelli (2010a), Spatial-temporal variability of soil moisture and its estimation across scales, *Water Resources Research*, **46**(2), doi:10.1029/2009WR008016.

Brocca, L., F. Melone, T. Moramarco, W. Wagner, and S. Hasenauer (2010b), ASCAT soil wetness index validation through in situ and modeled soil moisture data in central italy, *Remote Sensing of Environment*, **114**(11), p. 2745–2755, doi:10.1016/j.rse.2010.06.009.

Brocca, L., S. Hasenauer, T. Lacava, F. Melone, T. Moramarco, W. Wagner, W. Dorigo, P. Matgen, J. Martinez-Fernandez, P. Llorens, J. Latron, C. Martin, and M. Bittelli (2011), Soil moisture estimation through ASCAT and AMSR-E sensors: An intercomparison and validation study across europe, *Remote Sensing of Environment*, **115**(12), p. 3390–3408, doi:10.1016/j.rse.2011.08.003.

Brocca, L., T. Tullo, F. Melone, T. Moramarco, and R. Morbidelli (2012), Catchment scale soil moisture spatial-temporal variability, *Journal of Hydrology*, **422-423**, p. 63–75, doi:10.1016/j.jhydrol.2011.12.039.

Brodzik, M. J., B. Billingsley, T. Haran, B. Raup, and M. H. Savoie (2012), EASE-Grid 2.0: Incremental but significant improvements for earth-gridded data sets, *ISPRS International Journal of Geo-Information*, **1**(1), p. 32–45, doi:10.3390/ijgi1010032.

Burgin, M. S., A. Colliander, E. G. Njoku, S. K. Chan, F. Cabot, Y. H. Kerr, R. Bindlish, T. J. Jackson, D. Entekhabi, and S. H. Yueh (2017), A comparative study of the SMAP passive soil moisture product with existing satellite-based soil moisture products, *IEEE Transactions on Geoscience and Remote Sensing*, **55**(5), p. 2959–2971, doi:10.1109/TGRS.2017.2656859.

Caires, S., and A. Sterl (2003), Validation of ocean wind and wave data using triple collocation, *Journal of Geophysical Research: Oceans*, **108**(C3), doi:10.1029/2002JC001491.

Chauhan, N. S., S. Miller, and P. Ardanuy (2003), Spaceborne soil moisture estimation at high resolution: a microwave-optical/ir synergistic approach, *International Journal of Remote Sensing*, **24**(22), p. 4599–4622, doi:10.1080/0143116031000156837.

Chen, F., W. T. Crow, A. Colliander, M. H. Cosh, T. J. Jackson, R. Bindlish, R. H. Reichle, S. K. Chan, D. D. Bosch, P. J. Starks, et al. (2017), Application of triple collocation in ground-based validation of soil moisture active/passive (SMAP) level 2 data products, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **10**(2), p. 489–502, doi:10.1109/JSTARS.2016.2569998.

Chen, F., W. T. Crow, R. Bindlish, A. Colliander, M. S. Burgin, J. Asanuma, and K. Aida (2018), Global-scale evaluation of SMAP, SMOS and ASCAT soil moisture products using triple collocation, *Remote Sensing of Environment*, **214**, p. 1–13, doi:10.1016/j.rse.2018.05.008.

Colliander, A., T. J. Jackson, R. Bindlish, S. Chan, N. Das, S. Kim, M. Cosh, R. Dunbar, L. Dang, L. Pashaian, et al. (2017), Validation of SMAP surface soil moisture products with core validation sites, *Remote sensing of environment*, **191**, p. 215–231, doi:10.1016/j.rse.2017.01.021.

Corey, D. M., W. P. Dunlap, and M. J. Burke (1998), Averaging correlations: Expected values and bias in combined pearson rs and fisher's z transformations, *The Journal of general psychology*, **125**(3), p. 245–261, doi:10.1080/00221309809595548.

Cosh, M. H., T. J. Jackson, R. Bindlish, and J. H. Prueger (2004), Watershed scale temporal and spatial stability of soil moisture and its role in validating satellite estimates, *Remote sensing of Environment*, **92**(4), p. 427–435, doi:10.1016/j.rse.2004.02.016.

Cosh, M. H., T. J. Jackson, P. Starks, and G. Heathman (2006), Temporal stability of surface soil moisture in the little washita river watershed and its applications in satellite soil moisture product validation, *Journal of Hydrology*, **323**(1–4), p. 168–177, doi:10.1016/j.jhydrol.2005.08.020.

Crow, W. T., A. A. Berg, M. H. Cosh, A. Loew, B. P. Mohanty, R. Panciera, P. de Rosnay, D. Ryu, and J. P. Walker (2012), Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products, *Rev. Geophys.*, **50**(2), p. RG2002, doi:10.1029/2011RG000372.

Cuenca, R. H., D. E. Stangel, and S. F. Kelly (1997), Soil water balance in a boreal forest, *Journal of Geophysical Research-Atmospheres*, **102**(D 24), p. 29,355–29,365, doi:10.1029/97JD02312.

Das, N. N., D. Entekhabi, S. Kim, T. Jagdhuber, S. Dunbar, S. Yueh, and A. Colliander (2017), High-resolution enhanced product based on smap active-passive approach using sentinel 1a and 1b sar data, in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, p. 2543–2545, IEEE, doi:10.1109/IGARSS.2017.8127513.

Dawdy, D., and N. Matalas (1964), *Statistical and probability analysis of hydrologic data, part III: Analysis of variance, covariance and time series*, McGraw-Hill.

De Lannoy, G. J., and R. H. Reichle (2016), Assimilation of SMOS brightness temperatures or soil moisture retrievals into a land surface model, *Hydrology and Earth System Sciences*, **20**(12), p. 4895–4911, doi:10.5194/hess-20-4895-2016.

de Nijs, A. H., R. M. Parinussa, R. A. de Jeu, J. Schellekens, and T. R. Holmes (2015), A methodology to determine radio-frequency interference in AMSR2 observations, *Geoscience and Remote Sensing, IEEE Transactions on*, **53**(9), p. 5148–5159, doi:10.1109/TGRS.2015.2417653.

De Rosnay, P., J.-C. Calvet, Y. Kerr, J.-P. Wigneron, F. Lemaître, M. J. Escorihuela, J. M. Sabater, K. Saleh, J. Barrié, G. Bouhours, et al. (2006), SMOSREX: A long term field campaign experiment for soil moisture and land surface processes remote sensing, *Remote Sensing of Environment*, **102**(3-4), p. 377–389, doi:10.1016/j.rse.2006.02.021.

Dee, D. P. (2005), Bias and data assimilation, *Quarterly Journal of the Royal Meteorological Society*, **131**(613), p. 3323–3343, doi:10.1256/qj.05.137.

Djamai, N., R. Magagi, K. Goïta, M. Hosseini, M. H. Cosh, A. Berg, and B. Toth (2015), Evaluation of SMOS soil moisture products over the CanEx-SM10 area, *Journal of hydrology*, **520**, p. 254–267, doi:10.1016/j.jhydrol.2014.11.026.

Dorigo, W., P. van Oevelen, W. Wagner, M. Drusch, S. Mecklenburg, A. Robock, and T. Jackson (2011a), A new international network for in situ soil moisture data, *Eos Transactions AGU*, **92**(17), p. 141–142, doi:10.1029/2011EO170001.

Dorigo, W., R. de Jeu, D. Chung, R. Parinussa, Y. Liu, W. Wagner, and D. Fernández-Prieto (2012), Evaluating global trends (1988–2010) in harmonized multi-satellite surface soil moisture, *Geophysical Research Letters*, **39**(18), doi:10.1029/2012GL052988.

Dorigo, W., A. Xaver, M. Vreugdenhil, A. Gruber, H. A, A. Sanchis-Dufau, D. Zamojski, C. Cordes, W. Wagner, and M. Drusch (2013), Global automated quality control of in situ soil moisture data from the international soil moisture network, *Vadose Zone Journal*, **12**(3), doi:10.2136/vzj2012.0097.

Dorigo, W., A. Gruber, R. De Jeu, W. Wagner, T. Stacke, A. Loew, C. Albergel, L. Brocca, D. Chung, R. Parinussa, et al. (2015), Evaluation of the ESA CCI soil moisture product using ground-based observations, *Remote Sensing of Environment*, **162**, p. 380–395, doi:10.1016/j.rse.2014.07.023.

Dorigo, W., W. Wagner, C. Albergel, F. Albrecht, G. Balsamo, L. Brocca, D. Chung, M. Ertl, M. Forkel, A. Gruber, et al. (2017), ESA CCI soil moisture for improved earth system understanding: state-of-the art and future directions, *Remote Sensing of Environment*, **203**, p. 185–215, doi:10.1016/j.rse.2017.07.001.

Dorigo, W. A., K. Scipal, R. M. Parinussa, Y. Y. Liu, W. Wagner, R. A. M. de Jeu, and V. Naeimi (2010), Error characterisation of global active and passive microwave soil moisture datasets, *Hydrol. Earth Syst. Sci.*, **14**(12), p. 2605–2616, doi:10.5194/hessd-7-5621-2010.

Dorigo, W. A., W. Wagner, R. Hohensinn, S. Hahn, C. Paulik, A. Xaver, A. Gruber, M. Drusch, S. Mecklenburg, P. van Oevelen, A. Robock, and T. Jackson (2011b), The international soil moisture network: a data hosting facility for global in situ soil moisture measurements, *Hydrol. Earth Syst. Sci.*, **15**(5), p. 1675–1698, doi:10.5194/hess-15-1675-2011.

Draper, C., and R. Reichle (2015), The impact of near-surface soil moisture assimilation at subseasonal, seasonal, and inter-annual timescales, *Hydrology and Earth System Sciences*, **19**(12), p. 4831, doi:10.5194/hess-19-4831-2015.

Draper, C., R. Reichle, G. De Lannoy, and Q. Liu (2012), Assimilation of passive and active microwave soil moisture retrievals, *Geophysical Research Letters*, **39**(4), doi:10.1029/2011GL050655.

Draper, C., R. Reichle, R. de Jeu, V. Naeimi, R. Parinussa, and W. Wagner (2013), Estimating root mean square errors in remotely sensed soil moisture over continental scale domains, *Remote Sensing of Environment*, **137**, p. 288–298, doi:10.1016/j.rse.2013.06.013.

Efron, B., and R. Tibshirani (1986), Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, *Statistical science*, **1**(1), p. 54–75, doi:10.1214/ss/1177013815.

Entekhabi, D., E. Njoku, P. O'Neill, K. Kellogg, W. Crow, W. Edelstein, J. Entin, S. Goodman, T. Jackson, J. Johnson, J. Kimball, J. Piepmeier, R. Koster, N. Martin, K. McDonald, M. Moghaddam, S. Moran, R. Reichle, J. Shi, M. Spencer, S. Thurman, L. Tsang, and J. Van Zyl (2010a), The soil moisture active passive (SMAP) mission, *Proceedings of the IEEE*, **98**(5), p. 704–716, doi:10.1109/JPROC.2010.2043918.

Entekhabi, D., R. H. Reichle, R. D. Koster, and W. T. Crow (2010b), Performance metrics for soil moisture retrievals and application requirements, *J. Hydrometeor*, **11**(3), p. 832–840, doi:10.1175/2010JHM1223.1.

Famiglietti, J. S., D. Ryu, A. A. Berg, M. Rodell, and T. J. Jackson (2008), Field observations of soil moisture variability across scales, *Water Resour. Res.*, **44**(1), p. W01,423, doi:10.1029/2006WR005804.

Figa-Saldaña, J., J. J. Wilson, E. Attema, R. Gelsthorpe, M. Drinkwater, and A. Stoffelen (2002), The advanced scatterometer (ASCAT) on the meteorological operational (MetOp) platform: A follow on for european wind scatterometers, *Canadian Journal of Remote Sensing*, **28**(3), p. 404–412, doi:10.5589/m02-035.

Fox, N. (2010), A guide to "reference standards" in support of quality assurance requirements of GEO, *Tech. Rep. QA4EO-QAEO-GEN-DQK-003, v4.0*, QA4EO, http://qa4eo.org/docs/QA4EO-QAEO-GEN-DQK-003_v4.0.pdf, last access: 1 July 2019.

Gelaro, R., W. McCarty, M. J. Suárez, R. Todling, A. Molod, L. Takacs, C. A. Randles, A. Darmenov, M. G. Bosilovich, R. Reichle, et al. (2017), The modern-era retrospective analysis for research and applications, version 2 (MERRA-2), *Journal of Climate*, **30**(14), p. 5419–5454, doi:10.1175/JCLI-D-16-0758.1.

Gelman, A., and H. Stern (2006), The difference between "significant" and "not significant" is not itself statistically significant, *The American Statistician*, **60**(4), p. 328–331, doi:10.1198/000313006X152649.

Gilleland, E. (2010), Confidence intervals for forecast verification, *NCAR Technical Note*, **TN-479**, doi:10.5065/D6WD3XJM.

GMAO (2015), Global Modeling and Assimilation Office (GMAO), MERRA-2 tavg1_2d_lnd_Nx: 2d,1-Hourly,Time-Averaged,Single-Level,Assimilation,Land Surface Diagnostics V5.12.4, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed: 1 Nov 2018, doi:10.5067/RKPHT8KC1Y1T.

