AGU PUBLICATIONS

Reviews of Geophysics

REVIEW ARTICLE
10.1002/2017RG000562

Key Points:
- First review of EO validation approaches across different Geoscience communities
- Validation approaches depend on the intermittency and inhomogeneity of the geophysical variables
- Enhanced traceability in EO validation approaches required

Correspondence to:
T. Verhoelst,
tijl.verhoelst@aeronomie.be

Citation:
Loew, A., et al. (2017), Validation practices for satellite-based Earth observation data across communities, *Rev. Geophys.*, *55*, 779–817, doi:10.1002/2017RG000562.

Received 17 MAR 2017
Accepted 4 JUN 2017
Accepted article online 6 JUN 2017
Published online 21 SEP 2017

Validation practices for satellite-based Earth observation data across communities

Alexander Loew[1,2], William Bell[3], Luca Brocca[4], Claire E. Bulgin[5], Jörg Burdanowitz[6], Xavier Calbet[7], Reik V. Donner[8], Darren Ghent[9], Alexander Gruber[10], Thomas Kaminski[11], Julian Kinzel[12], Christian Klepp[13], Jean-Christopher Lambert[14], Gabriela Schaepman-Strub[15], Marc Schröder[12], and Tijl Verhoelst[14]

[1]Department of Geography, Ludwig-Maximilians-Universität München (LMU), Munich, Germany, [2]Deceased 2 July 2017,, [3]MetOffice, Reading, UK, [4]Research Institute for Geo-Hydrological Protection-National Research Council, Perugia, Italy, [5]Department of Meteorology, University of Reading, Reading, UK, [6]Max Planck Institute for Meteorology, Hamburg, Germany, [7]Spanish Meteorological Agency, AEMET, Madrid, Spain, [8]Potsdam Institute for Climate Impact Research, Potsdam, Germany, [9]Department of Physics and Astronomy, University of Leicester, Leicester, UK, [10]Department of Geodesy and Geoinformation, TU Wien, Vienna, Austria, [11]Inversion Lab, Hamburg, Germany, [12]Deutscher Wetterdienst, Offenbach, Germany, [13]Initiative Pro Klima, University of Hamburg, CliSAP/CEN, Hamburg, Germany, [14]Royal Belgian Institute for Space Aeronomy (BIRA-IASB), Brussels, Belgium, [15]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland

Abstract Assessing the inherent uncertainties in satellite data products is a challenging task. Different technical approaches have been developed in the Earth Observation (EO) communities to address the validation problem which results in a large variety of methods as well as terminology. This paper reviews state-of-the-art methods of satellite validation and documents their similarities and differences. First, the overall validation objectives and terminologies are specified, followed by a generic mathematical formulation of the validation problem. Metrics currently used as well as more advanced EO validation approaches are introduced thereafter. An outlook on the applicability and requirements of current EO validation approaches and targets is given.

1. Introduction

Errors in satellite data products are known unknowns. However, quantifying the quality of these products by decomposing the inherent uncertainty components can be a very challenging task. Various and conceptually different approaches have been developed in the Earth Observation (EO) communities to address the validation problem for satellite-based data products. This has led to the development of different methodological approaches as well as partly different nomenclature, skill scores, and terminology, all to answer the same question: How good is a particular data set?

Space agencies are tackling validation at the producer side for major ongoing global EO efforts.

Examples include the environmental data records being developed by the National Oceanographic and Atmospheric Administration (NOAA) and products or climate data records from the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) [*Yost*, 2016], the European Space Agency's (ESA) Climate Change Initiative (CCI) [*Hollmann et al.*, 2013; *Dorigo et al.*, 2017], and NASA's Earth system data records [*Justice et al.*, 2013]. At the user's side, product accuracy requirements are being formulated by international programs such as the Global Climate Observing System (GCOS) [*Bojinski et al.*, 2014]. There are efforts under way to increasingly standardize and make product validation procedures and results traceable to the data user community. These efforts range from the establishment of variable-specific validation protocols and frameworks by, e.g., the Land Product Validation subgroup of the Committee on Earth Observation Satellites (CEOS LPV) to the establishment of traceability chains being developed for the climate change services in the framework of the European Earth observation program Copernicus.

An overview of general validation practices for different levels of EO data is provided by *Zeng et al.* [2015] with a particular focus on the generation of long-term data records for Essential Climate Variables (ECV) [*Bojinski et al.*, 2014]. However, from these activities emerged an urgent need for a unification of uncertainty

terminology and validation metrics across producer communities representing all spheres (land, atmosphere, and ocean). Such a coordination will also be beneficial for the consistent formulation of accuracy requirements across products as developed internationally for example by GCOS (for climate) and for 13 other weather and climate-related application areas within the WMO rolling review of requirements (http://www.wmo.int/pages/prog/www/OSY/GOS-RRR.html).

This paper presents results from two workshops at the International Space Science Institute (ISSI) where experts from very different research fields met. During these workshops different and complementary strategies in quantifying the quality of satellite data products were identified and discussed. Currently, there is a lack of documentation of satellite validation strategies and methods across the different communities. This paper therefore reviews the existing state-of-the-art methods in satellite validation and documents their similarities and differences.

The major objectives of the present paper are the following: (1) to provide a review about the state of the art in the field of EO data validation with respect to the validation of satellite remote sensing-based data sets; (2) to document different EO validation techniques applied in rather different communities including their limitations and inherent assumptions; (3) to provide a summary of commonly applied terminologies and their usage within different research communities; and (4) to provide selected examples for EO data validation techniques.

Based on the inventory made and examples given, we will discuss the applicability of generic EO data accuracy requirements and the need for more clearly defined product requirement terms, e.g., those defined within World Meteorological Organization (WMO) requirements for essential climate variable data sets [*World Meteorological Organization (WMO)*, 2011].

Starting from a very generic description of the EO validation problem, we will first introduce the general components of an EO validation framework and the associated nomenclature (section 2), followed by a thorough description of the mathematical basis and basic metrics for validation, with a clear summary of their assumptions (section 3). Advanced EO validation approaches from different research communities will be introduced thereafter (section 4). A discussion on the applicability and limitations for different EO validation strategies is introduced (section 5), followed by the summary, conclusions, and recommendations.

Future satellite missions, satellite data product validation, and in situ data collection efforts are expected to benefit from this joint community effort by a clear definition of validation targets and terminology, raising awareness on existing validation practices and methods within different communities. This might also help to develop improved sampling designs and approaches for continuous collection of reference data as well as the definition of field validation campaigns.

2. Validation Framework

2.1. Underlying Concepts and Terminology

Validation and uncertainty assessment is a crucial requirement from the end user perspective of a satellite data product [*Otto et al.*, 2016]. The term *validation* can, however, mean many things. For the land product domain, *Justice et al.* [2000] defined validation as the process of evaluating by independent means the accuracy of satellite-derived land products and quantifying their uncertainties by analytical comparison with reference data.

Rodgers [2000] defined the purpose of validation as the confirmation that the theoretical characterization and error analysis actually represent the properties of the real data. In the metrology (i.e., "measurement science") community, validation is understood to be a verification against requirements which ensure that the data are adequate for an intended use [see *Joint Committee for Guides in Metrology (JCGM)*, 2012]. Verification, similar to evaluation here means confirmation, through the provision of objective evidence, that requirements have been fulfilled. In other words, validation is a specific case of verification, which takes into account the intended use of the data products. Clearly, this is a more demanding interpretation. For instance: not only does one test whether reported uncertainties are realistic, they also have to be small enough for the envisaged use case. Across the different Earth observation communities, a key component in any validation exercise is the consistency check (comparison) of the remote sensing data with reference measurements which are assumed to be representative of the truth, at least within their own reported uncertainties.

The EO and reference data are in an ideal case both linked to a stated metrological reference through an unbroken chain of calibrations or comparisons, each contributing to the stated measurement uncertainty. This is the so-called traceability of the measurements, as defined in the International Vocabulary of Metrology, the VIM [*JGCM*, 2012]. In practice, reference data can range from fully SI (Système International) traceable to loosely agreed community standards. Ground measurements that are fully characterized and traceable are often called fiducial reference measurements (FRM) [*Donlon et al.*, 2014]. Examples of visualized traceability chains for measurements of several land and atmosphere ECVs were produced in the EC FP7 project QA4ECV (http://www.qa4ecv.eu). Exemplary work to ensure the traceability of the GRUAN (GCOS Reference Upper Air Network) radiosondes was published by *Dirksen et al.* [2014], and similar work is ongoing, for instance, for the Network for Detection of Atmospheric Composition Change (NDACC) and for the Total Carbon Column Observing Network (TCCON).

The EO data are in practice rarely fully traceable, for instance, because fundamental calibrations done in the laboratory prelaunch cannot be repeated in space. Consequently, the comparison against reference measurements in a validation exercise is often the only way to link the EO data back to an agreed standard.

It is important to clarify here what exactly is understood by measurement uncertainty, as the terms "error" and "uncertainty" are often used interchangeably within the scientific community. The VIM [*JCGM*, 2012] defines the measurement uncertainty as a nonnegative parameter describing the dispersion of the quantity values attributed to a measurand. The measurement error on the other hand is the difference between the measured value and the true value, i.e., a single draw from the probability density function (PDF) determined by the measurement uncertainty. The measurement error can contain both a random and a systematic component. While the former averages out over multiple measurements, the latter does not. A summary of relevant terminology is provided in Table 1.

Uncertainties in the reference and EO measurements are derived from a consideration of the calibration chain in each system and the statistical properties of outputs of the measurement system (cf. the traceability chains discussed earlier). The Guide to the expression of Uncertainty in Measurement (GUM) [*JCGM*, 2008, 2009, 2011] prescribes how these component uncertainties are combined to give an overall uncertainty associated with measurements from each system, denoted in the current paper with u_x and u_y.

In the validation process (see section 2.2), also the uncertainty due to imperfect spatiotemporal collocation must be taken into account in the consistency check. This uncertainty is not related to an individual measurement but to the differences in sampling and smoothing properties of different measurement systems.

It must be noted that the current paper focuses on that part of validation that deals with the comparison with reference data. Those communities that follow the broader metrological definition of validation also include verification against user requirements not directly related to the measurement accuracy such as spatiotemporal coverage and resolution of the data set (e.g., *Keppens et al.* [2015], for the atmospheric domain).

2.2. The Validation Process

While the validation aim is in principle straightforward, the actual implementation represents an extensive process in which each individual step is subject to various assumptions and potentially requires user decisions which might make it a subjective approach.

Within most communities, detailed validation protocols have been established, tailored to the specific products and validation aims. Some examples can be found in the early work on SAGE II ozone profile validation using ozonesondes by *Cunnold et al.* [1989], the recent work on IASI radiance validation, from temperature and water vapor profiles, developed by *Calbet et al.* [2011, 2016], and the detailed processing model for the future QA/Validation service developed in QA4ECV [*Compernolle et al.*, 2016].

Figure 1 shows the generic structure of the comparison part within a validation process. In the following, we will describe the different steps, highlighting user decisions and assumptions. In particular, these assumptions are discussed which can cause additional uncertainties affecting the validation results.

2.2.1. Data

From an idealized perspective the input data *x* and *y* (e.g., satellite data and reference data) to the validation process would be traceable to SI reference standards (see section 2.1). In practice this is rarely the case, and the choice of reference data, in particular, is often a pragmatic decision. Typical considerations in this regard

Table 1. Definitions and Their Source for Terms Commonly Used in EO Validation[a]

Term	Definition	Source
Verification	Confirmation, through the provision of objective evidence, that specified requirements have been fulfilled	ISO:9000
Validation	(1) The process of assessing, by independent means, the quality of the data products derived from the system outputs	CEOS/ISO:19159
	(2) Verification, where the specified requirements are adequate for an intended use	VIM/ISO:99
evaluation	The process of judging something's quality	The Cambridge Dictionary
Quality	Degree to which a set of inherent characteristics of an object fulfills requirements	ISO:9000
Quality indicator	A means of providing a user of data or derived product with sufficient information to assess its suitability for a particular application	QA4EO
Accuracy	Closeness of agreement between a measured quantity value and a true quantity value of a measurand. Note that it is not a quantity and it is not given a numerical quantity value.	VIM/ISO:99, GUM
Precision	Closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions	VIM/ISO:99
Measurement error	Measured quantity value minus a reference quantity value	VIM def. 2.16
	difference of quantity value obtained by measurement and true value of the measurand	CEOS/ISO:19159
Random measurement error	Component of measurement error that in replicate measurements varies in an unpredictable manner	VIM def. 2.19
Systematic measurement error	Component of measurement error that in replicate measurements remains constant or varies in a predictable manner	VIM def. 2.17
Measurement uncertainty	Nonnegative parameter characterizing the dispersion of the quantity values being attributed to a measurand	VIM def. 2.26
Standard uncertainty	Measurement uncertainty expressed as a standard deviation	VIM/ISO:99
Measurement bias	Estimate of a systematic measurement error	VIM def. 2.18
Traceability	(Metrological traceability) Property of a measurement result relating the result to a stated metrological reference (free definition and not necessarily SI) through an unbroken chain of calibrations of a measuring system or comparisons, each contributing to the stated measurement uncertainty	VIM/ISO:99

Table 1. (continued)

Term	Definition	Source
Fiducial reference measurements (FRM)	Suite of independent, fully characterized, and traceable ground measurements that follow the guidelines outlined by the GEO/CEOS Quality Assurance framework for Earth Observation (QA4EO)	[*Donlon et al.*, 2014], ESA's SPPA division
Representativeness	The extent to which a set of measurements taken in a given space-time domain reflect the actual conditions in the same or different space-time domain	[*Nappo et al.*, 1982]
Temporal stability	(a) Ability of a data record to detect long-term trends, (b) property of a measuring instrument to provide similar measurements when the measurand remains constant in time	GCOS / VIM def.4.19

[a]These sources are, in particular, the ISO Quality Management Principles (http://www.iso.org/iso/pub100080.pdf), the International Vocabulary of Metrology (VIM) [*JGCM*, 2012], and associated Guide to the expression of Uncertainty in a Measurement (GUM) [*JCGM*, 2008], the Terms, Definitions, and Cal/Val Best Practices adopted by the Committee on Earth Observation Satellites (CEOS) (http://calvalportal.ceos.org), and the WMO Quality Management Framework (QMF) (http://www.bom.gov.au/wmo/quality_management.shtml).

include the following questions: (1) Do the data provide scientifically meaningful estimates of the investigated geophysical quantity? (2) Do these data sufficiently cover the potential parameter space? (3) Are the data expected to be accurate enough to be able to draw desired conclusions from the validation process? (4) Are the data publicly available and accessible?

These points are all critical in maintaining end-to-end traceability in the validation process. Currently, many of these considerations may not be adequately addressed and the choice of data is often based either on practicality or on what is considered to be best suited for the validation procedure. These decisions, therefore, affect the overall credibility of the validation results.

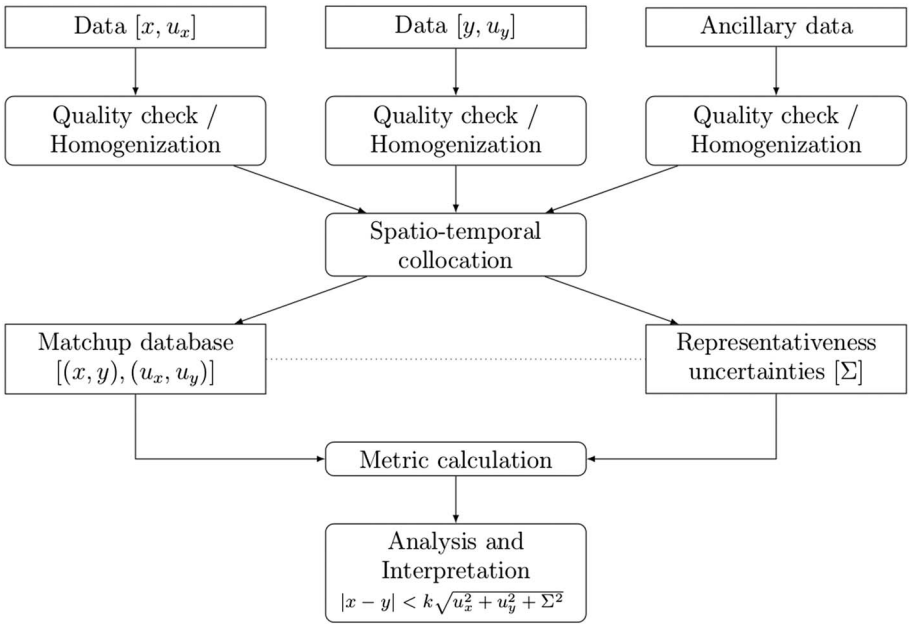

Figure 1. Schematic overview of the general validation process.

2.2.2. Quality Checking

Once a data set has been chosen, a quality assessment is required, and this is applicable also to any ancillary data. Many data providers include information on data quality within their data sets. Data quality information can be provided as simple binary flags (good/poor data quality) or gradated flags representing a number of different quality levels. Alternatively, data may be provided with quantitative uncertainty measures, and in some cases, these can also be matched to a quality level description. Thresholds are commonly applied to quality level descriptors or uncertainty information to determine which data to use in the validation process. In some cases, additional checks may be necessary such as checking the physical plausibility of a given measurement, visual inspection of the data, or tests for temporal consistency. In some communities, data may also be verified against a climatology to ensure that values lie within normal seasonal limits.

The decisions made here are generally application specific and may be based on the volume of data available. They may result in a reduction in spatiotemporal coverage of the available data and can affect the representativeness of the data set.

2.2.3. Spatiotemporal Collocation

Spatiotemporal collocation is one of the more challenging aspects of a validation exercise. The following concerns are to be addressed:

1. Collocated measurements should be close to each other relative to the spatiotemporal scale on which the variability of the geophysical field becomes comparable to the measurement uncertainties, especially when intermittent and highly variable parameters are considered (e.g., convective precipitation).
2. If possible, differences in spatiotemporal resolution (horizontal, vertical, and temporal) should be minimized.
3. The collocation criteria should take into account the need for sufficient collocated pairs for robust statistical analysis. This need is often at odds with the first concern, and a compromise must be made.

Broadly speaking, there are two categories of collocation methods: (1) those that keep the data on their original grids and select the closest matches and (2) those that use interpolation and aggregation techniques (e.g., regridding, resampling, and kriging) to bring both data sets onto the same grid and temporal scale.

The interpretation of "closest" in the former approach can lead to a wide variety of colocation criteria, from simple metrics based on physical separation (distance, time) to more advanced methods such as the use of backward trajectories or auxiliary data, providing additional guarantees that the geophysical field is measured under similar circumstances (e.g., potential vorticity constraints in the atmospheric domain). This can include also statistical tests for the representativeness of a data set based on, e.g., geostatistical analysis. *Roman et al.* [2009] provide an example for the characterization of the spatial representativeness of in situ, point-like observations for the validation of satellite surface albedo data products. Their approach is based on a semivariogram analysis.

Also, when regridding a data set, several options are available, ranging from linear interpolations to advanced upscaling and downscaling schemes that enforce mass/surface/energy conservation and harmonization of the actual measurement resolution. Note that the latter is not always equal/similar to the adopted sampling grid.

Whichever colocation approach is chosen, a remaining mismatch is almost inevitable, either due to differences in resolution and field of view, or even just a plain offset in measurement location, like, e.g., caused due to the satellite sounder having a sampling pattern that does not result in an exact overpass at the ground station. When the resulting differences cannot be corrected for in the comparison, an additional uncertainty term must be taken into account in the consistency check represented here by Σ.

This collocation uncertainty can be estimated with various methods such as triple collocation analysis [*Gruber et al.*, 2013], the use of structure functions or uncertainties derived from Observation System Simulation Experiments (OSSE), etc. Since these methods are so specific to particular applications, a more detailed explanation can be found in the particular examples shown in section 4.