Greenland, S., S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman (2016), Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations, *European journal of epidemiology*, **31**(4), p. 337–350, doi:10.1007/s10654-016-0149-3.

Gruber, A., W. Dorigo, S. Zwieback, A. Xaver, and W. Wagner (2013a), Characterizing coarse-scale representativeness of in situ soil moisture measurements from the international soil moisture network, *Vadose Zone Journal*, **12**(2), doi:10.2136/vzj2012.0170.

Gruber, A., W. Wagner, A. Hegyiova, F. Greifeneder, and S. Schlaffer (2013b), Potential of sentinel-1 for high-resolution soil moisture monitoring, in *Geoscience and Remote Sensing Symposium (IGARSS), 2013 IEEE International*, p. 4030–4033, IEEE, doi:10.1109/IGARSS.2017.8127513.

Gruber, A., W. Crow, W. Dorigo, and W. Wagner (2015), The potential of 2D Kalman filtering for soil moisture data assimilation, *Remote Sensing of Environment*, **171**, p. 137–148, doi:10.1016/j.rse.2015.10.019.

Gruber, A., C.-H. Su, S. Zwieback, W. Crow, W. Dorigo, and W. Wagner (2016a), Recent advances in (soil moisture) triple collocation analysis, *International Journal of Applied Earth Observation and Geoinformation*, **45**, p. 200–211, doi:10.1016/j.jag.2015.09.002.

Gruber, A., C.-H. Su, W. Crow, S. Zwieback, W. Dorigo, and W. Wagner (2016b), Estimating error cross-correlations in soil moisture data sets using extended collocation analysis, *Journal of Geophysical Research: Atmospheres*, **121(3)**, p. 1208–1219, doi:10.1002/2015JD024027.

Gruber, A., W. A. Dorigo, W. Crow, and W. Wagner (2017), Triple collocation-based merging of satellite soil moisture retrievals, *IEEE Transactions on Geoscience and Remote Sensing*, **55**(12), p. 6780–6792, doi:10.1109/TGRS.2017.2734070.

Gruber, A., W. Crow, and W. Dorigo (2018), Assimilation of spatially sparse in situ soil moisture networks into a continuous model domain, *Water Resources Research*, **54**(2), p. 1353–1367, doi:10.1002/2017WR021277.

Gruber, A., T. Scanlon, R. van der Schalie, W. Wagner, and W. Dorigo (2019a), Evolution of the esa cci soil moisture climate data records and their underlying merging methodology, *Earth System Science Data*, **11**(2), p. 717–739, doi:10.5194/essd-11-717-2019.

Gruber, A., G. D. Lannoy, and W. Crow (2019b), A monte carlo based adaptive kalman filtering framework for soil moisture data assimilation, *Remote Sensing of Environment*, **228**, p. 105 – 114, doi:10.1016/j.rse.2019.04.003.

Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez (2009), Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *Journal of Hydrology*, **377**(1), p. 80–91, doi:10.1016/j.jhydrol.2009.08.003.

H-SAF (2017), Product validation report (PVR) h111 metop ASCAT soil moisture, *Tech. Rep. SAF/HSAF/CDOP3/PVR/H111, v0.3*, EUMETSAT H SAF reports, http://hsaf.meteoam.it/documents/PVR/H111_ASCAT_SSM_CDR_PVR_v0.3.pdf (last access: 1 July 2019).

H-SAF (2018), Algorithm theoretical baseline document (ATBD) soil moisture data records, metop ASCAT soil moisture time series, *Tech. Rep. SAF/HSAF/CDOP3/ATBD, v0.7*, EUMETSAT H SAF reports, http://hsaf.meteoam.it/documents/ATDD/ASCAT_SSM_CDR_ATBD_v0.7.pdf (last access: 1 July 2019).

Jackson, T., M. Cosh, R. Bindlish, P. Starks, D. Bosch, M. Seyfried, D. Goodrich, M. Moran, and J. Du (2010), Validation of advanced microwave scanning radiometer soil moisture products, *Geoscience and Remote Sensing, IEEE Transactions on*, **48**(12), p. 4256–4272, doi:10.1109/TGRS.2010.2051035.

Jackson, T., A. Colliander, J. Kimball, R. Reichle, W. Crow, D. Entekhabi, and P. Neill (2012), Science data calibration and validation plan, *SMAP Mission, NASA Jet Propuls. Lab.*

Jackson, T. J., D. M. Le Vine, C. T. Swift, T. J. Schmugge, and F. R. Schiebe (1995), Large area mapping of soil moisture using the ESTAR passive microwave radiometer in washita'92, *Remote sensing of Environment*, **54**(1), p. 27–37, doi:10.1016/0034-4257(95)00084-E.

Jackson, T. J., D. M. Le Vine, A. Y. Hsu, A. Oldak, P. J. Starks, C. T. Swift, J. D. Isham, and M. Haken (1999), Soil moisture mapping at regional scales using microwave radiometry: The southern great plains hydrology experiment, *IEEE transactions on geoscience and remote sensing*, **37**(5), p. 2136–2151, doi:10.1109/36.789610.

Jackson, T. J., R. Bindlish, A. J. Gasiewski, B. Stankov, M. Klein, E. G. Njoku, D. Bosch, T. L. Coleman, C. A. Laymon, and P. Starks (2005), Polarimetric scanning radiometer C- and X-band microwave observations during SMEX03, *IEEE Transactions on Geoscience and Remote Sensing*, **43**(11), p. 2418–2430, doi:10.1109/TGRS.2005.857625.

JCGM (2008), Evaluation of measurement data–guide to the expression of uncertainty in measurement (GUM), *Tech. Rep. JCGM 100:2008*, Bureau International des Poids et Mesures (BIPM), Joint Committee for Guides in Metrology (JCGM), URL: https://www.bipm.org/en/publications/guides/gum.html, last access: 1 July 2019.

JCGM (2012), International vocabulary of metrology–basic and general concepts and associated terms (VIM 3rd edition), *Tech. Rep. JCGM 200:2012*, Bureau International des Poids et Mesures (BIPM), Joint Committee for Guides in Metrology (JCGM), URL: https://www.bipm.org/en/publications/guides/vim.html, last access: 1 July 2019.

Justice, C., A. Belward, J. Morisette, P. Lewis, J. Privette, and F. Baret (2000), Developments in the 'validation' of satellite sensor products for the study of the land surface, *International Journal of Remote Sensing*, **21**(17), p. 3383–3390, doi:10.1080/014311600750020000.

Kerr, Y., P. Waldteufel, J.-P. Wigneron, S. Delwart, F. Cabot, J. Boutin, M. Escorihuela, J. Font, N. Reul, C. Gruhier, S. Juglea, M. Drinkwater, A. Hahne, M. Martin-Neira, and S. Mecklenburg (2010), The SMOS mission: New tool for monitoring key elements ofthe global water cycle, *Proceedings of the IEEE*, **98**(5), p. 666–687, doi:10.1109/JPROC.2010.2043032.

Kerr, Y. H., P. Waldteufel, J.-P. Wigneron, J. Martinuzzi, J. Font, and M. Berger (2001), Soil moisture retrieval from space: The soil moisture and ocean salinity (SMOS) mission, *IEEE transactions on Geoscience and remote sensing*, **39**(8), p. 1729–1735, doi:10.1109/36.942551.

Kerr, Y. H., P. Waldteufel, P. Richaume, J. P. Wigneron, P. Ferrazzoli, A. Mahmoodi, A. Al Bitar, F. Cabot, C. Gruhier, S. E. Juglea, et al. (2012), The SMOS soil moisture retrieval algorithm, *IEEE Transactions on Geoscience and Remote Sensing*, **50**(5), p. 1384–1403, doi:10.1109/TGRS.2012.2184548.

Kerr, Y. H., A. Al-Yaari, N. Rodriguez-Fernandez, M. Parrens, B. Molero, D. Leroux, S. Bircher, A. Mahmoodi, A. Mialon, P. Richaume, et al. (2016), Overview of SMOS performance in terms of global soil moisture monitoring after six years in operation, *Remote Sensing of Environment*, **180**, p. 40–63, doi:10.1016/j.rse.2016.02.042.

Kolassa, J., P. Gentine, C. Prigent, F. Aires, and S. Alemohammad (2017), Soil moisture retrieval from amsr-e and ascat microwave observation synergy. part 2: Product evaluation, *Remote Sensing of Environment*, **195**, p. 202 – 217, doi:https://doi.org/10.1016/j.rse.2017.04.020.

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, *Journal of Geophysical Research: Atmospheres*, **105**(D20), p. 24,809–24,822, doi:10.1029/2000JD900327.

Koster, R. D., Z. Guo, R. Yang, P. A. Dirmeyer, K. Mitchell, and M. J. Puma (2009), On the nature of soil moisture in land surface models, *Journal of Climate*, **22**(16), p. 4322–4335, doi:10.1175/2009JCLI2832.1.

Kumar, S. V., R. H. Reichle, K. W. Harrison, C. D. Peters-Lidard, S. Yatheendradas, and J. A. Santanello (2012), A comparison of methods for a priori bias correction in soil moisture data assimilation, *Water Resour. Res.*, **48**(3), p. W03,515, doi:10.1029/2010WR010261.

Lahoz, W. A., and G. J. De Lannoy (2014), Closing the gaps in our knowledge of the hydrological cycle over land: Conceptual problems, *Surveys in Geophysics*, **35**(3), p. 623–660, doi:10.1007/s10712-013-9221-7.

Loew, A., W. Bell, L. Brocca, C. E. Bulgin, J. Burdanowitz, X. Calbet, R. V. Donner, D. Ghent, A. Gruber, T. Kaminski, et al. (2017), Validation practices for satellite-based earth observation data across communities, *Reviews of Geophysics*, **55**(3), p. 779–817, doi:10.1002/2017RG000562.

Macelloni, G., M. Brogioni, P. Pampaloni, A. Cagnati, and M. R. Drinkwater (2006), DOMEX 2004: An experimental campaign at Dome-C antarctica for the calibration of spaceborne low-frequency microwave radiometers, *IEEE transactions on geoscience and remote sensing*, **44**(10), p. 2642–2653, doi:10.1109/TGRS.2006.882801.

Magagi, R., A. A. Berg, K. Goïta, S. Bélair, T. J. Jackson, B. Toth, A. Walker, H. McNairn, P. E. O'Neill, M. Moghaddam, et al. (2013), Canadian experiment for soil moisture in 2010 (CanEx-SM10): Overview and preliminary results, *IEEE Transactions on Geoscience and Remote Sensing*, **51**(1), p. 347–363, doi:10.1109/TGRS.2012.2198920.

McColl, K. A., J. Vogelzang, A. G. Konings, D. Entekhabi, M. Piles, and A. Stoffelen (2014), Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target, *Geophysical Research Letters*, **41**(17), p. 6229–6236, doi:10.1002/2014GL061322.

McColl, K. A., A. Roy, C. Derksen, A. G. Konings, S. H. Alemohammed, and D. Entekhabi (2016), Triple collocation for binary and categorical variables: Application to validating landscape freeze/thaw retrievals, *Remote Sensing of Environment*, **176**, p. 31–42, doi:10.1016/j.rse.2016.01.010.

McNairn, H., T. J. Jackson, G. Wiseman, S. Belair, A. Berg, P. Bullock, A. Colliander, M. H. Cosh, S.-B. Kim, R. Magagi, et al. (2015), The soil moisture active passive validation experiment 2012 (SMAPVEX12): Prelaunch calibration and validation of the SMAP soil moisture algorithms, *IEEE Transactions on Geoscience and Remote Sensing*, **53**(5), p. 2784–2801, doi: 10.1109/TGRS.2014.2364913.

Merchant, C. J., F. Paul, T. Popp, M. Ablain, S. Bontemps, P. Defourny, R. Hollmann, T. Lavergne, A. Laeng, G. d. Leeuw, et al. (2017), Uncertainty information in climate data records from earth observation, *Earth System Science Data*, **9**(2), p. 511–527, doi: 10.5194/essd-9-511-2017.

Miralles, D. G., W. T. Crow, and M. H. Cosh (2010), Estimating spatial sampling errors in coarse-scale soil moisture estimates derived from point-scale observations, *J. Hydrometeor*, **11**(6), p. 1423–1429, doi:10.1175/2010JHM1285.1.

Miyaoka, K., A. Gruber, F. Ticconi, S. Hahn, W. Wagner, J. Figa-Saldana, and C. Anderson (2017), Triple collocation analysis of soil moisture from Metop-A ASCAT and SMOS against JRA-55 and ERA-Interim, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **10**(5), p. 2274–2284, doi:10.1109/JSTARS.2016.2632306.

Moghaddam, M., D. Entekhabi, Y. Goykhman, K. Li, M. Liu, A. Mahajan, A. Nayyar, D. Shuman, and D. Teneketzis (2010), A wireless soil moisture smart sensor web using physics-based optimal control: Concept and initial demonstrations, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **3**(4), p. 522–535, doi: 10.1109/JSTARS.2010.2052918.

Molero, B., D. Leroux, P. Richaume, Y. Kerr, O. Merlin, M. Cosh, and R. Bindlish (2018), Multi-timescale analysis of the spatial representativeness of in situ soil moisture data within satellite footprints, *Journal of Geophysical Research: Atmospheres*, **123**(1), p. 3–21, doi:10.1002/2017JD027478.