2.2.4. Homogenization

In many cases, further homogenization between the two data sets is necessary before actual differences can be computed. For instance, unit and other representation conversions are often required, and these can introduce additional uncertainties, in particular, when ancillary data are used. For comparisons of measurements

Table 2. Projects and Activities Related to EO Data Validation and Available Resources

Project	Description	References
IPWG	Validation resources and server for precipitation data validation	http://www.isac.cnr.it/~ipwg/calval.html
GPM	Ground Validation server for precipitation	http://gpm-gv.gsfc.nasa.gov/
GAIA-CLIM (H2020)	Improving nonsatellite atmospheric reference data, supporting their use in satellite validation, and identifying gaps	http://www.gaia-clim.eu; The GAIA-CLIM Virtual Observatory; Deliverables D2.8, D3.2, D3.4, D6.2, and D6.7
QA4ECV (FP7)	Quality assurance for essential climate variables	http://www.qa4ecv.eu; online traceability chains; The QA4ECV Validation Server; D2.4 (DPM)
OLIVE	Evaluation and cross comparison of land variables	*Weiss et al.* [2014], http://calvalportal.ceos.org/web/olive
FIDUCEO (H2020)	Fidelity and uncertainty in climate data records from Earth Observations following metrological principles	http://www.fiduceo.eu
NORS	CAMS NDACC validation server	http://nors-server.aeronomie.be
MACC (now CAMS)	Copernicus Atmospheric Monitoring Service	D153.1: Atmospheric Service Validation Protocol V2
CEOS-LPV	Protocols for variable-specific validation and coordination of validation across land product products	https://lpvs.gsfc.nasa.gov/
ESA CCI	CCI Project guidelines	http://cci.esa.int/sites/default/files/ESA_CCI_Project_Guidlines_V1.pdf
ESA WACMOS-MED	Quality assessment of different satellite products (water and energy cycle) over the Mediterranean area	http://wacmosmed.estellus.fr
ESA GlobVapour	Quality assessment of satellite-based water vapor products and requirements baseline documentation	http://globvapour.info/
CM SAF (EUMETSAT)	Protocols for the validation of satellite data records, i.e., requirements review documents	http://www.cmsaf.eu/docs/
H SAF (EUMETSAT)	Product validation service for precipitation, soil moisture and snow products	http://hsaf.meteoam.it
GEWEX water vapor assessment	Overview of satellite, in situ, and ground-based water vapor data records, an archive of global long-term water vapor data records on common grid and period for intercomparison and collocated radiosonde data from two different multistation long-term radiosonde archive	http://gewex-vap.org/
GEWEX cloud assessment	Overview of satellite cloud data records and an archive of global cloud data records on common grid for intercomparison	http://climserv.ipsl.polytechnique.fr/gewexca/index.html
RFA	Overview of satellite and ground-based radiation data records on common grid for intercomparison	https://gewex-rfa.larc.nasa.gov/
NPROVS	NOAA Products Validation Server, colocating and comparing selected satellites, sondes and NWP on a daily basis	https://www.star.nesdis.noaa.gov/smcd/opdb/nprovs/

Table 3. Software Tools Developed for EO Data Validation and Geoscientific Data Quality Assessment

Name	Description	Programming Language	Reference
ESMValTool	Comprehensive tool for model evaluation and data cross comparison for different ECVs	python, NCL	http://esmvaltool.org, *Eyring et al.* [2016]; *Lauer et al.* [2017]
Pytesmo	Geospatial (soil moisture) time series validation toolbox	python	http://rs.geo.tuwien.ac.at/validation_tool/pytesmo/docs/index.html
Poets	Geospatial image resampling toolbox	python	http://rs.geo.tuwien.ac.at/poets/docs/introduction.html
HARP	Atmospheric validation toolbox	C, python	https://cdn.rawgit.com/stcorp/harp/master/doc/html/index.html
GeoVal	Geoscientific data analysis and plotting package	python	https://github.com/pygeo/geoval

that rely on retrieval methods such as optimal estimation, this step could include the harmonization in terms of resolution and prior contribution using the Averaging Kernels, for instance, following the recipe by *Rodgers and Connor* [2003].

2.2.5. Metric Calculation

The choice of metrics used within the validation process depends on the application and data available. A range of metrics that can be applied in the validation process is discussed in detail in section 3.4.

2.2.6. Analysis and Interpretation

Once the final metrics have been obtained, it needs to be judged if the results are compliant with the requirements. Following the definition of validation (see Table 1), this implies verification that the data set is suitable for a specific application. However, in many cases a single application does not exist and requirements may be numerous and, thus, validation targets would need to be defined, which could then be checked for compliance on an individual basis.

In its most fundamental form, the consistency check between the differences between two measurements and the reported measurement uncertainties can be written as [*Immler et al.*, 2010]

$$|x - y| < k\sqrt{u_x^2 + u_y^2 + \Sigma^2}, \tag{1}$$

where x and y are the reference and EO measurements, u_x and u_y their respective uncertainties, k the so-called coverage factor, and Σ the additional variance of the differences due to colocation mismatch, i.e., differences in representativeness of both measurements. The coverage factor allows the combined uncertainties to be scaled to a particular confidence level. Where $k = 1$, the combined uncertainty is consistent with 1 standard deviation. The value $k = 2$ is frequently used to give a confidence level of 95% (assuming a normal distribution of the combined uncertainty) [*Bell*, 2001].

2.3. Activities and Tools for Quality Assurance and EO Data Validation

A couple of research projects have been devoted to establish best practices for the production of traceable, quality-assured EO data products as well as practices for their validation. The deliverables of these projects and international activities are not necessarily published in the peer reviewed literature but represent nevertheless a rich set of valuable information. Table 2 therefore provides references to project activities related to the validation of EO data and the associated resources available. Related to these are several dedicated software packages that are being offered to the communities (see Table 3).

3. Mathematical Basis

3.1. Differences Between Data Sets

Let us denote a continuous geophysical field in space and time $\theta(t, r)$, where t and r denote time and space, respectively. A sampled discrete or (spatially and/or temporally) averaged measurement x of this field is then

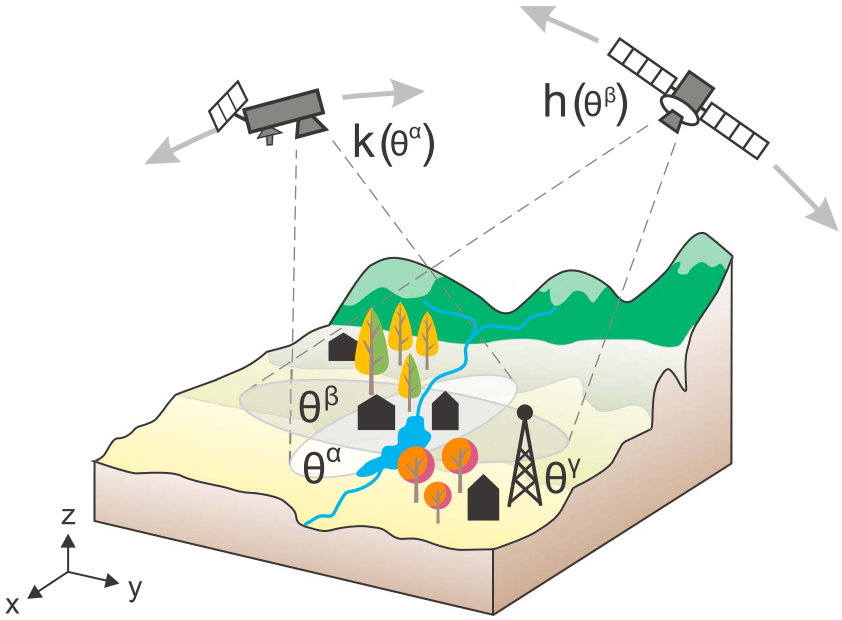

Figure 2. General sketch of the validation problem. A true, but unknown field of a geophysical variable θ is observed by different measurement systems on different spatial (and temporal) scales, denoted as α, β, and γ, using nonlinear mapping functions (h, k).

given by the random variable

$$\vec{x} = \left[h(\theta_1^\alpha) + u_{x,1}, h(\theta_2^\alpha) + u_{x,2}, \ldots, h(\theta_n^\alpha) + u_{x,n}\right]^T = h(\vec{\theta}^\alpha) + \vec{u}_x \tag{2}$$

where the subscript index denotes discrete instances in time and/or space associated with observations of x, the superscript multi-index α summarizes information on temporal and spatial resolution/averaging, and n is the number of samples in both space and time, h is a (nonlinear) mapping function, and u_x is the measurement error.

The quantity $\vec{x}$ might be either point-like (1-D), an area (2-D), or a volume (3-D). Examples might include point-like measurements of scalar surface properties, e.g., soil moisture in situ measurements (1-D), or sea surface temperature (2-D), or atmospheric moisture content in a volume (3-D) (Figure 2).

The mapping function h relates the true geophysical field to the actual measurement. This is typically required for any kind of measurement system, including in situ data. Note, that x might be autocorrelated in both, space and time. The error $\vec{e}_x$ is then defined as the difference between the sampled observation and the true, but unknown, state of the geophysical field,

$$\vec{e}_x = \vec{x} - \vec{\theta}^\alpha = h(\vec{\theta}^\alpha) + \vec{u}_x - \vec{\theta}^\alpha. \tag{3}$$

Since the truth commonly remains unknown, one typically has to use a second, (fiducial) reference data set containing measurements y of the same geophysical variable (e.g., in situ reference data) with spatiotemporal resolution β and supposedly better constrained uncertainty than x, which might be sampled in different ways at different times, yielding

$$\vec{e}_y = \vec{y} - \vec{\theta}^\beta = k(\vec{\theta}^\beta) + \vec{u}_y - \vec{\theta}^\beta. \tag{4}$$

The difference between the two samples is then obviously given as

$$\vec{\delta} = \vec{x} - \vec{y} = \left[h(\vec{\theta}^\alpha) + \vec{u}_x\right] - \left[k(\vec{\theta}^\beta) + \vec{u}_y\right] \tag{5}$$

As can be seen from equation (5), the difference between two measurements of θ is a function of both the individual measurement approaches, indicated by the measurement operators (h, k), and the associated measurement errors (u_x, u_y) as defined on different spatial and temporal scales (α, β). Provided that the

error properties of one of the data sets (e.g., fiducial reference data set) are well characterized, one may use equation (5) to obtain estimates for the uncertainties of the data set under validation.

Note that the errors in both measurements x and y include both systematic and random contributions. Both of these contributions may be characterized by the analysis of the differences between the measurements (equation (5)) as will be shown in the following. However, in addition to individual measurement uncertainties, the different spatiotemporal discretization/averaging induces an intrinsic difference

$$\vec{d}_\theta = \vec{\theta}^\alpha - \vec{\theta}^\beta = h^{-1}\left(\vec{x} - \vec{u}_x\right) - k^{-1}\left(\vec{y} - \vec{u}_y\right) \tag{6}$$

between the differently sampled representations of the common truth underlying both data sets. Please note that the superscript for h and k should not be interpreted as a strict inversion of the mapping functions in a mathematical sense, as this is often not possible due to the nonbijective nature of these functions. Spatiotemporal scale matching techniques have to be applied in order to correct for these differences before a meaningful interpretation of the differences between two data sets is possible (section 2.2.3).

3.2. Probability Density Functions

For validation purposes, we are not primarily interested in the differences for each individual pair of observations, but the underlying probability density function (PDF) of δ, f_δ, which is defined as

$$f_\delta(\delta) = f_\delta(h(\theta^\alpha) - k(\theta^\beta) + u_x - u_y) \tag{7}$$

and can be described as a function of the individual (conditional) PDFs of the four terms $h(\theta^\alpha)$, $k(\theta^\beta)$, u_x, and u_y. Many attempts to validate a satellite data product are based on the PDF of the differences δ.

However, the uncertainty assessment requires detailed understanding of the individual contributions to the PDF, i.e.,

1. systematic effects due to individual measurement devices and retrieval procedures (h, k), including large-scale systematic effects due to instrument calibration or harmonization;
2. random measurement and retrieval errors; and
3. differences in the representativeness of different data sets (again including random and systematic components).

If all these different components are fully understood, one can develop an error model M that allows to predict the PDF of δ as a function of the above mentioned factors,

$$f_\delta(\delta) = M(h(\theta^\alpha), k(\theta^\beta), u_x, u_y), \tag{8}$$

based upon certain assumptions on the functional form and (in)dependence of the four individual contributions.

3.3. Statistical Characteristics

The statistical properties of δ (i.e., the differences between the considered data sets) are commonly expressed in terms of the corresponding zeroth and first-order moments (mean, variance and covariance), and can be used to derive uncertainty estimates for these products.

The variance of δ, σ_δ^2, is defined as

$$\mathrm{Var}(\delta) = \sigma_\delta^2 = \int (\delta - \mu_\delta)^2 f_\delta(\delta)\mathrm{d}\,\delta = \int \delta^2 f_\delta(\delta)\mathrm{d}\delta - \mu_\delta^2 \tag{9}$$

where

$$\mu_\delta = \int \delta f_\delta(\delta)\mathrm{d}\delta = E[\delta] \tag{10}$$

with E being the expectation operator. By replacing the integrals by sums over individual pairwise differences, we obtain empirical estimators (denoted by the hat symbol) of the mean ($\hat{\mu}_\delta$) and variance ($\hat{\sigma}_\delta^2$), which can be further used to derive estimators for the bias and statistical spread of an observational data set with respect to a given reference.

Table 4. Summary of Common Metrics Definition As Well As Their Inherent Assumptions and Sensitivities[a]

Metric	Definition	Assumptions	Sensitive to
Measures of systematic differences			
Bias	$\hat{\mu}_x - \hat{\mu}_y$	S, G	sm, sr
Median difference	$p_x^{50} - p_y^{50}$	S	sm, sr
Measures of statistical spread			
RMSD	$\sqrt{E[(x-y)^2]}$	A, S, G, L, IE, O	sm, sr, rm, rr
cRMSD	$\sqrt{E[((x-\hat{\mu}_x)-(y-\hat{\mu}_y))^2]}$	A, S, G, L, IE, O	sm, sr, rm, rr
Triple collocation measures			
R^t	$\sqrt{\frac{\hat{\sigma}_{xy}\hat{\sigma}_{xz}}{\hat{\sigma}_x^2\hat{\sigma}_{yz}}}$	A, S, G, L, IE, O	rm, rr
RMSE	$\sqrt{\hat{\sigma}_x^2 - \frac{\hat{\sigma}_{xy}\hat{\sigma}_{xz}}{\hat{\sigma}_{yz}}}$	A, S, G, L, IE, O	rm, rr
SNR (dB)	$-10\ \log\left(\frac{\hat{\sigma}_x^2\hat{\sigma}_{yz}}{\hat{\sigma}_{xy}\hat{\sigma}_{xz}} - 1\right)$	A, S, G, L, IE, O	rm, rr
Statistical dependency measures			
Pearson's R	$\frac{\hat{\sigma}_{xy}}{\sqrt{\hat{\sigma}_x^2\hat{\sigma}_y^2}}$	A, S, G, L, IE, O	rm, rr
Spearman's ρ	$\frac{\hat{\sigma}_{r_{xy}}}{\sqrt{\hat{\sigma}_{r_x}^2\hat{\sigma}_{r_y}^2}}$	A, S, IE, O	rm, rr
Kendall's τ	$\frac{n_c - n_d}{n_0}$	A, S, IE, O	rm, rr
Mutual information, I	$\iint f_{x,y}(x,y)\log\left(\frac{f_{x,y}(x,y)}{f_x(x)f_y(y)}\right)\mathrm{d}x\mathrm{d}y$	A, S, IE, O	rm, rr
Temporal stability measures			
Absolute temporal stability		L	

[a]Abbreviations: S, stationarity; G, Gaussianity; L, linearity; IE, independence of error terms; O, orthogonality; A, additive error model; sm, systematic measurement uncertainties; sr, systematic representativeness differences; rm, random measurement uncertainties; rr, random representativeness differences.

In full analogy to the mean and variance of the difference δ, we can define the corresponding properties for the two observational data sets x and y, as well as their associated errors u_x and u_y. We can also define the covariance between the data sets as a measure for their statistical dependency as

$$\mathrm{Cov}(x,y) = \sigma_{xy} = \iint (x-\mu_x)(y-\mu_y)f_{x,y}(x,y)\mathrm{d}x\mathrm{d}y = \iint xyf_{x,y}(x,y)\mathrm{d}x\mathrm{d}y - \mu_x\mu_y. \tag{11}$$

An empirical estimator of the covariance between the two data sets ($\hat{\sigma}_{xy}$) can again be obtained by replacing the integral by the sum over collocated data pairs. Note that so far no assumptions on the statistical properties of the PDFs of the different variables have been made.

3.4. Validation Metrics

In the following, we provide a summary of some common validation metrics that are used in (or appear useful for) EO data. Note that these methods are generic in the sense that they can also be used for more general data intercomparison problems where none of the two considered data sets is a fiducial reference. A summary of the metrics and their assumptions is provided in Table 4.

3.4.1. Elementary Pairwise Metrics

The most commonly used pairwise methods are the correlation coefficient, the bias, and the root-mean-square deviation. Even though often considered complementary, these metrics are strongly interrelated, as will be shown below. Note that all pairwise methods are calculated between two data sets, both of which are subject to errors. However, for validation purposes, they are commonly calculated with respect to a fiducial reference data set with traceable uncertainty characteristics. Therefore, they are assumed to represent an estimate of the various uncertainty aspects of the data set under validation.

Correlation and other pairwise similarity characteristics. The correlation between two data sets is a normalized measure of statistical dependency between two variables. The most commonly used correlation metric is

the linear (Pearson) product-moment correlation coefficient, which describes first-order linear statistical dependency as the covariance between the data sets normalized by their respective standard deviations:

$$R(x,y) = \frac{\hat{\sigma}_{xy}}{\sqrt{\hat{\sigma}_x^2 \hat{\sigma}_y^2}}. \tag{12}$$

Characterizing linear dependencies is sufficient for many EO applications, mainly because metrics capable of capturing nonlinear relationships are often difficult to interpret. However, there may be situations where such nonlinear dependencies need to be taken into account.

The most commonly used nonlinear correlation metric is the nonparametric Spearman's rank correlation coefficient. Its definition is very similar to that of the Pearson correlation coefficient with the exception that it uses the variances and covariance of the ranks (r) of the data sets rather than the values directly:

$$\rho(x,y) = \frac{\hat{\sigma}_{r_{xy}}}{\sqrt{\hat{\sigma}_{r_x}^2 \hat{\sigma}_{r_y}^2}} \tag{13}$$

As a result, ρ quantifies how well the relationship between the data sets can be characterized by a monotonic function rather than by a linear function, as it is the case for R.

Another common nonparametric, nonlinear correlation metric is Kendall's Tau, which is mostly used as hypothesis test whether two variables are statistically dependent rather than for quantifying the quality of a regression. It is defined as

$$\tau(x,y) = \frac{n_c - n_d}{n_0} \tag{14}$$

where n_c is the number of data pairs with concordant ranks; n_d is the number of pairs with discordant ranks; and n_0 is the number of all possible data pairs.

While the previous three correlation characteristics essentially focus on monotonic relationships between the values of two data sets, mutual information (MI) [*Kraskov et al.*, 2004] — a concept originating from information theory — relieves this restriction by accounting for general statistical dependencies without any specific assumption of the corresponding functional form. It is defined based upon the joint and marginal probability distributions of the data sets:

$$I(x,y) = \int\int f_{x,y}(x,y) \log\left(\frac{f_{x,y}(x,y)}{f_x(x) f_y(y)}\right) \mathrm{d}x\mathrm{d}y. \tag{15}$$

In addition to the aforementioned metrics based upon linear and nonlinear dependency measures, for the classical correlation coefficients, associated significance levels are usually computed using a Student's t test in order to check whether a calculated correlation level really represents the statistical relationship between the variables under evaluation, or whether it was merely achieved by coincident data behavior which is increasingly likely for small sample sizes. Note that the suitability of the confidence limits provided by the t test requires the absence of serial correlations in the data; otherwise, the variance of the test statistics inflates and needs to be corrected analytically or numerically [*Hamed*, 2011, 2016].