Naeimi, V., K. Scipal, Z. Bartalis, S. Hasenauer, and W. Wagner (2009), An improved soil moisture retrieval algorithm for ERS and METOP scatterometer observations, *Geoscience and Remote Sensing, IEEE Transactions on*, **47**(7), p. 1999–2013, doi:10.1109/TGRS.2008.2011617.

Naeimi, V., C. Paulik, A. Bartsch, W. Wagner, R. Kidd, S.-E. Park, K. Elger, and J. Boike (2012), ASCAT surface state flag (SSF): Extracting information on surface freeze/thaw conditions from backscatter data using an empirical threshold-analysis algorithm, *Geoscience* *and Remote Sensing, IEEE Transactions on*, **50**(7), p. 2566–2582, doi:10.1109/TGRS.2011.2177667.

Narapusetty, B., T. DelSole, and M. K. Tippett (2009), Optimal estimation of the climatological mean, *Journal of Climate*, **22**(18), p. 4845–4859, doi:10.1175/2009JCLI2944.1.

Neyman, J. (1937), X—outline of a theory of statistical estimation based on the classical theory of probability, *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, **236**(767), p. 333–380, doi:10.1098/rsta.1937.0005.

Nicolai-Shaw, N., M. Hirschi, H. Mittelbach, and S. I. Seneviratne (2015), Spatial representativeness of soil moisture using in situ, remote sensing, and land reanalysis data, *Journal of Geophysical Research: Atmospheres*, **120**(19), p. 9955–9964, doi:10.1002/2015JD023305.

Noilhan, J., P. Lacarrère, and P. Bougeault (1991), An experiment with an advanced surface parameterization in a mesobeta-scale model. part III: Comparison with the HAPEX-MOBILHY dataset, *Monthly weather review*, **119**(10), p. 2393–2413, doi:10.1175/1520-0493(1991)119⟨2393:AEWAAS⟩2.0.CO;2.

Ochsner, T. E., M. H. Cosh, R. H. Cuenca, W. A. Dorigo, C. S. Draper, Y. Hagimoto, Y. H. Kerr, E. G. Njoku, E. E. Small, M. Zreda, et al. (2013), State of the art in large-scale soil moisture monitoring, *Soil Science Society of America Journal*, **77**(6), p. 1888–1919, doi:10.2136/sssaj2013.03.0093.

Ólafsdóttir, K., and M. Mudelsee (2014), More accurate, calibrated bootstrap confidence intervals for estimating the correlation between two time series, *Mathematical Geosciences*, **46**(4), p. 411–427, doi:10.1007/s11004-014-9523-4.

O'Neill, P., S. Chan, E. Njoku, T. Jackson, and R. Bindlish (2012), SMAP level 2 & 3 soil moisture (passive) algorithm theoretical basis document (ATBD), *Initial Release, version*, **1**.

O'Neill, P., S. Chan, E. Njoku, T. Jackson, and R. Blindish (2018), SMAP L2 radiometer half-orbit 36 km EASE-grid soil moisture, version 5, *Boulder, Colorado USA. NASA National Snow and ice Data Center Distributed Active Archive Center*, doi:https://doi.org/10.5067/SODMLCE6LGLL.

Pan, M., C. K. Fisher, N. W. Chaney, W. Zhan, W. T. Crow, F. Aires, D. Entekhabi, and E. F. Wood (2015), Triple collocation: Beyond three estimates and separation of structural/nonstructural errors, *Remote Sensing of Environment*, **171**, p. 299–310, doi:doi.org/10.1016/j.rse.2015.10.028.

Panciera, R., J. P. Walker, J. D. Kalma, E. J. Kim, J. M. Hacker, O. Merlin, M. Berger, and N. Skou (2008), The NAFE'05/CoSMOS data set: Toward SMOS soil moisture retrieval, downscaling, and assimilation, *IEEE Transactions on Geoscience and Remote Sensing*, **46**(3), p. 736–745, doi:10.1109/TGRS.2007.915403.

Parinussa, R. M., A. G. Meesters, Y. Y. Liu, W. Dorigo, W. Wagner, and R. A. De Jeu (2011), Error estimates for near-real-time satellite soil moisture as derived from the land parameter retrieval model, *Geoscience and Remote Sensing Letters, IEEE*, **8**(4), p. 779–783, doi:10.1109/LGRS.2011.2114872.

Parinussa, R. M., T. R. Holmes, N. Wanders, W. A. Dorigo, and R. A. de Jeu (2015), A preliminary study toward consistent soil moisture from AMSR2, *Journal of Hydrometeorology*, **16**(2), p. 932–947, doi:10.1175/JHM-D-13-0200.1.

Pathe, C., W. Wagner, D. Sabel, M. Doubkova, and J. B. Basara (2009), Using envisat asar global mode data for surface soil moisture retrieval over oklahoma, usa, *IEEE Transactions on Geoscience and Remote Sensing*, **47**(2), p. 468–480, doi:10.1109/TGRS.2008.2004711.

Peischl, S., J. P. Walker, C. Rüdiger, N. Ye, Y. H. Kerr, E. Kim, R. Bandara, and M. Allahmoradi (2012), The AACES field experiments: SMOS calibration and validation across the murrumbidgee river catchment., *Hydrology & Earth System Sciences Discussions*, **9**(3), doi:10.5194/hessd-9-2763-2012.

Peng, J., A. Loew, S. Zhang, J. Wang, and J. Niesel (2015), Spatial downscaling of satellite soil moisture data using a vegetation temperature condition index, *IEEE Transactions on Geoscience and Remote Sensing*, **54**(1), p. 558–566, doi:10.1109/TGRS.2015.2462074.

Peng, J., A. Loew, O. Merlin, and N. E. Verhoest (2017), A review of spatial downscaling of satellite remotely sensed soil moisture, *Reviews of Geophysics*, **55**(2), p. 341–366, doi:10.1002/2016RG000543.

Pierdicca, N., F. Fascetti, L. Pulvirenti, and R. Crapolicchio (2017), Error characterization of soil moisture satellite products: Retrieving error cross-correlation through extended quadruple collocation, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **10**(10), p. 4522–4530, doi:10.1109/JSTARS.2017.2714025.

QA4EO (2010), *A Quality Assurance Framework for Earth Observation: Principles*, version 4.0 ed.

Reichle, R. H., and R. D. Koster (2004), Bias reduction in short records of satellite soil moisture, *Geophys. Res. Lett.*, **31**(19), p. L19,501, doi:10.1029/2004GL020938.

Reichle, R. H., R. D. Koster, G. J. De Lannoy, B. A. Forman, Q. Liu, S. P. Mahanama, and A. Touré (2011), Assessment and enhancement of MERRA land surface hydrology estimates, *Journal of climate*, **24**(24), p. 6322–6338, doi:10.1175/JCLI-D-10-05033.1.

Reichle, R. H., C. S. Draper, Q. Liu, M. Girotto, S. P. Mahanama, R. D. Koster, and G. J. De Lannoy (2017a), Assessment of MERRA-2 land surface hydrology estimates, *Journal of Climate*, **30**(8), p. 2937–2960, doi:10.1175/JCLI-D-16-0720.1.

Reichle, R. H., G. J. De Lannoy, Q. Liu, R. D. Koster, J. S. Kimball, W. T. Crow, J. V. Ardizzone, P. Chakraborty, D. W. Collins, A. L. Conaty, et al. (2017b), Global assessment of the SMAP level-4 surface and root-zone soil moisture product using assimilation diagnostics, *Journal of Hydrometeorology*, **18**(12), p. 3217–3237, doi:10.1175/JHM-D-17-0130.1.

Reichle, R. H., G. J. De Lannoy, Q. Liu, J. V. Ardizzone, A. Colliander, A. Conaty, W. Crow, T. J. Jackson, L. A. Jones, J. S. Kimball, et al. (2017c), Assessment of the SMAP level-4 surface and root-zone soil moisture product using in situ measurements, *Journal of hydrometeorology*, **18**(10), p. 2621–2645, doi:10.1175/JHM-D-17-0063.1.

Rodell, M., P. Houser, U. e. a. Jambor, J. Gottschalck, K. Mitchell, C. Meng, K. Arsenault, B. Cosgrove, J. Radakovich, M. Bosilovich, et al. (2004), The global land data assimilation system, *Bulletin of the American Meteorological Society*, **85**(3), p. 381–394, doi:10.1175/BAMS-85-3-381.

Rüdiger, C., A. W. Western, J. P. Walker, A. B. Smith, J. D. Kalma, and G. R. Willgoose (2010), Towards a general equation for frequency domain reflectometers, *Journal of hydrology*, **383**(3-4), p. 319–329, doi:10.1016/j.jhydrol.2009.12.046.

Sabaghy, S., J. Walker, L. Renzullo, R. Akbar, S. Chan, J. Chaubell, N. Das, R. Dunbar, D. Entekhabi, A. Gevaert, T. Jackson, A. Loew, O. Merlin, M. Moghaddam, J. Peng, J. Peng,

J. Piepmeier, C. Rüdiger, V. Stefan, X. Wu, N. Ye, and S. Yueh (in review), Comprehensive analysis of alternative downscaled soil moisture products, *Remote Sensing of Environment*.

Sahoo, A. K., G. J. D. Lannoy, R. H. Reichle, and P. R. Houser (2013), Assimilation and downscaling of satellite observed soil moisture over the little river experimental watershed in georgia, usa, *Advances in Water Resources*, **52**, p. 19 – 33, doi:10.1016/j.advwatres.2012.08.007.

Scanlon, T., J. Nightingale, F. Boersma, J.-P. Muller, C. Farquhar, S. Compernolle, and J.-C. Lambert (2017), Outline of QA4ECV quality assurance service (version 2.0), *Tech. rep.*, QA4ECV, http://www.qa4ecv.eu/qa-system, last access: 1 July 2019.

Scipal, K., M. Drusch, and W. Wagner (2008a), Assimilation of a ERS scatterometer derived soil moisture index in the ECMWF numerical weather prediction system, *Advances in water resources*, **31**(8), p. 1101–1112, doi:10.1016/j.advwatres.2008.04.013.

Scipal, K., T. Holmes, R. de Jeu, V. Naeimi, and W. Wagner (2008b), A possible solution for the problem of estimating the error structure of global soil moisture data sets, *Geophys. Res. Lett.*, **35**(24), p. L24,403, doi:10.1029/2008GL035599.

Starks, P. J., G. C. Heathman, T. J. Jackson, and M. H. Cosh (2006), Temporal stability of soil moisture profile, *Journal of Hydrology*, **324**, p. 400–411, doi:10.1016/j.jhydrol.2005.09.024.

Stoffelen, A. (1998), Toward the true near-surface wind speed: Error modeling and calibration using triple collocation, *J. Geophys. Res.*, **103**(C4), p. 7755–7766, doi:10.1029/97JC03180.

Su, C.-H., and D. Ryu (2015), Multi-scale analysis of bias correction of soil moisture, *Hydrology and Earth System Sciences*, **19**(1), p. 17–31, doi:10.5194/hess-19-17-2015.

Su, C.-H., D. Ryu, W. T. Crow, and A. W. Western (2014), Beyond triple collocation: Applications to soil moisture monitoring, *Journal of Geophysical Research: Atmospheres*, **119**(11), p. 6419–6439, doi:10.1002/2013JD021043.

Su, Z., W. Timmermans, Y. Zeng, J. Schulz, V. O. John, R. A. Roebeling, P. Poli, D. Tan, F. Kaspar, A. K. Kaiser-Weiss, E. Swinnen, C. Toté, H. Gregow, T. Manninen, A. Riihelä, J.-C. Calvet, Y. Ma, and J. Wen (2018), An overview of european efforts in generating climate data records, *Bulletin of the American Meteorological Society*, **99**(2), p. 349–359, doi:10.1175/BAMS-D-16-0074.1.

Tong, C. (2019), Statistical inference enables bad science; statistical thinking enables good science, *The American Statistician*, **73**(sup1), p. 246–261, doi:10.1080/00031305.2018.1518264.

Ulaby, F. T., D. G. Long, W. J. Blackwell, C. Elachi, A. K. Fung, C. Ruf, K. Sarabandi, H. A. Zebker, and J. Van Zyl (2014), *Microwave radar and radiometric remote sensing*, vol. 4, University of Michigan Press Ann Arbor.

Vachaud, G., A. Passerat De Silans, P. Balabanis, and M. Vauclin (1985), Temporal stability of spatially measured soil water probability density function, *Soil Sci. Soc. Am. J.*, **49**(4), p. 822–828, doi:10.2136/sssaj1985.03615995004900040006x.

van der Schalie, R., R. de Jeu, R. Parinussa, N. Rodríguez-Fernández, Y. Kerr, A. Al-Yaari, J.-P. Wigneron, and M. Drusch (2018), The effect of three different data fusion approaches on the quality of soil moisture retrievals from multiple passive microwave sensors, *Remote Sensing*, **10**(1), p. 107, doi:10.3390/rs10010107.

Van Leeuwen, P. J. (2015), Representation errors and retrievals in linear and nonlinear data assimilation, *Quarterly Journal of the Royal Meteorological Society*, **141**(690), p. 1612–1623.