It should also be noted that correlation metrics are not affected by systematic measurement or representativeness errors or, more precisely, by systematic errors within the orders of statistical dependency that are being characterized.

Pairwise difference-based metrics. The most common metric to quantify relative systematic differences between data sets is the bias, which is defined as the difference in the mean values and can identify a systematic overestimation or underestimation of one data set with respect to the other:

$$b(x,y) = \mu_x - \mu_y \approx \hat{\mu}_x - \hat{\mu}_y. \tag{16}$$

As opposed to the correlation coefficient, the bias is not sensitive to random measurement errors, since these have, per definition, a zero mean.

Another widely used metric for characterizing the pairwise agreement between two data sets is the root-mean-square deviation (RMSD), which is defined as

$$\mathrm{RMSD}(x,y) = \sqrt{E[(\hat{\mu}_\delta)^2]}. \tag{17}$$

Taken together, correlation coefficient, bias, and RMSD form the basis for the vast majority of validation studies. However, contrary to the way they are commonly perceived, these metrics are not fully independent and only partially complementary [*Taylor*, 2001]. As shown by *Gupta et al.* [2009], the RMSD can be decomposed into two bias related terms (a bias in mean and a bias in standard deviation) and a correlation-dependent term:

$$\mathrm{RMSD}(x,y) = \sqrt{(\hat{\mu}_x - \hat{\mu}_y)^2 + (\hat{\sigma}_x - \hat{\sigma}_y)^2 + 2\hat{\sigma}_x\hat{\sigma}_y(1 - R(x,y))}. \tag{18}$$

Often, the mean bias term is removed by subtracting the mean values of the data sets before calculating the RMSD. The resulting metric is referred to as centered RMSD (cRMSD). Notice that from this decomposition it becomes obvious that the cRMSD and the RMSD are sensitive to both systematic and random errors. *Taylor* [2001] has further shown that the cRMSD, the difference in the standard deviations of both data sets, and the Pearson correlation coefficient share a geometric relationship through the law of cosines, which allows to plot these three metrics on a two-dimensional diagram. This again demonstrates that these characteristics are not mutually independent.

3.4.2. Triple Collocation (TC)

As mentioned above, pairwise intercomparison metrics are sensitive to errors in both considered data sets. By using one fiducial reference data set with traceable uncertainty characteristics, one can estimate the uncertainties of a second data set under validation. However, this traceability is often hard to achieve and fiducial reference sites are for some geophysical parameters limited to relatively few sites globally. Therefore, so-called triple collocation (TC) analysis has become a standard method in various EO communities for uncertainty characterization if no fiducial reference data are available [*Stoffelen*, 1998; *McColl et al.*, 2014; *Loew and Schlenz*, 2011; *Gruber et al.*, 2016; *Su et al.*, 2014a; *Kinzel et al.*, 2016]. By introducing a third data set z into the comparison, this method allows to estimate the random uncertainties in all three data sets as well as their individual signal-to-noise ratio (SNR) properties. Various metrics can be derived from TC analysis, the most important ones being the correlation with respect to the unknown truth (R^t), the root-mean-square error (RMSE), and the logarithmic SNR expressed in decibel units [*Gruber et al.*, 2016]:

$$R^t(x) = \sqrt{\frac{\hat{\sigma}_{xy}\hat{\sigma}_{xz}}{\hat{\sigma}_x^2\hat{\sigma}_{yz}}} \tag{19}$$

$$\mathrm{RMSE}(x) = \sqrt{\hat{\sigma}_x^2 - \frac{\hat{\sigma}_{xy}\hat{\sigma}_{xz}}{\hat{\sigma}_{yz}}} \tag{20}$$

$$\mathrm{SNR}(x) = -10\,\log\left(\frac{\hat{\sigma}_x^2\hat{\sigma}_{yz}}{\hat{\sigma}_{xy}\hat{\sigma}_{xz}} - 1\right) \tag{21}$$

Recall that all three data sets contain errors, but none of them have to be a fiducial reference data set. R^t and RMSE estimate the same kind of measurement uncertainties as the Pearson correlation coefficient and the RMSD, respectively, whereby the latter are sensitive to uncertainties in both contributing data sets while the former are only sensitive to uncertainties in one individual data set. For the same reason we use—following common practice in literature—the RMSE for estimated uncertainties of single data sets and the RMSD for estimated spreads of differences between two data sets (which are driven by the uncertainties of two data sets). For a detailed derivation of TC-based estimators and their relation to pairwise metrics, we refer to *Gruber et al.* [2016].

3.4.3. Metric Assumptions

So far we have ignored that individual metrics rely on assumptions themselves. Assumptions related to the different metrics are therefore summarized in Table 4. It should be mentioned that certain assumptions need to be made for, e.g., the estimators in equation (19) to be unbiased, the most important ones being independence between the random measurement errors of the different data sets, independence between the random measurement errors and the state of the true signal, and the lack of nonlinearities in the data. Against common perception, these requirements are equally important for pairwise metrics, yet it is often easier to find just two data sets which meet those assumptions than it is to find three.

Notice, however, that some of these assumptions are more critical for some geophysical variables than for others. The assumption of independence between the random errors and the state of the variable, for example, is hardly ever questioned for soil moisture data sets, whereas precipitation data are often assumed to have such dependency.

3.4.4. Validation by PDF Intercomparison

One widely used assumption beyond the use of elementary validation metrics is that the PDFs of the different factors affecting f_δ all follow a normal distribution.

One way of circumventing problems due to deviations from this assumption is replacing quantities based on mean and variance by such based on medians and squared mean absolute differences, respectively. In more detail, quantile-based statistics can be used to further characterize differences between the PDFs of, e.g., x and y (like differences between their interquartile ranges).

While the basic validation metrics discussed above provide good indicators for the agreement between two data sets in case of Gaussian distributions, they might encounter systematic biases when the PDF of the data sets or their difference is non-Gaussian, especially in case of small sample sizes. Intercomparing the PDFs of the individual products (or characterizing the PDF of their differences) might therefore provide additional valuable information. This is particularly important if one is interested in the capability of a satellite data set to observe extremes of the geophysical variable investigated, i.e., when one is interested in the tails of distributions.

There are multiple complementary ways to compare the PDFs of two different data sets supposed to represent the same underlying variable. Beyond established statistical tests for the equality of the mean or median (commonly summarized as analysis of variance, ANOVA), there is a plethora of homogeneity tests which characterize the dissimilarity between the full PDFs. Corresponding quantifiers include the Kolmogorov-Smirnov statistics, Kullback-Leibler divergence, and many others [*Lehmann and Romano*, 2005; *Press et al.*, 1992].

3.4.5. Temporal Stability

The long-term stability of multidecadal satellite data records is a key requirement for their applicability for climate studies. The temporal stability β is defined either as a measure of the ability of a data record to detect long-term trends [*WMO*, 2011] or as the property of a measuring instrument or data set to provide consistent measurements of a geophysical variable, which is known to remain constant over time [*JCGM*, 2012].

Note that while the first definition is only based on a satellite time series itself, the second definition actually requires a priori knowledge about the measurand as reference data, assuming however that this is constant over time. An estimator for the stability is in both cases the temporal derivative, and as the reference is assumed to be constant in time ($\mathrm{d}y/\mathrm{d}t = 0$), this implies that both definitions are mathematically equivalent, as

$$\beta = \frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}(x - y) = \frac{\mathrm{d}x}{\mathrm{d}t} - \frac{\mathrm{d}y}{\mathrm{d}t} = \frac{\mathrm{d}x}{\mathrm{d}t} - 0. \tag{22}$$

As it is also possible that the reference measurements show also changes on longer time scales ($\mathrm{d}y/\mathrm{d}t \neq 0$) (e.g., change in surface solar irradiance and change in surface temperature), an additional definition for the temporal stability is commonly used, which defines the temporal stability as the change in the bias of a data set over time, namely

$$\beta = \frac{\mathrm{d}}{\mathrm{d}t}(x - y) = \frac{\mathrm{d}x}{\mathrm{d}t} - \frac{\mathrm{d}y}{\mathrm{d}t}, \tag{23}$$

which is obviously different from equation (22). While equations (22) and (23) provide measures for the absolute temporal stability, the relative temporal stability is also often used, which normalizes the absolute change with respect to some reference (e.g., mean of the time series). This is of particular interest if one is interested in relative changes of the observations that might be, e.g., caused by degradations of the observing system.

The ability to detect a temporal trend in an environmental data record and thus as a consequence to be able to quantify the temporal stability depends on the natural variability of the signal itself as well as the autocorrelation of the time series and its length [*Tiao et al.*, 1990; *Weatherhead et al.*, 1998; *Loew*, 2014].