Vogelzang, J., and A. Stoffelen (2012), Triple collocation, *EUMETSAT Report. Available at http://research.metoffice.gov.uk/research/interproj/nwpsaf/scatterometer/TripleCollocation_NWPSAF_TR_ last access: 1 July 2019.*

Wagner, W., G. Lemoine, and H. Rott (1999), A method for estimating soil moisture from ERS scatterometer and soil data, *Remote Sensing of Environment*, **70**(2), p. 191–207, doi:10.1016/S0034-4257(99)00036-X.

Wagner, W., L. Brocca, V. Naeimi, R. Reichle, C. Draper, R. de Jeu, D. Ryu, C.-H. Su, A. Western, J.-C. Calvet, et al. (2014), Clarifications on the “comparison between SMOS, VUA, ASCAT, and ECMWF soil moisture products over four watersheds in US”, *IEEE Transactions on Geoscience and Remote Sensing*, **52**(3), p. 1901–1906, doi:10.1109/TGRS.2013.2282172.

Walker, J. P., G. R. Willgoose, and J. D. Kalma (2004), In situ measurement of soil moisture: a comparison of techniques, *Journal of Hydrology*, **293**, p. 85–99, doi:10.1016/j.jhydrol.2004.01.008.

Wang, G., D. Garcia, Y. Liu, R. De Jeu, and A. J. Dolman (2012), A three-dimensional gap filling method for large geophysical datasets: Application to global satellite soil moisture observations, *Environmental Modelling & Software*, **30**, p. 139–142, doi:10.1016/j.envsoft.2011.10.015.

Wasserstein, R. L., and N. A. Lazar (2016), The ASA's statement on p-values: context, process, and purpose, *The American Statistician*, **70**(2), p. 129–133, doi:10.1080/00031305.2016.1154108.

Wasserstein, R. L., A. L. Schirm, and N. A. Lazar (2019), Moving to a world beyond "p¡0.05", *The American Statistician*, **73**(sup1), p. 1–19, doi:10.1080/00031305.2019.1583913.

Wilks, D. S. (2011), *Statistical Methods in the Atmospheric Sciences*, vol. 100, 3rd ed., Academic Press.

WMO (2016), The global observing system for climate: Implementation needs, *Implementation Plan GCOS-200*, World Meteorological Organization.

Yilmaz, M. T., and W. T. Crow (2013), The optimality of potential rescaling approaches in land data assimilation., *Journal of Hydrometeorology*, **14**(2), doi:10.1175/JHM-D-12-052.1.

Yilmaz, M. T., and W. T. Crow (2014), Evaluation of assumptions in soil moisture triple collocation analysis, *Journal of Hydrometeorology*, **15**(3), p. 1293–1302, doi:10.1175/JHM-D-13-0158.1.

Zeng, Y., Z. Su, J.-C. Calvet, T. Manninen, E. Swinnen, J. Schulz, R. Roebeling, P. Poli, D. Tan, A. Riihelä, C.-M. Tanis, A.-N. Arslan, A. Obregon, A. Kaiser-Weiss, V. John, W. Timmermans, J. Timmermans, F. Kaspar, H. Gregow, A.-L. Barbu, D. Fairbairn, E. Gelati, and C. Meurey (2015), Analysis of current validation practices in europe for space-based climate data records of essential climate variables, *International Journal of Applied Earth Observation and Geoinformation*, **42**, p. 150 – 161, doi:https://doi.org/10.1016/j.jag.2015.06.006.

Zreda, M., W. Shuttleworth, X. Zeng, C. Zweck, D. Desilets, T. Franz, and R. Rosolem (2012), COSMOS: the cosmic-ray soil moisture observing system, *Hydrology and Earth System Sciences*, **16**(11), p. 4079–4099, doi:10.5194/hess-16-4079-2012.

Zribi, M., M. Pardé, J. Boutin, P. Fanise, D. Hauser, M. Dechambre, Y. Kerr, M. Leduc-Leballeur, G. Reverdin, N. Skou, S. Søbjærg, C. Albergel, J. C. Calvet, J. P. Wigneron, E. Lopez-Baeza, A. Rius, and J. Tenerelli (2011), Carols: A new airborne l-band radiometer for ocean surface and land observations, *Sensors*, **11**(1), p. 719–742, doi:10.3390/s110100719.

Zwieback, S., K. Scipal, W. Dorigo, and W. Wagner (2012), Structural and statistical properties of the collocation technique for error characterization, *Nonlin. Processes Geophys.*, **19**(1), p. 69–80, doi:10.5194/npg-19-69-2012.

Zwieback, S., A. Colliander, M. H. Cosh, J. Martínez-Fernández, H. McNairn, P. J. Starks, M. Thibeault, and A. Berg (2018), Estimating time-dependent vegetation biases in the SMAP soil moisture product, *Hydrology and Earth System Sciences*, **22**(8), p. 4473–4489, doi:10.5194/hess-22-4473-2018.

Table 1: Validation stages as defined by CEOS (modified from `https://lpvs.gsfc.nasa.gov/`; last access: 1 July 2019).

Validation Stage	Definition
0	No validation. Product accuracy has not been assessed. Product considered beta.
1	Product accuracy is assessed from a small (typically <30) set of locations and time periods by comparison with in situ or other suitable reference data.
2	Product accuracy is estimated over a considerable set of locations and time periods by comparison with reference in situ or other suitable reference data. Spatial and temporal consistency of the product and consistency with similar products has been evaluated over globally representative locations and time periods. Results are published in the peer-reviewed literature.
3	Uncertainties in the product and its associated structure are well quantified from comparison with reference in situ or other suitable reference data. Uncertainties are characterized in a statistically rigorous way over multiple locations and time periods representing global conditions. Spatial and temporal consistency of the product and with similar products has been evaluated over globally representative locations and periods. Results are published in the peer-reviewed literature.
4	Validation results for stage 3 are systematically updated when new product versions are released and as the time-series expands.

Table 2: Summary of publicly available reference data sources commonly used for satellite soil moisture validation (links last accessed: 1 July 2019).

Name	Description	Reference
ISMN	Data hosting facility for sparse soil moisture networks	`http://ismn.geo.tuwien.ac.at/` (*Dorigo et al.*, 2011a,b)
CVS	Openly available Core Validation Site (CVS) data that have been specifically processed for SMAP validation.	`https://nsidc.org/data/nsidc-0712`
GLDAS	NASA's global modelling and data assimilation system	`https://ldas.gsfc.nasa.gov/gldas/`
MERRA	NASA's global reanalysis data sets	`https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/`
ERA	ECMWF's global reanalysis data sets	`https://www.ecmwf.int/en/forecasts/datasets/browse-reanalysis-datasets/`

Table 3: Open-source software that can be used for satellite soil moisture validation (links last accessed: last access: 1 July 2019).

Name	Description	Language	Reference
	Source code used to produce validation examples in this publication in Appendix A	python, MATLAB	`https://github.com/alexgruber/validation_good_practice/`
pytesmo	Geospatial time series validation toolbox	python	`https://doi.org/10.5281/zenodo.1215760/`
poets	Geospatial image resampling toolbox	python	`https://pypi.org/project/poets/`

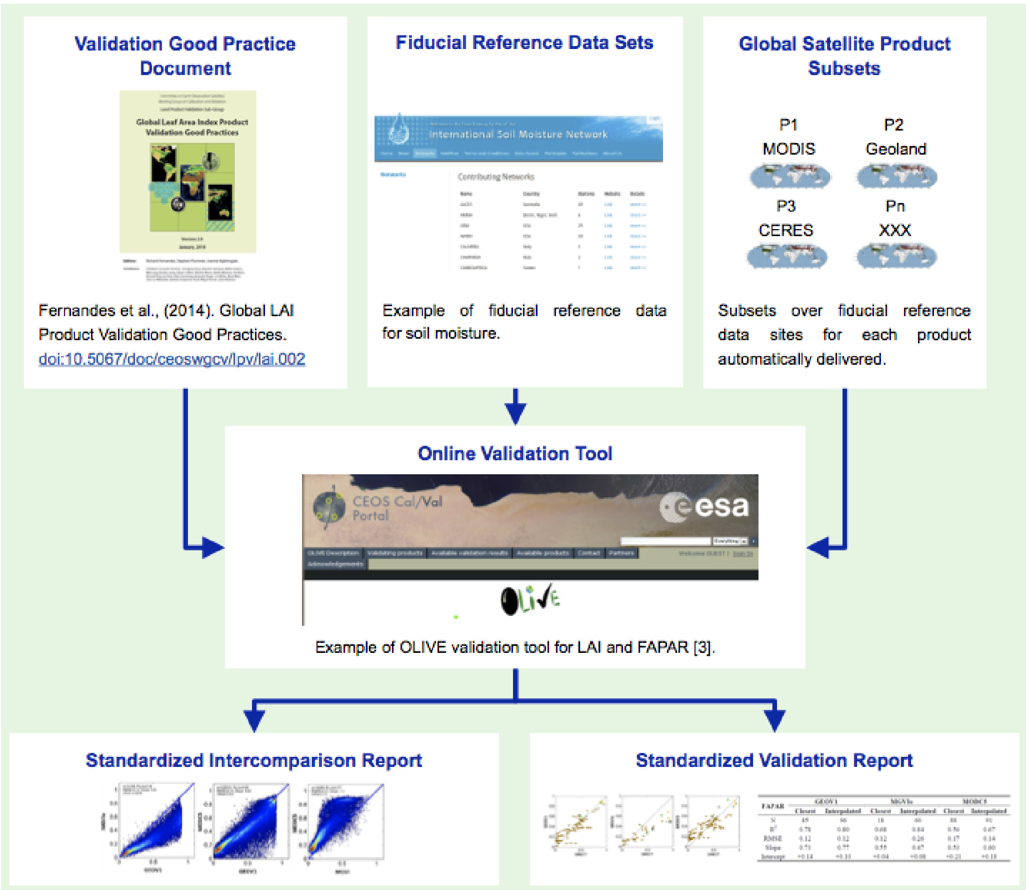

Figure 1: Validation framework as defined by CEOS (from `https://lpvs.gsfc.nasa.gov/`; last access: 1 July 2019).

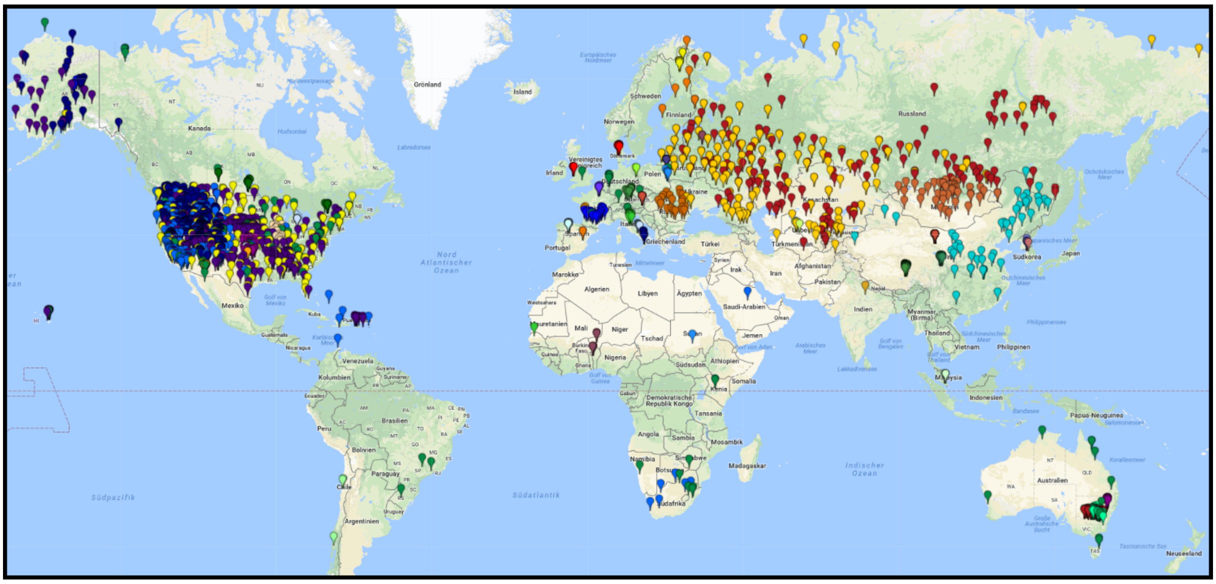

Figure 2: Currently available stations from sparse networks hosted by the ISMN. Colors represent different station hosting networks.