Let us consider the time series of measurements of a geophysical variable as function of time t as follows [*Weatherhead et al.*, 1998]:

$$x(t) = c + \beta t + S_t + U_t + N_t, \tag{24}$$

where c is a constant value, β is the slope of a linear trend, S_t is a seasonal signal with predefined periodicity, and $N_t = \phi N_{t-1} + \varepsilon_t$ with $|\phi| < 1$ is assumed to be the data noise represented by a stationary autoregressive process of order 1 (AR1). ε_t are independent random variables with zero mean and variance σ_ε^2. It is noted that the variance of the white noise process is directly related to the variance of the variance of N_t by $\sigma_N^2 = \mathrm{Var}(N_t) = \sigma_\varepsilon^2/(1-\phi^2)$. The term $U_t = \Delta\Theta(t-T_0)$ represents a shift in the mean of the time series at time $t = T_0$, which might be related to, e.g., abrupt changes in the observation system, with Δ being the magnitude of the shift and Θ the Heaviside function defines as

$$\Theta(t - T_0) = \begin{cases} 0, t < T_0 \\ 1, t \geq T_0 \end{cases} \tag{25}$$

For a deseasonalized time series, equation (24) reduces to

$$x(t) = c + \beta t + U_t + N_t. \tag{26}$$

Assessing the temporal stability of a data set requires the following steps:

1. *Seasonality*. The removal of seasonality effects is optional but important, if the methods for assessing the significance of a trend from an autocorrelated time series shall be assessed [*Weatherhead et al.*, 1998].
2. *Breakpoint detection*. It needs to be checked if the time series contains abrupt changes, which might be an artifact of an inhomogeneous data record. Their detection is essential as they might introduce spurious trend estimates. Several methods have been developed and applied for the detection of abrupt changes, so-called breakpoints, within different EO communities [*Wang*, 2008a, 2008b; *Verbesselt et al.*, 2010; *Schröder et al.*, 2016a]. A recent overview of available methods to detect and remove breakpoints is given in *Venema et al.* [2012]. Further information is available at www.homogenisation.org and from the WMO task team on homogenization.
3. *Stability estimation*. There exist a variety of approaches to quantify the temporal stability of a data series. The method chosen should reflect the actual definition of temporal stability considered. A brief summary of different studies using different approaches is therefore given in the following. All methods are based on a statistical estimator of the temporal stability parameter, $\hat{\beta}$, which depends on the uncertainties of the investigated geophysical measurement and is therefore uncertain by itself. It is affected both by the SNR of the signal as well as its autocorrelation structure [*Weatherhead et al.*, 1998; *Zhang and Zwiers*, 2004; *Morin*, 2011]. Both need to be taken into account when estimating β. The uncertainties of $\hat{\beta}$ are, however, often neglected when comparing the estimated temporal stability against some user or satellite mission requirements [*WMO*, 2011].

A robust estimation of β and its significance needs to (a) take into account the autocorrelation structure of the data or noise in both, the reference measurements, and the measurements of the geophysical variable, (b) provide robust results in case of outliers in the time series, (c) take into account uncertainty information in individual data samples, if available [*Hubert et al.*, 2016; *Loew et al.*, 2016], and (d) be robust against data sparse periods.

An analytical approximation for the estimation of the minimum number of years required to detect a significant trend of a given magnitude is provided by *Weatherhead et al.* [1998]. It has been applied by *Mieruch et al.* [2014] to global time series of water vapor and by *Roman et al.* [2014] for time series of precipitable total water content of the atmosphere. The time needed to detect a specific trend depends, among other things, on the measurement uncertainty. *Roman et al.* [2016] shows that the measurement uncertainty is a function of the variable itself. A different approach was proposed by *Morin* [2011] who used a Monte Carlo method to detect the minimum detectable trend in global precipitation records. Instead of taking into account the autocorrelation structure of the data itself, they prewhiten the time series [*Zhang and Zwiers*, 2004] to obtain new time series that are not serially autocorrelated in time.

Damadeo et al. [2014] investigated the impact of nonuniform spatial and temporal sampling in ozone data sets and their impact on the long-term trend assessment. They found that long-term trend estimators can deviate by up to 10% when neglecting the temporal and spatial distribution of data gaps. They therefore proposed an alternative regression approach for the estimation of the long-term trend.

The temporal stability estimator $\hat{\beta}$ is typically obtained through a linear regression on equation (24). *Loew et al.* [2016] analyzed the temporal stability of the METEOSAT surface albedo data product [*Govaerts et al.*, 2008;

Loew and Govaerts, 2010] and showed that more robust estimates of β are obtained using a weighted least square (WLS) approach that takes into account the quantitative uncertainty information provided together with the EO data product.

We note that beyond the time-dependent bias ($b = x - y$) described above, changes in the sensitivity of measurement devices could also manifest in trends of statistical quantities beyond the mean of b, like variance, higher-order moments or, most generally, quantiles. In principle, other elementary characteristics like the RMSD can be estimated in a time-dependent fashion in full analogy to the bias by considering linear regression to window-wise RMSD estimates. Especially in the case of strongly non-Gaussian PDFs, a convenient way to describe changes in arbitrary quantiles of the distribution of (instantaneous) differences between the two considered data sets are (linear) quantile regression models [*Koenker*, 2005], which can be used to estimate (mean) rates of changes β_τ for any arbitrary quantile τ of the PDF of the differences.

The use of global analysis systems as an effective transfer medium to facilitate the comparison of diverse EO data sets (including Level 1 satellite data as well as in situ data), covering aspects including stability, is an active area of research. A more detailed summary of recent work is given in section 4.6.

4. Advanced Validation Approaches and Strategies

In section 3, the mathematical background of basic EO validation metrics has been detailed. In the following, a brief description of various more advanced EO validation approaches will be given. If applicable, respective examples for each method are provided in Appendix A for different application domains. The list of methods summarized in the following is not supposed to be conclusive, and alternative approaches exist as well.

4.1. Multiple TCA

As described in section 3.4.2, triple collocation analysis (TCA) allows for estimating random error variances of three collocated data sets of the same geophysical variable [*Stoffelen*, 1998]. Depending on the complexity of the used error model, the concept of TCA needs to be expanded, in case all independent random uncertainties cannot be quantified due to an underrepresentation of the system of equations. The multiple TCA (MTC) overcomes this challenge, as has been demonstrated by *Kinzel et al.* [2016]. By performing two TCAs simultaneously, which differ in the combination of satellite pixels and in situ point measurements, the fractional contributions of noise associated with each collocation procedure as well as the in situ measurement can be identified. This eventually allows for isolating instantaneous random retrieval uncertainties, which is beneficial when it comes to assessing the performance of a satellite data set. An example of such a random uncertainty decomposition is provided in section A1.2 for L2 near-surface specific humidities incorporated in the Hamburg Ocean and Atmosphere Parameters and Fluxes from Satellite (HOAPS) climatology [*Andersson et al.*, 2010; *Fennig et al.*, 2012].

4.2. Spectral Methods

Several studies in different communities have demonstrated approaches to apply spectral-based methods as validation techniques over a broad range of data sources and timescales.

4.2.1. Fourier-Based Approaches

One of these approaches finds expression in the standard Fourier transform, which resolves the frequencies underlying the signal of interest.

Yoo [2002], for example, presented a spectral technique for validating remotely sensed soil moisture (SM) retrieval algorithms. The author provided a mathematical description of the mean-square error (MSE) as an index of accuracy in estimating the ground measurement (point gauges) using spaceborne measurements. In this regard, it has been demonstrated that the (random) SM field can be transformed into the frequency domain by using its normalized spectral density function. The resulting expected MSE for a single visit depends on the time smoothing and the expected value of the design filter function. This expected MSE decreases with the number of measurements and the lagged correlation of the visiting interval. The method has been applied to the Little Washita watershed (USA) using a simple SM dynamics model [*Entekhabi and Rodriguez-Iturbe*, 1994]. The SM RMSE was shown to be very sensitive to the number of point gauges as well as the size of the sensor's field of view.

A Fourier-based error estimation method, the so-called spectral fitting approach introduced by *Su et al.* [2013], was investigated by *Su et al.* [2014b]. It detects characteristic features of erroneous SM estimates from EO data

in the frequency domain. Using a simplified linearized water balance model and an additive Gaussian white noise model, the stochastic random errors are estimated by comparing the spectral properties of the SM time series and the water balance model. Results indicate that higher retrieval error standard deviations go along with higher vegetation/leaf area index, rainfall, and SM. By contrast, error standard deviations are higher over dry areas for the AMSR-E sensor. For both AMSR-E and ASCAT sensors, the signal-to-noise ratios are low over dry areas, indicating a clear dependency of retrieval performance on SM content.

4.2.2. Wavelet-Based Approaches

A second approach is given by the wavelet transform (WT). In contrast to conventional Fourier analysis, wavelets are localized waveforms and functions in both time and scale. This enables the visualization of the time series' frequency content in the temporal and spatial domain concurrently.

WT has become a common analysis tool for geophysical applications, particularly when it comes to the investigation of spatial and temporal variability patterns. However, only few studies have utilized WT as a validation metric to date. Nevertheless, the broad range of application fields across communities highlights the potential of WT analysis for validation purposes. One of the first WT overview studies in the field of geophysics has been published by *Praveen and Foufoula-Georgiou* [1997].

A overview paper on wavelet methods with respect to atmospheric applications is *Domingues et al.* [2005]. *Elsayed* [2010] provided an overview from an oceanographic perspective.

A validation approach specifically based on WT has been presented in *Hosoda and Kawamura* [2004]. The authors applied WT to examine the "New Generation SST" (NGSST), a merged microwave-infrared SST product produced by *Guan and Kawamura* [2004] with respect to moored buoy SST in the vicinity of the Kuroshio Current. The region was chosen to assess whether NGSST is capable of correctly representing the observed strong seasonal SST gradients, oceanic eddies, and oceanic frontal disturbances. Results of the WT indicated that aliasing effects seem to be inherent to the cloud-free microwave SST, as suggested by a higher SST variance. Wavelet coherency analysis revealed that frequent cloud cover during winter and spring is the cause for an observed phase difference between both data sources. The cloud cover issue also targets the poor thermal representation of cold oceanic eddies and localized warm streamers.

A further validation example utilizing WT is given by *Stevenson et al.* [2010], who chose a wavelet-based probabilistic approach to test the performance of a set of climate model performances regarding El Niño – Southern Oscillation (ENSO) variability. Among others, the authors were able to predict the necessary length for the model simulations to yield robust statistics. It was concluded that 250 years of model simulation are sufficient to capture 90% of ENSO behavior. From an observational point of view, 240 years of SST measurements would be required, which is considerably longer than present in situ SST records.

A comprehensive overview regarding assumptions and limitations of WT in validation analysis is given in *Torrence and Compo* [1998].

4.3. Field Intercomparison and Functional Network Analysis

In section 3, we have already discussed various metrics that can be used for validating EO products with respect to a given reference data set with traceable uncertainty. In common cases, such data sets comprise spatial fields of time series of a certain observation of interest, but the intercomparison between observation and reference data set is commonly made pointwise (i.e., separately for each grid point), utilizing any of the previously discussed metrics and providing spatial fields of local bias, RMSD, correlation coefficients or any other characteristic of interest.

In the following, we address the problem of EO validation from a different viewpoint by making explicit use of the spatial structures of the data sets of interest. Specifically, we do not only consider pointwise statistical similarity between observation and reference data, but the spatiotemporal interdependence structure among the two spatial fields of interest. This interdependence can be characterized by correlation measures (linear Pearson correlation, Spearman rank-order correlation, Kendall's Tau) as well as nonlinear generalizations thereof (mutual information and its generalizations) and also other statistical similarity concepts that quantify the similarity of specific types of temporal variability patterns embedded in the individual time series obtained at each grid point (e.g., ordinal pattern-based mutual information [*Deza et al.*, 2013] or phase synchronization in case of oscillatory patterns dominating the variability of the signal [*Yamasaki et al.*, 2009]).

Having thus computed two matrices of pairwise similarity measures $\mathbf{S}^x$ and $\mathbf{S}^y$ for all pairs of grid points, referring to observation and reference data sets, respectively, we can proceed with intercomparing these matrices. In many cases, the values of all matrix entries will be bound (e.g., to the interval $[-1, 1]$ for correlation measures) or can be normalized accordingly. In such cases, useful metrics are provided by the mean, median, and maximum differences between the colocated matrix entries.

One inherent problem of this straightforward approach is the intrinsic noise level (comprising the different types of uncertainties described in section 2), which commonly reduces the statistical similarity between pairs of time series in comparison with the "ground truth." Consequently, the results of the aforementioned metrics may be biased. In order to suppress the effect of stochastic factors on the spatial patterns to be compared, dimensionality reduction (e.g., using empirical orthogonal function (EOF) analysis) is a widespread concept especially in climatological applications [*von Storch and Zwiers*, 2003]. In this case, the complete spatial correlation structure of a given field of observations is approximated by a few spatial modes with the strongest variability (which are forced to be orthonormal in case of EOF analysis but may also exhibit other types of statistical independence when more sophisticated methods are used). As a result, the problem of intercomparing the spatial covariability patterns in observation and reference data set is reduced to a comparison between the corresponding EOF patterns or their analogs provided by other decomposition methods.

A more recent approach to reduce the covariability patterns of spatiotemporal data sets to their "dynamical backbone" is based on concepts from complex network theory [*Tsonis and Roebber*, 2004; *Donges et al.*, 2009, 2015]. In this case, the similarity matrices $\mathbf{S}^x$ and $\mathbf{S}^y$ are reduced to binary matrices $\mathbf{A}^x$ and $\mathbf{A}^y$ encoding the pairs of time series with the strongest (or statistically most significant) statistical dependencies in both data sets [*Paluš et al.*, 2011; *Radebach et al.*, 2013]. Mathematically, this can be expressed as

$$A^x_{i,j} = \Theta\left(S^x_{i,j} - S^{x,*}_{i,j}\right) \tag{27}$$

(and the same for the reference data set $\vec{y}$), where $\Theta(\cdot)$ is again the Heaviside function, i and j indicate the matrix elements representing fixed grid points, and $S^{x,*}_{i,j}$ is either a global threshold value (taken the same for all i and j) for identifying, say, a given percentage of the strongest statistical similarities among all pairs of grid points, or an individual threshold for all pairs of i and j that corresponds to the same significance level of a certain statistical test for interdependence (e.g., a t test) [*Paluš et al.*, 2011]. If $\vec{x}$ is a climatological field variable, $\mathbf{A}^x$ is referred to as the adjacency matrix of the functional climate network representation of $\vec{x}$. Note that if $\mathbf{S}^x$ represents a correlation matrix of much the same similarity measure that can reveal both positive and negative interdependencies, one commonly uses the absolute value of all matrix entries for the definition of the adjacency matrix in equation (27) in order to account for both types (i.e., positive and negative) of relationship.

Since the resulting adjacency matrices for both fields are binary, they allow employment of associated statistics based on categorical data analysis for their analysis, including the Hamming distance [*Hamming*, 1950; *Radebach et al.*, 2013]

$$H^{x,y} = \frac{1}{N^2}\sum_{i,j=1}^{N}\left(1 - \left|A^x_{i,j} - A^y_{i,j}\right|\right) \tag{28}$$

as a global characteristic, and the Jaccard index (or matching index [*Donner et al.*, 2010])

$$J^{x,y}_i = \frac{\sum_{j=1}^{N} A^x_{i,j}A^y_{i,j}}{\sum_{j=1}^{N} A^x_{i,j} + \sum_{j=1}^{N} A^y_{i,j} - \sum_{j=1}^{N} A^x_{i,j}A^y_{i,j}} \tag{29}$$

or conceptually related characteristics as local counterparts. Beyond these simple characteristics based on a comparison of the neighborhoods of "nodes" (grid points) in the associated networks, one may further utilize other local or global network characteristics encoding multiple-point statistical interdependencies to intercompare the spatial covariability structures of the two fields of interest [*Donges et al.*, 2009; *Radebach et al.*, 2013].

Even though a corresponding investigation has not yet been performed in the context of EO validation, the introduction of concepts from multivariate statistics and complex network theory provides great potential for characterizing present day and future EO products, especially regarding the identification of possible

improvement potentials for future algorithms. A more detailed exploration of the associated prospects and challenges is, however, beyond the scope of the present work.

It has to be noted that the aforementioned techniques require the spatial grid and temporal sampling underlying the observation and reference data set to be the same. Otherwise, additional preprocessing may be necessary to unify at least the spatial positions of grid points in both data sets, whereas different sampling can be dealt with by making use of appropriate estimators for the chosen statistical interdependence measure (e.g., Gaussian kernel-based correlation estimates [*Rehfeld and Kurths*, 2014]).

4.4. Consistency Through Process Models

Process models reflect our understanding of the relevant processes governing the system under observation and of their interaction. In particular, they can ensure properties such as conservation of mass, energy, or momentum. Such models can be used in the validation context in various ways. For example, one can exploit their capability of providing relations among variables and domains in space and time. One aspect of such a relation is that it allows using the observation of a particular variable $\vec{x}$ defined over a given domain in space and time, to obtain information about another variable $\vec{y}$ defined over another domain in space and time that we wish to validate (indirect validation). We may express such a relation in algebraic form as follows:

$$0 = R(\vec{x}, \vec{y}, X) \tag{30}$$

In the above expression, we have also included a vector X which denotes a set of further uncertain quantities (control variables) that determine the behavior of the model. In a dynamical model, this is typically some combination of initial and boundary conditions and process parameters, i.e., constants in the model equations. For details on the nomination of the control variables, see, e.g., *Kaminski et al.* [2012a] or *Rayner et al.* [2016].

In cases where the model can be formulated such that $\vec{x}$ can be expressed as a function of $\vec{y}$ and X, we can use the observation of $\vec{x}$ for validation of $\vec{y}$. For this purpose, we employ the model to propagate the joint PDF of $\vec{y}$ and X to a PDF of $\vec{x}$. If we neglect the uncertainty in X, we are likely to produce a PDF in $\vec{x}$ with too small spread and wrong center. An example for this approach to validation from the land community is provided in section A5 which describes the Two-Stream-Inversion Package (JRC-TIP) [*Pinty et al.*, 2007]. The variable $\vec{x}$ for which observations are available is the direct transmission through the canopy. In the example, the state of the vegetation is described by a set of (partly spectrally dependent) state variables, for example, the canopy's single-scattering albedo, leaf area index, or the reflectance of the soil background. All of these variables are simultaneously retrieved (including a formal uncertainty propagation) from a broadband albedo input. If our goal is to validate the retrieval of one of the state variables, say, the single-scattering albedo, then in the above formalism, that variable would take the role of $\vec{y}$, and the remaining state variables would compose X.

Often we are lacking a direct functional dependency of $\vec{x}$ on $\vec{y}$, but we can extract both from a model simulation, typically using suitable observational operators [*Kaminski and Mathieu*, 2017]. In such situation, we apply a two-step procedure. The first step (inversion step) derives a PDF of X that is consistent with three pieces of information: the PDF of $\vec{y}$, the PDF of the model (reflecting the residual model error, i.e., the error that cannot be attributed to inaccuracies of X), and a PDF of X that quantifies any available prior information on X, i.e., describes our state of information on X before the inverse step [*Kaminski et al.*, 2012a; *Rayner et al.*, 2016]. In a second step, the derived PDF of X (termed posterior PDF) is used together with the PDF of the model to derive a PDF for $\vec{x}$, which can be compared with the PDF that characterizes the observation of $\vec{x}$. An example from the above mentioned JRC-TIP is the situation where the direct transmission observation is not used to validate a retrieved state variable but another radiant flux simulated from the state variables, e.g., the fraction of photosynthetically active radiation (FAPAR), which in the formalism takes the role of $\vec{y}$. In a joint retrieval system like JRC-TIP, the nomination of a particular variable to be validated is, however, conceptually not required, and we only made it for didactic purposes. We should consider the validation to address all variables that are jointly retrieved, i.e., all state variables and all radiant fluxes derived from the state variables.

Another example for the above formalism is provided by the Carbon Cycle Data Assimilation System (CCDAS) [*Rayner et al.*, 2005]. The system is built around the terrestrial biosphere model BETHY [*Knorr*, 2000] and simulates, among other variables, fluxes of carbon and water between land and atmosphere. The CCDAS control vector X is composed of the model's process parameters plus initial and (potentially also) boundary conditions. As with JRC-TIP, we consider the validation to address the entire control vector and all simulated variables, in particular the above mentioned fluxes. *Knorr et al.* [2010] present an example, where a FAPAR

product [*Gobron et al.*, 2007] is simultaneously assimilated over seven sites. FAPAR is also simulated over an extra site, and the FAPAR product over that site (taking the role of $\vec{x}$) is used for validation. *Kaminski et al.* [2012b] present the joint assimilation of the same FAPAR product and flask samples of atmospheric carbon dioxide in a CCDAS setup at global scale. For validation they employ flask samples at extra sites not used for assimilation. *Scholze et al.* [2016] assimilate a soil moisture product [*Kerr et al.*, 2010] at global scale together with flask samples of atmospheric carbon dioxide. A similar approach to validation is used in the Earth Observation-Land Data Assimilation System (EO-LDAS) [*Lewis et al.*, 2012].