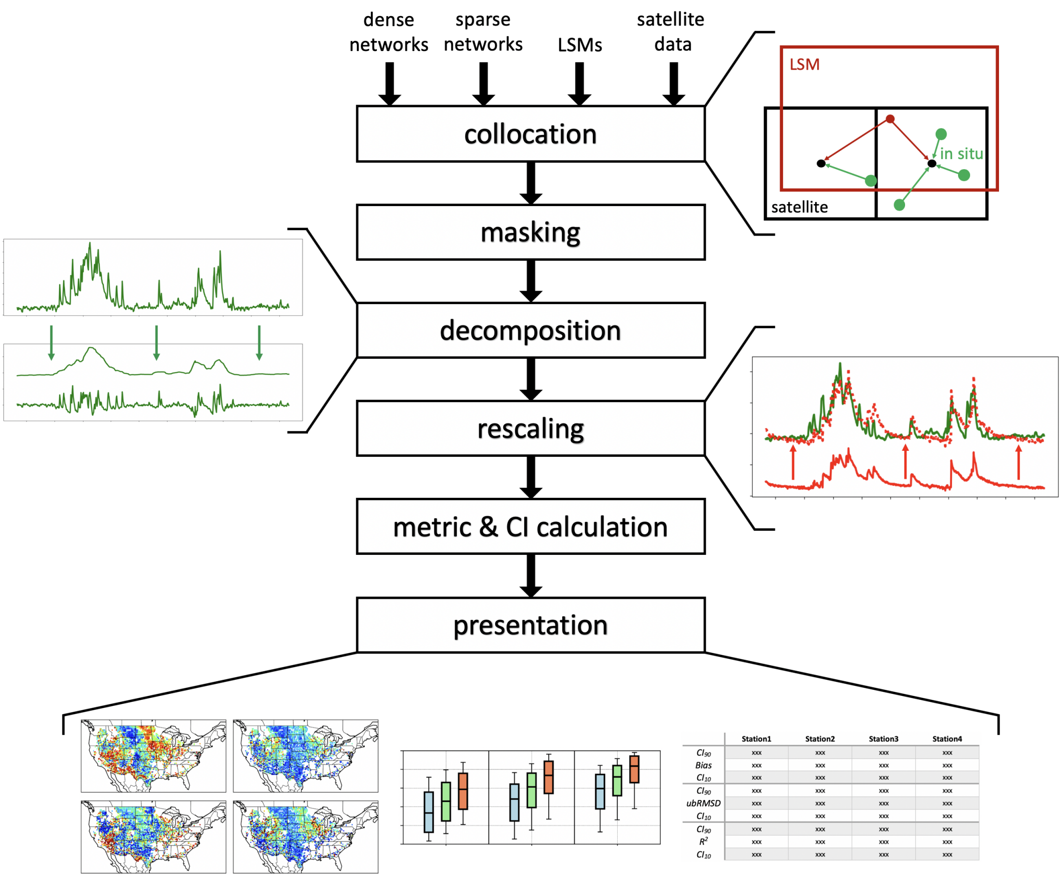

Figure 3: Validation good practice protocol illustration.

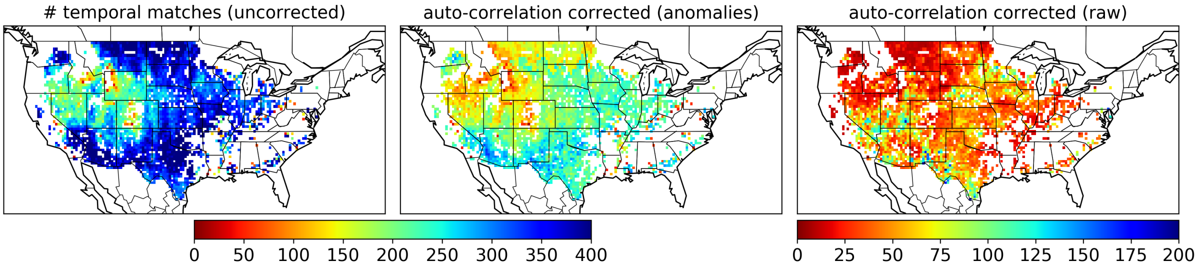

Figure A.1: Sample size for temporal matches between ASCAT, SMOS, SMAP and MERRA-2 between 2015 and 2018 (left), effective sample size when correcting for anomaly auto-correlation (middle), and effective sample size when correcting for auto-correlation in the raw time series (right).

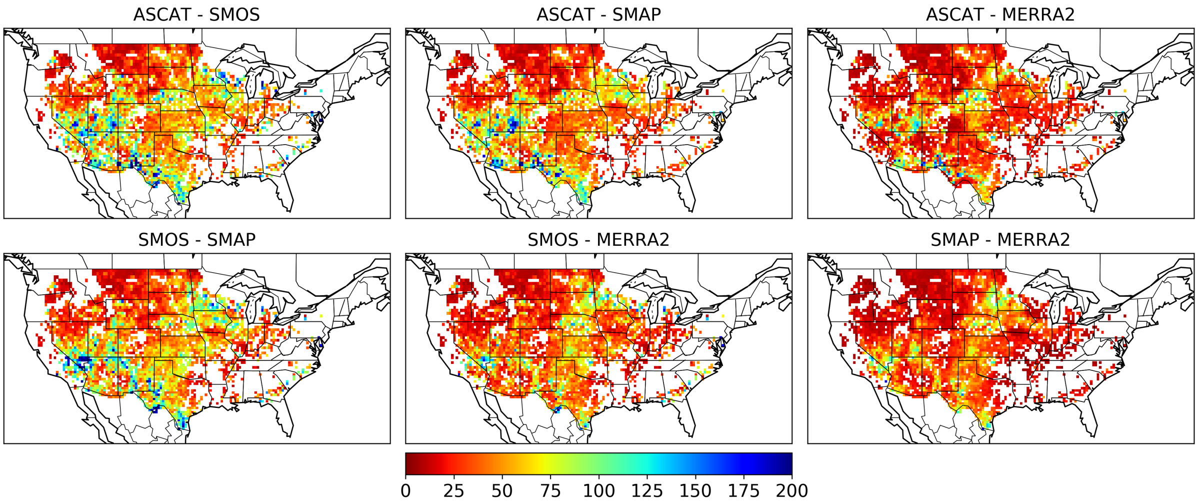

Figure A.2: Effective, raw time series auto-correlation corrected sample size for different data set combinations.

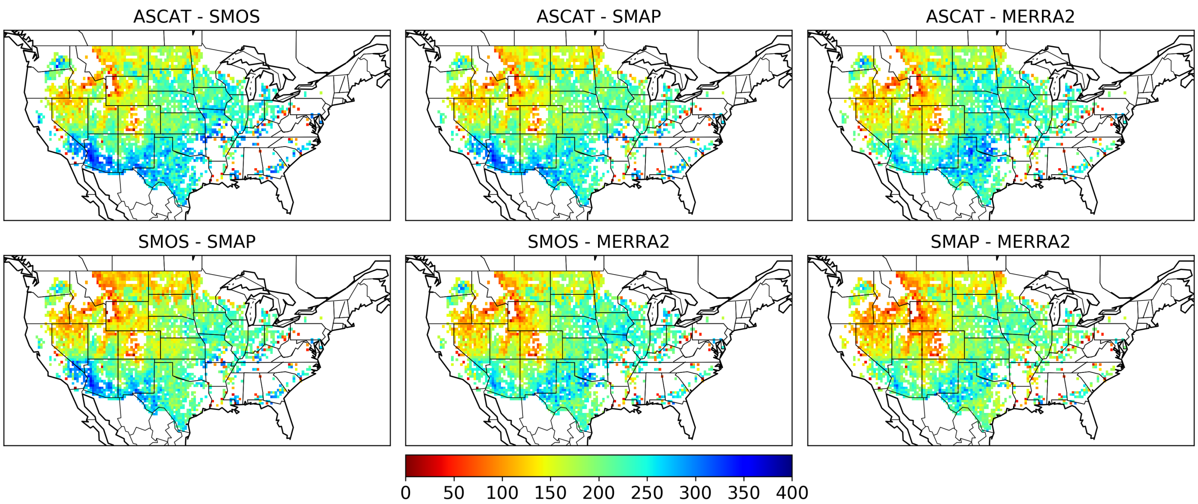

Figure A.3: Effective, anomaly auto-correlation corrected sample size for different data set combinations.

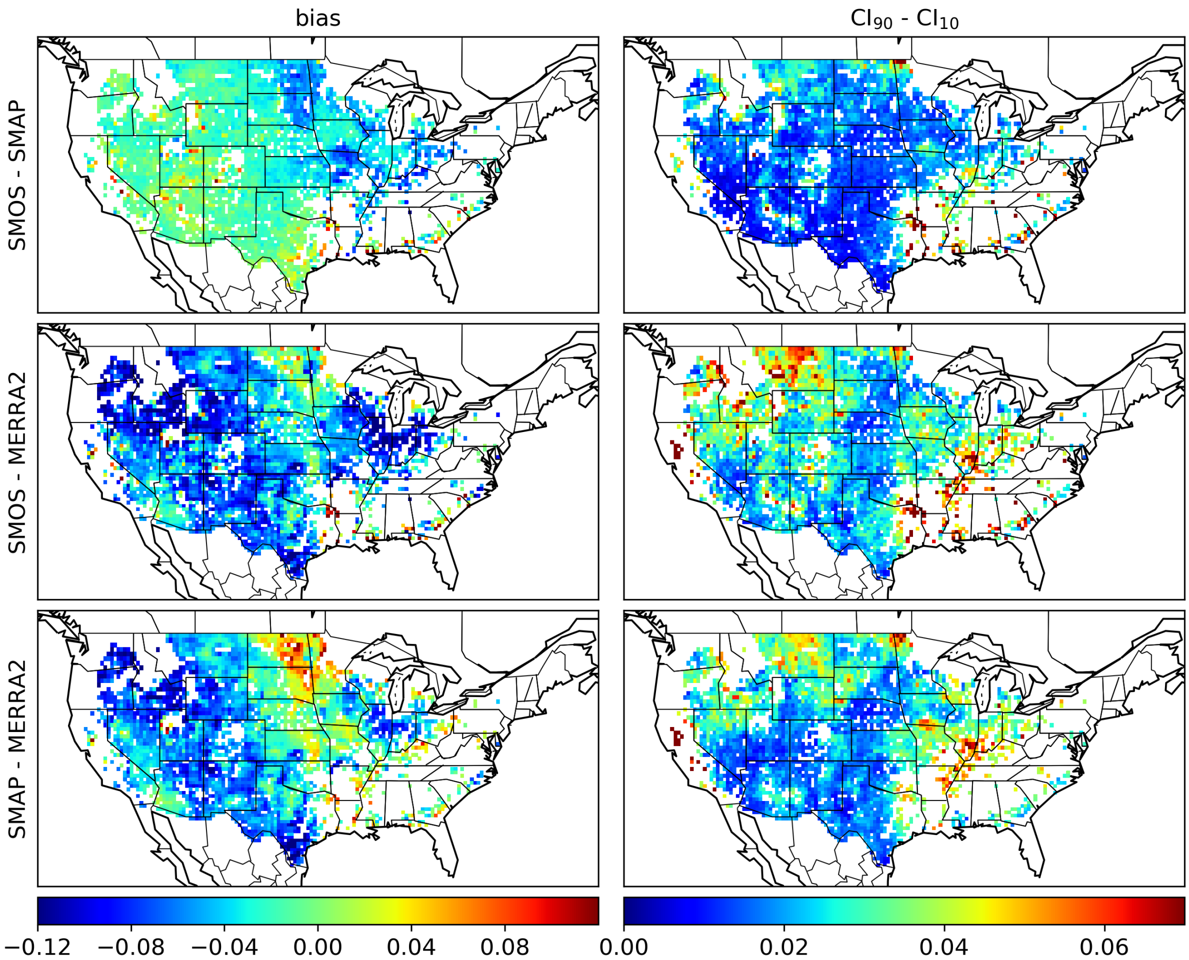

Figure A.4: Temporal mean biases $[m^3m^{-3}]$ (left) and associated 80% confidence intervals (right) between raw soil moisture estimates of SMOS, SMAP and MERRA-2.

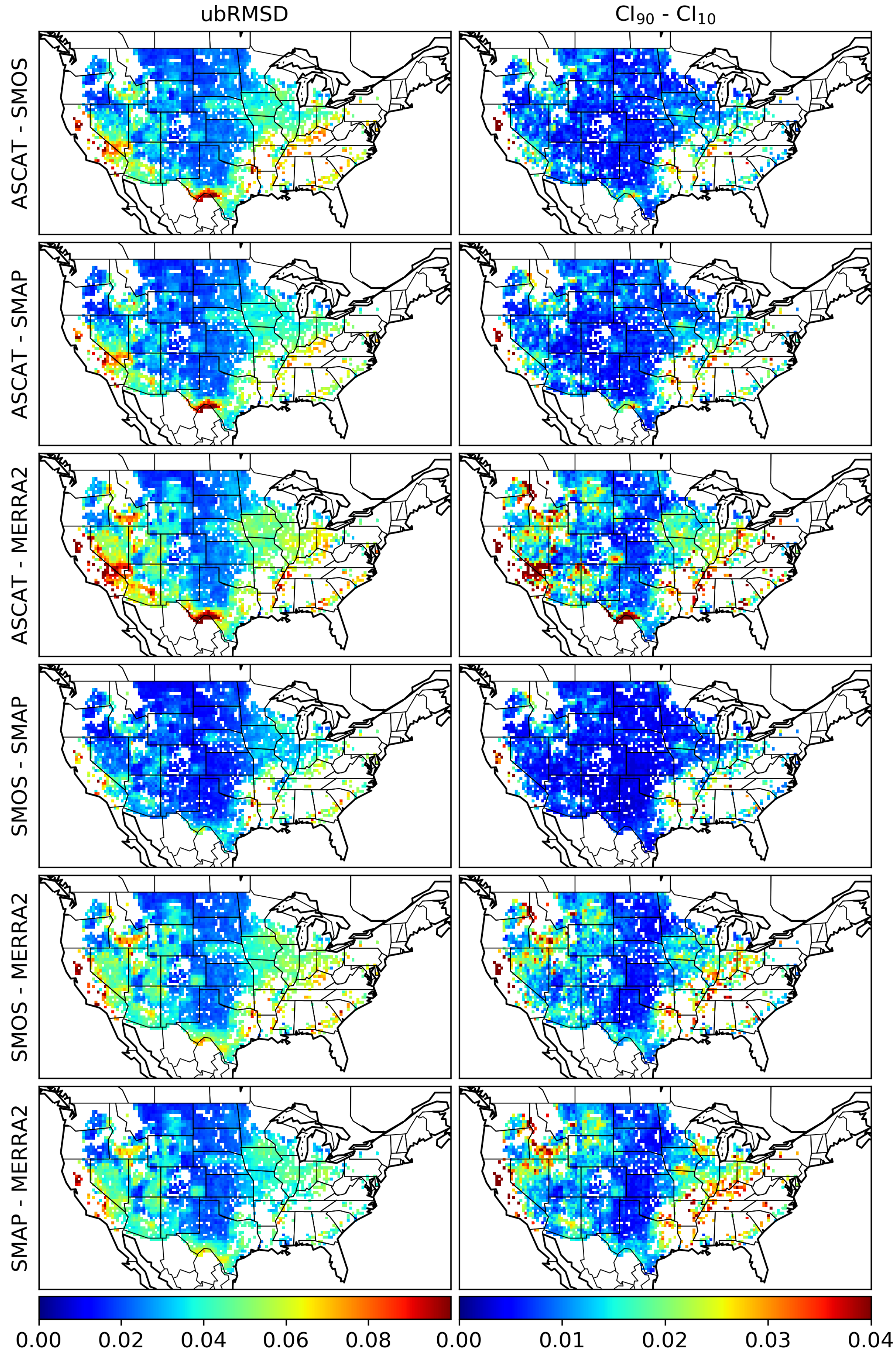

Figure A.5: Unbiased (in mean and standard deviation) root-mean-square-differences $[m^3m^{-3}]$ (left) and associated 80% confidence intervals (right) between raw soil moisture estimates of ASCAT, SMOS, SMAP and MERRA-2.

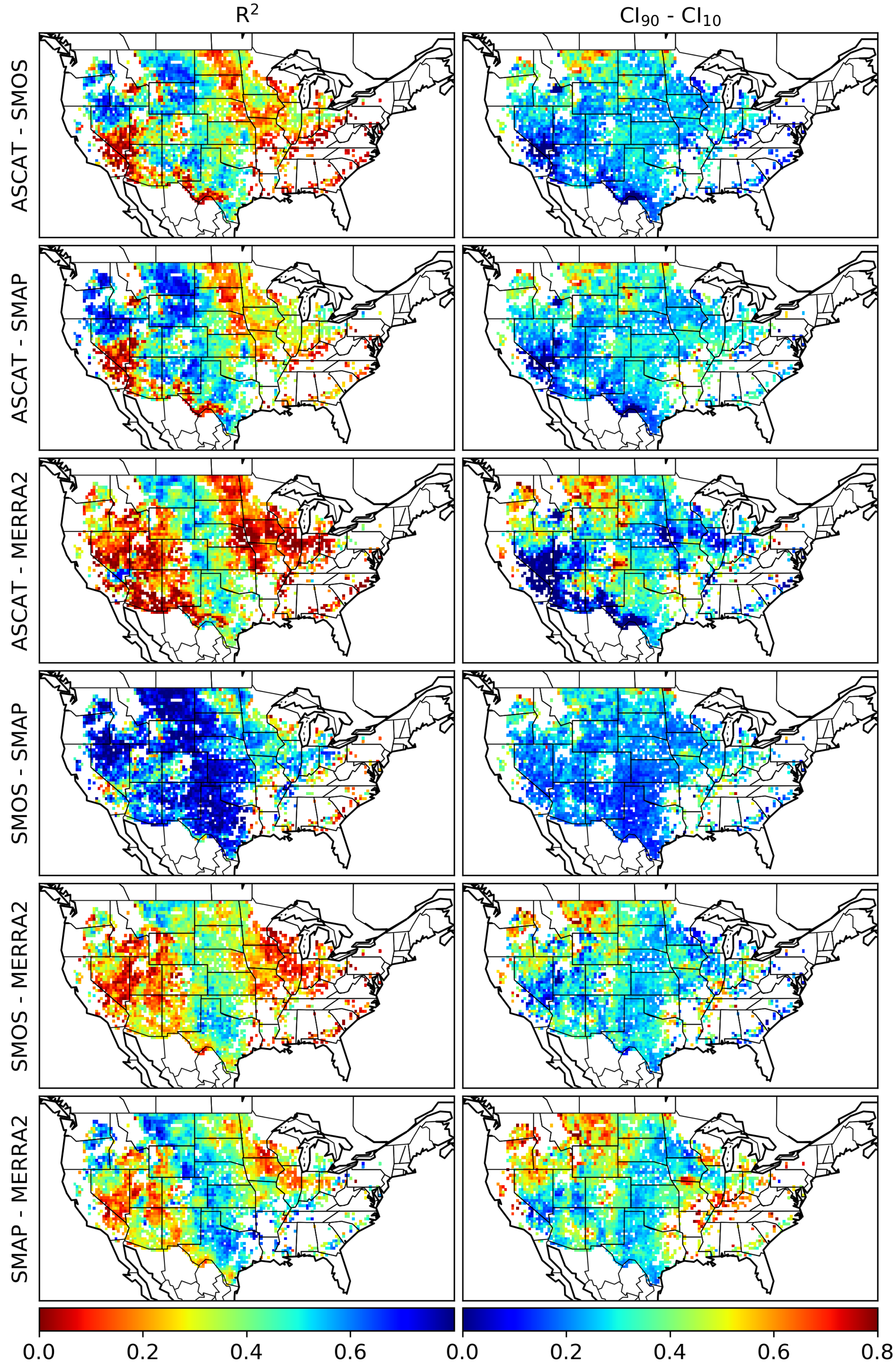

Figure A.6: Coefficients of determination [-] (left) and associated 80% confidence intervals (right) between raw soil moisture estimates of ASCAT, SMOS, SMAP and MERRA-2.

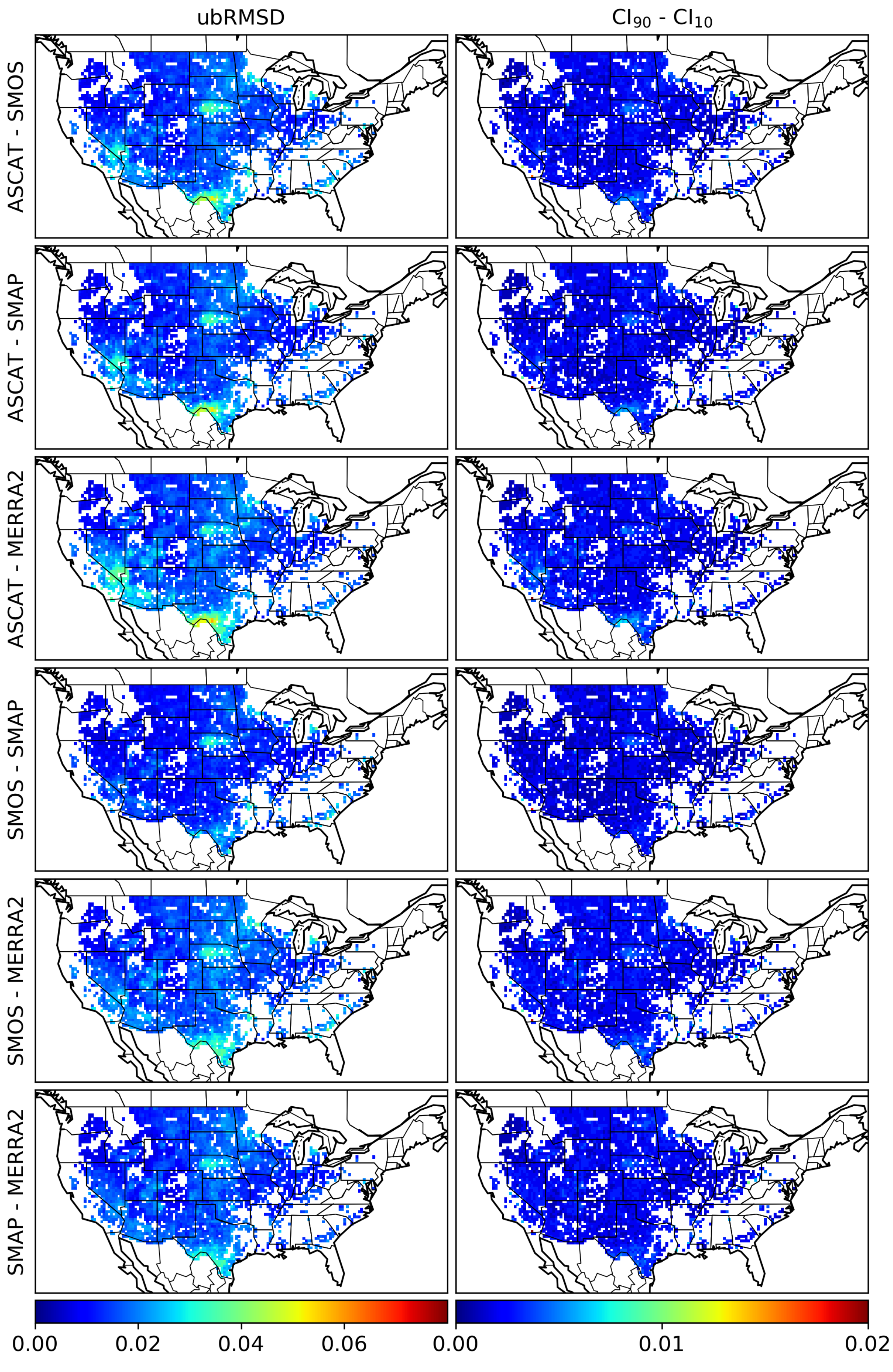

Figure A.7: Unbiased (in mean and standard deviation) $[m^3m^{-3}]$ root-mean-square-differences (left) and associated 80% confidence intervals (right) between soil moisture anomaly estimates of ASCAT, SMOS, SMAP and MERRA-2.

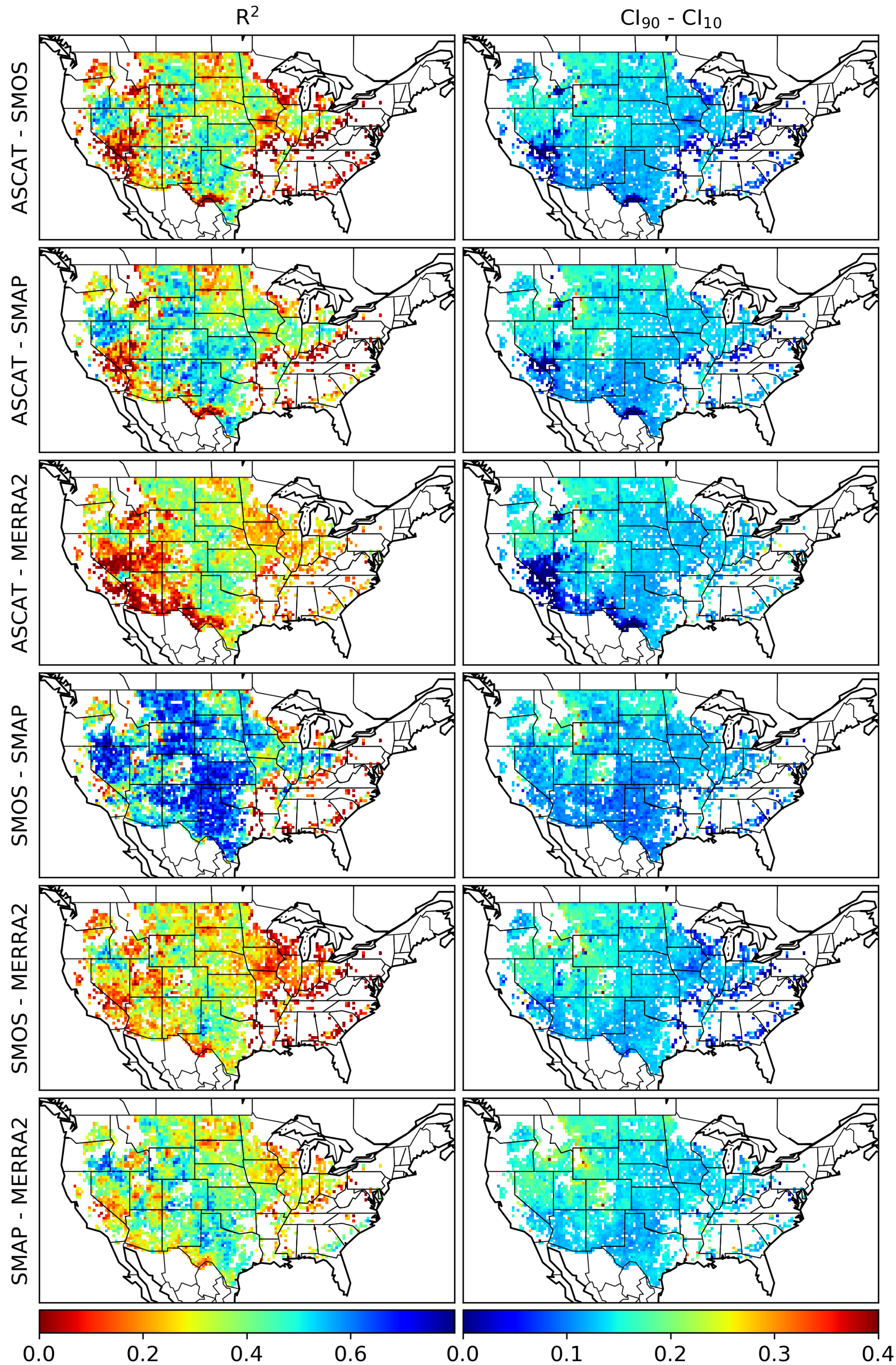

Figure A.8: Coefficients of determination [-] (left) and associated 80% confidence intervals (right) between soil moisture anomaly estimates of ASCAT, SMOS, SMAP and MERRA-2.

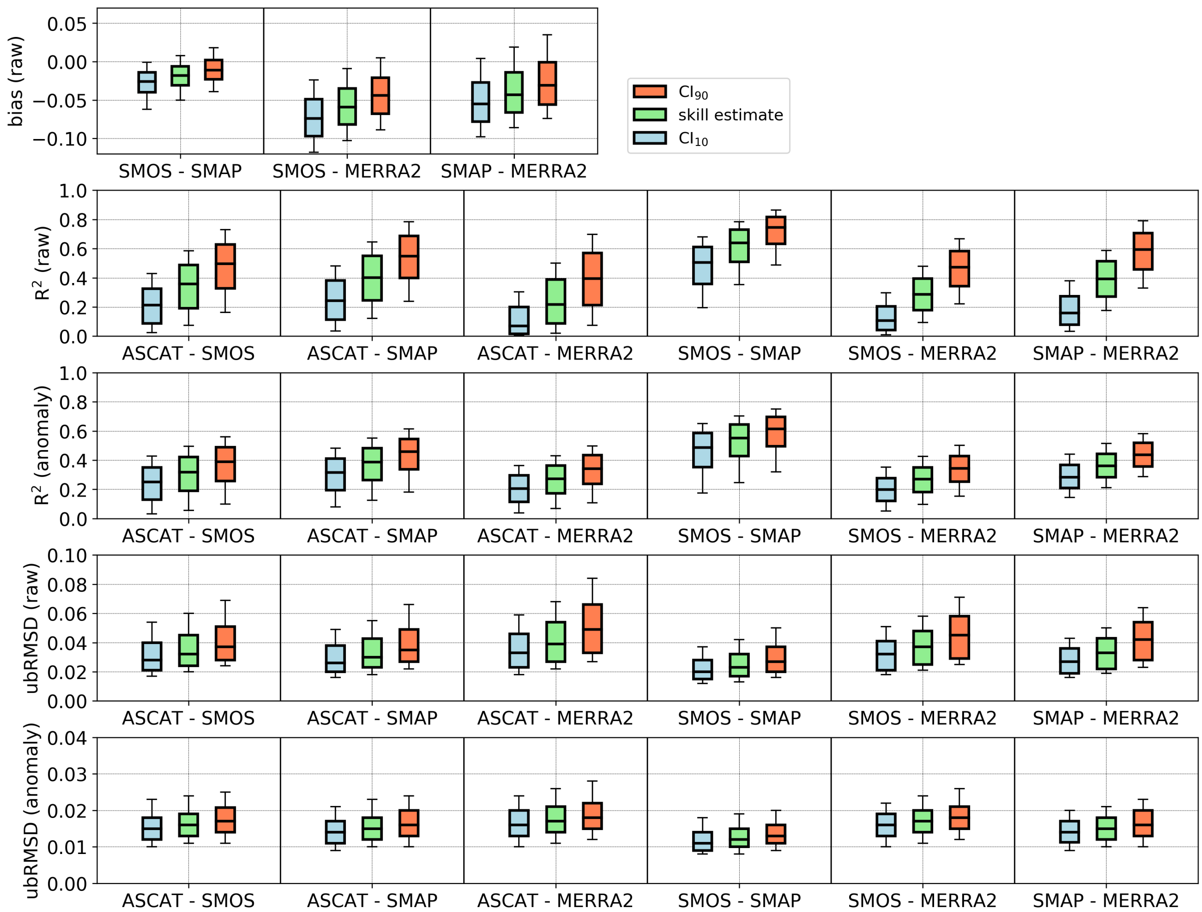

Figure A.9: Spatial summary statistics of biases $[m^3m^{-3}]$, ubRMSDs $[m^3m^{-3}]$, and coefficients of determination [-] and their 10% and 90% confidence limits, respectively, for raw soil moisture estimates and soil moisture anomalies of ASCAT, SMOS, SMAP and MERRA-2. Boxes represent the (spatial) median and inter-quartile-range and whiskers represent the 5 and 95 percentiles.

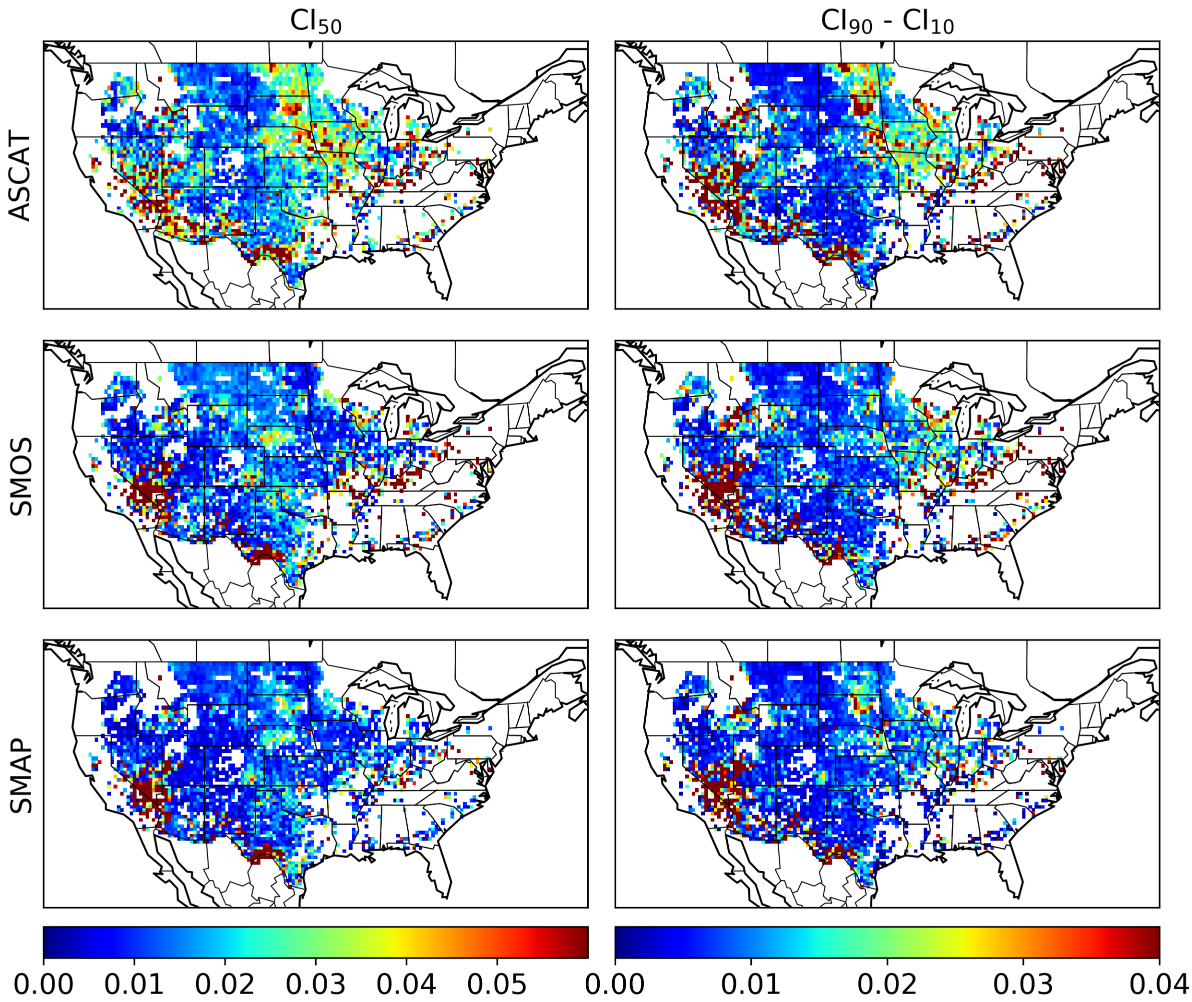

Figure A.10: Median of the bootstrapped TCA-based ubRMSEs $[m^3m^{-3}]$ (left) and associated 80% confidence intervals (right) of raw soil moisture estimates of ASCAT, SMOS, and SMAP.

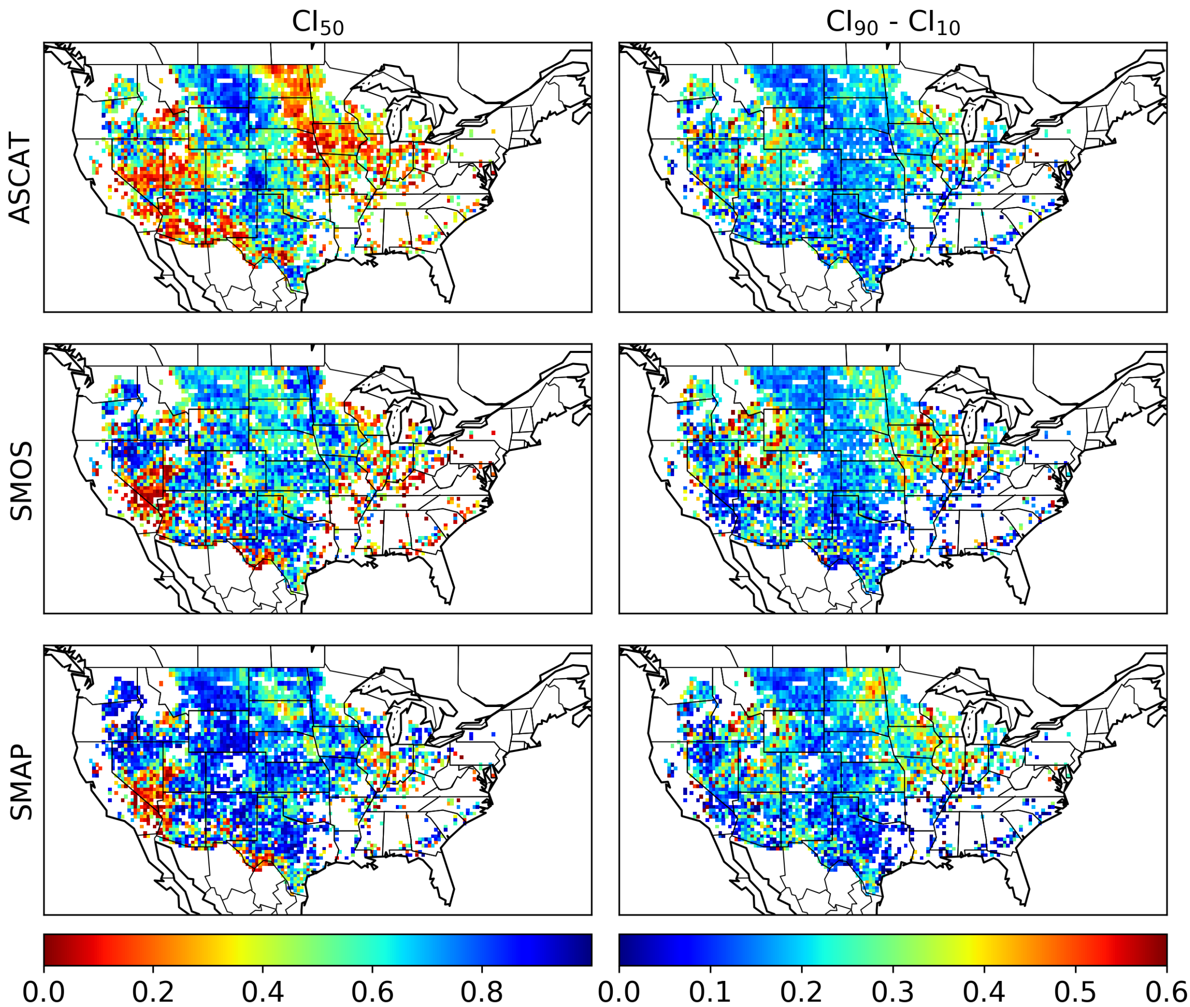

Figure A.11: Median of the bootstrapped TCA-based R^2 estimates [-] (left) and associated 80% confidence intervals (right) of raw soil moisture estimates of ASCAT, SMOS, and SMAP.

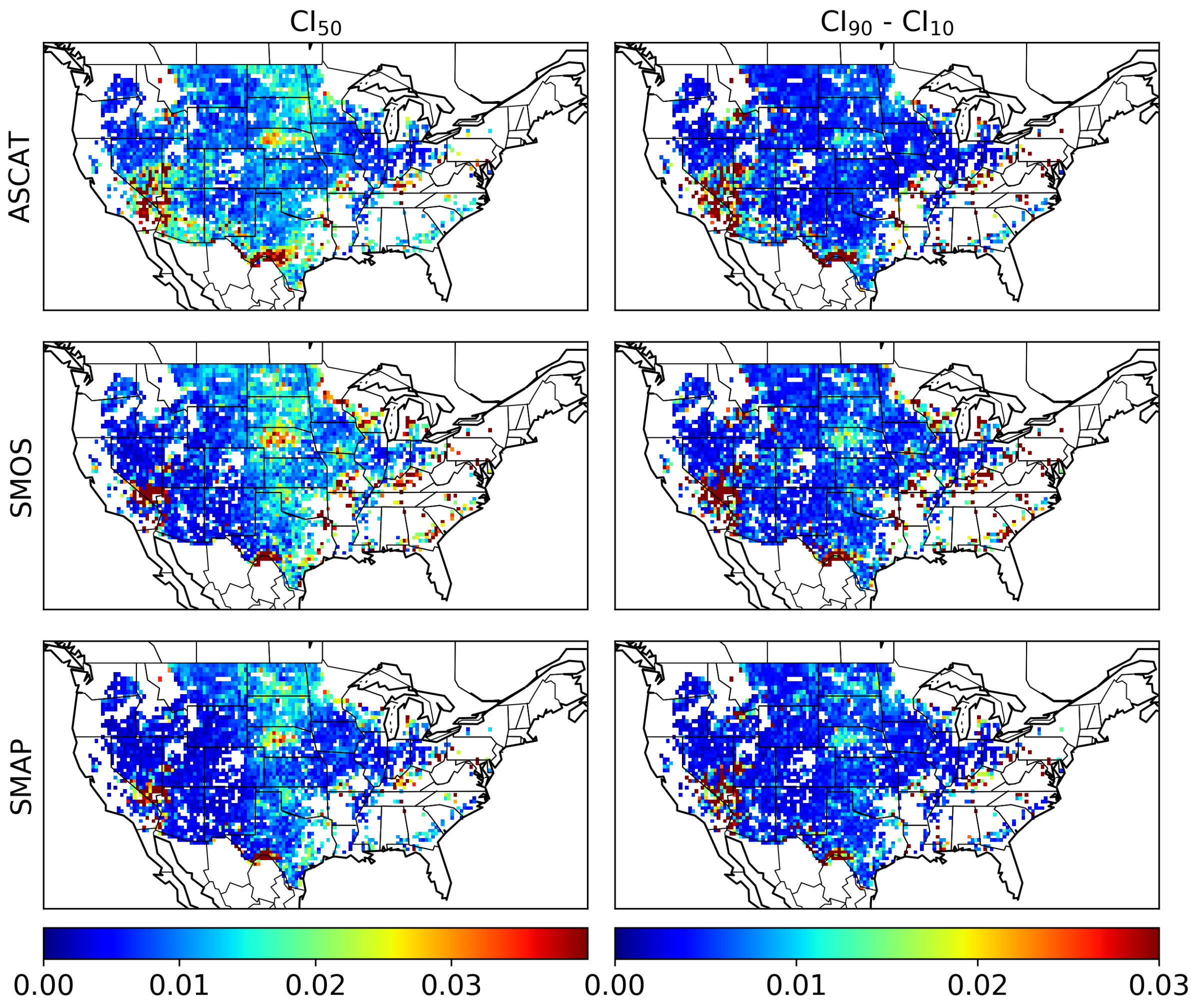

Figure A.12: Median of the bootstrapped TCA-based ubRMSEs $[m^3m^{-3}]$ (left) and associated 80% confidence intervals (right) of soil moisture anomaly estimates of ASCAT, SMOS, and SMAP.

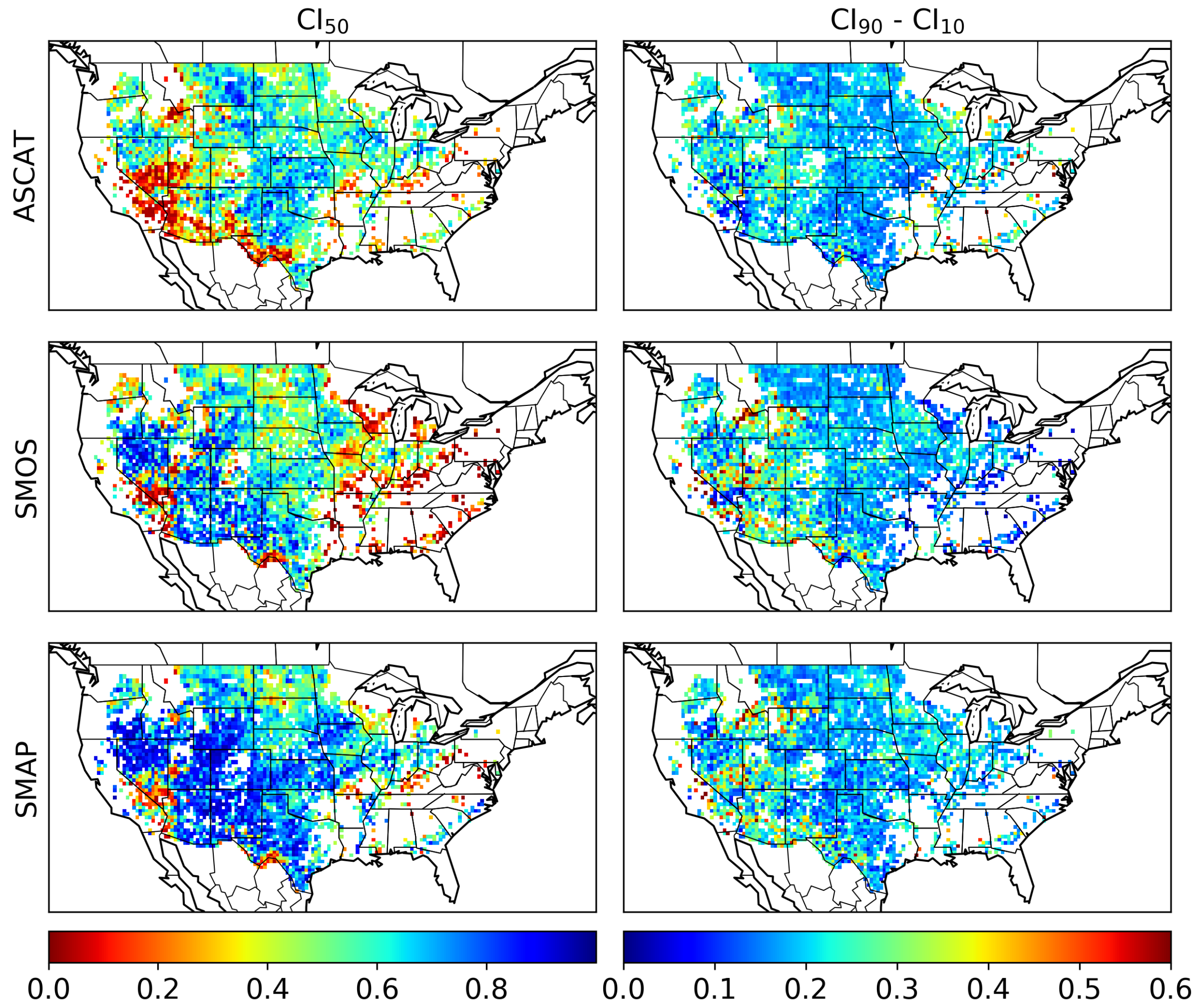

Figure A.13: Median of the bootstrapped TCA-based R^2 estimates [-] (left) and associated 80% confidence intervals (right) of soil moisture anomaly estimates of ASCAT, SMOS, and SMAP.

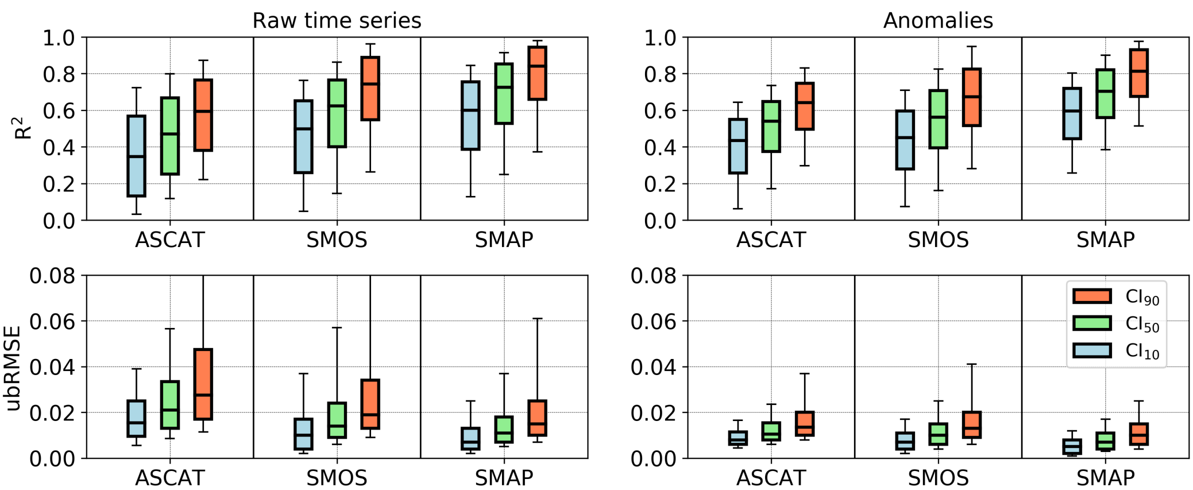

Figure A.14: Spatial summary statistics of the median of the bootstrapped TCA-based ubRMSEs $[m^3m^{-3}]$, and R^2 estimates [-] and their 10% and 90% confidence limits, respectively, for raw soil moisture estimates and soil moisture anomalies of ASCAT, SMOS, and SMAP. Boxes represent the (spatial) median and inter-quartile-range and whiskers represent the 5 and 95 percentiles.

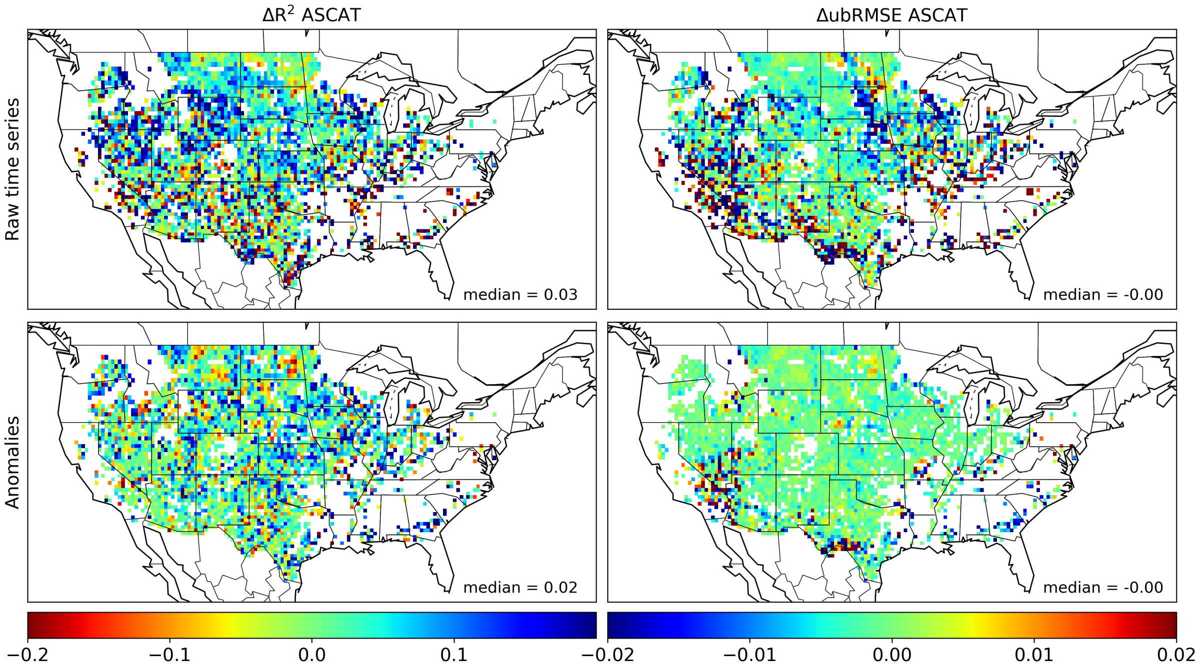

Figure A.15: Difference in TCA-based ubRMSE [m^3m^{-3}] and R^2 estimates [-] for raw soil moisture estimates (top) and soil moisture anomaly estimates (bottom) of ASCAT when using SMOS as third data set minus when using SMAP as third data set in the triplet.