A further important step toward uncertainty assessment is provided by *Adams et al.* [2016], who have set up a "virtual laboratory" around a comprehensive Monte Carlo ray tracing model of the radiative transfer in the canopy-soil system. In their virtual laboratory they simulated field observations of surface albedo and assessed the accuracy of these measurements as a function of sensor location.

4.5. Indirect Validation

Indirect validation approaches map the observational data product to a secondary variable, which is then compared against reference data of the same geophysical variable. The performance of the secondary variable (in comparison with reference data) provides information on the quality of the observational data product (i.e., assuming that the inverse model has a negligible effect on the understanding of the quality of the observational data product).

A first study was carried out by *Crow et al.* [2010] who developed a data assimilation technique for evaluating satellite soil moisture data through rain gauge observations. Similarly, *Tuttle and Salvucci* [2014] assumed that under statistically stationary conditions, precipitation conditionally averaged according to soil moisture results in a sigmoidal curve and errors in soil moisture data degrade this relationship. The lower the agreement with the sigmoidal relationship, the lower the quality of the soil moisture product. A more recent and more direct example is the SM2RAIN algorithm developed by *Brocca et al.* [2013, 2014]. Satellite-based soil moisture estimates are used in combination with a simple water balance model to directly obtain estimates of the precipitation dynamics. The new precipitation data set [e.g., *Brocca et al.*, 2014] can then be validated against independent precipitation measurements.

In recent years, for estimating and correcting rainfall, similar studies were published in which the secondary variable included land surface temperature [*Wanders et al.*, 2015], Ku-band backscattering coefficient [*Turk et al.*, 2015], microwave surface emissivities [*Birman et al.*, 2015], and discharge [*Kirchner*, 2009; *Herrnegger et al.*, 2015]. Similarly, *Tian et al.* [2014] considered snow water equivalent observations for estimating snowfall at high latitudes. All these approaches can be exploited for inverse validation of satellite data, by benefiting from the larger availability of ground observation for the secondary variable (e.g., precipitation versus soil moisture). Further details are provided in section A4.

4.6. The Role of Data Assimilation for EO Validation

Satellite data plays a key role in current data assimilation systems within numerical weather prediction (NWP) models. The direct assimilation of passive microwave and infrared radiance data has had a particularly large beneficial impact on analysis and forecast quality [*Joo et al.*, 2013] over the last two decades. As the range of satellites exploited for this application grows, there is an ongoing requirement to validate new additions. Key challenges include the lack of traceable radiometric measurements from on-orbit satellite radiometers and a lack of traceable ground truth observations against which to validate the satellite observations.

Various strategies have been used (with some success) in recent years for the validation of satellite sounding observations, including comparison with collocated airborne radiometers flown as part of dedicated Cal/Val campaigns [*Newman et al.*, 2012] and collocated radiosondes [e.g., *Calbet et al.*, 2011, 2016].

In recent years, the use of simulated radiances (or brightness temperatures), generated from short range forecast fields interpolated in space and time to the location of satellite observations and projected into observation space by means of an observation operator, have proved to be effective in identifying biases in satellite radiances [e.g., *Bell*, 2015; *Lu and Bell*, 2014], pointing to the need for improved prelaunch characterizations of satellite radiometers. Such efforts further enable the development of bias models to reduce uncertainties in satellite data. These differences (observation-model) are routinely produced as part of the data assimilation process. The particular strength of these approaches lies in the complete global spatial coverage provided by the NWP models, coupled with the continuous coverage in time.

Quantitatively, uncertainties in model forecast fields projected into observation space (top of atmosphere brightness temperatures) are in the range of 0.05 K–0.20 K for radiances that are sensitive to temperature in the midtroposphere and lower stratosphere. For radiances sensitive to middle upper tropospheric humidity, NWP model uncertainties are around 1–2 K. For infrared and microwave radiances that are sensitive to ocean and land surfaces, the errors are around several Kelvin.

The aspiration of achieving a true validation of these satellite measurements is primarily dependent on progress in the development of traceable radiometric standards to support prelaunch characterization of satellite radiometers, as well as on-orbit calibration (to enable the determination of robust uncertainties in the satellite measurements). The development of methods to establish the uncertainties in NWP model fields through, for example, the use of reference measurements can be a third, valuable pathway to assess the quality of the satellite data.

There are a number of ongoing initiatives aimed at addressing these aims (e.g., GAIA-CLIM, CM SAF and efforts at UKMO, and ECMWF). These applications are described in more detail in *Bell* [2015], *Kobayashi et al.* [2017], and at www.gaia-clim.eu.

4.7. Assessing Uncertainty From Scale Mismatch

In the atmospheric domain, when measuring temperature, water vapor, or trace gases concentration, reference measurements are usually so sparsely distributed in space and time with respect to the broader resolution space-based measurements that it is difficult to find tight collocation criteria, upscaling methods, or even an estimation of the collocation uncertainty involved in the comparison. Moreover, atmospheric small scale features, i.e., scales of 10 km or smaller, have a stochastic behavior due to the atmosphere's turbulent nature at these scales. These facts make it impossible to precisely upscale just one reference measurement made at one physical location to satellite observations, which usually encloses an area covering only a few square kilometers.

Several strategies have been devised to overcome this issue, which can be summarized as follows:

1. Taking more than one reference point measurement. For example, by using two consecutive radiosonde measurements, separated by ~1 h, it is possible to make a time interpolation [*Tobin et al.*, 2006], providing an atmospheric state best estimate which has proven to be well collocated with satellite observations giving a nearly negligible collocation uncertainty [*Calbet et al.*, 2011]. It is possible to envisage other measurement strategies, involving, for example, ground-based remote sensing equipment such as lidar systems, which can achieve a similar goal.
2. Estimate the collocation errors (and resulting uncertainties) when comparing single-reference measurements with satellite data using, e.g., Numerical Weather Prediction (NWP) model fields. This has been applied with some success for ozone [*Verhoelst et al.*, 2015], see also section A3. This strategy has proven unattainable for other variables that have a much smaller variability scale such as temperature or water vapor.
3. When only a single-reference measurement is made, for example, a radiosonde measurement compared to a satellite-based one, the collocation uncertainties become nonnegligible [*Calbet et al.*, 2016]. The reason for this is the turbulent nature of the small-scale variability of temperature and water vapor. This can be seen by comparing the differences in temperature or water vapor between two consecutive sonde launches, usually launched about 1 h apart and known as dual sondes. The dependency of this difference with respect to distance is what is known in turbulent theory as structure function. The typical signatures of turbulence (inertial and energy injection ranges) can be derived from in situ measurements like radiosondes. The inertial range is derived from Kolmogorov's theory of turbulence. The fact that the atmosphere is highly turbulent at the scales of satellite observation make it also impossible to predict the real values of the parameters when using NWP model fields as auxiliary data. The only option left is to estimate the collocation uncertainty. This can be achieved either experimentally, with past data or using turbulence fields from an NWP model. Both of these methods are currently being investigated by various research groups.

4.8. Validation of Retrieval Uncertainties

In principle, most EO retrieval schemes allow a quantification of the uncertainty on the retrieved quantity, either within the retrieval system (see, e.g. *Rodgers* [2000, chapter 3] for the case of optimal estimation) or with statistical methods such as bootstrapping or Monte Carlo simulations, although in practice few EO retrieval

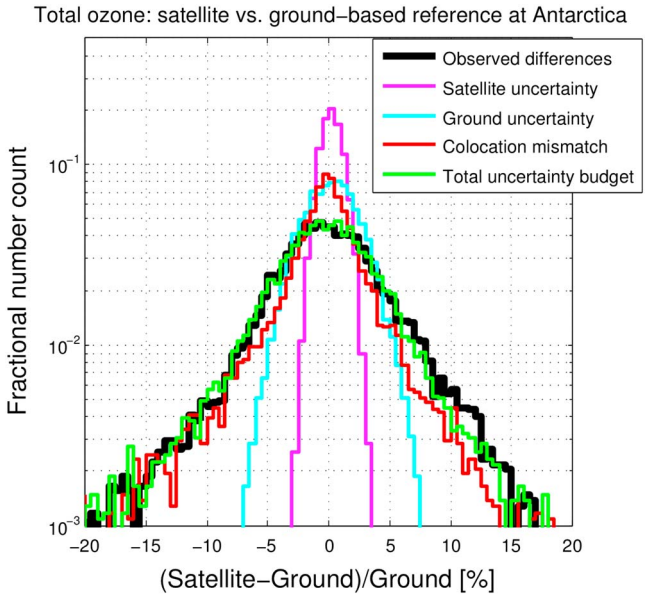

Figure 3. Comparison between (1) the PDF of the differences between satellite (GOME-2/MetOp-A) and ground-based reference (NDACC SAOZ) measurements of the total ozone column at Antarctica, and (2) the different known contributions to the error budget: measurement uncertainties (satellite and ground) and collocation mismatch (estimated with the OSSSMOSE system, *Verhoelst et al.*, 2015]), see also Appendix A3]. The good match between the PDF of the actual differences and the total of the different uncertainty terms instills confidence in the reported uncertainties.

algorithms of Earth surface properties implement this. In the atmospheric domain, this approach is already common practice [see, e.g., *Rodgers*, 1976; *Cunnold et al.*, 1989; *Boersma et al.*, 2004], with retrieved uncertainties that now have been provided for some products in other domains, e.g., for sea surface temperature [*Bulgin et al.*, 2016], lake and land surface temperature [*Ghent et al.*, 2016], and a set of further land surface variables including background albedo, effective LAI, and FAPAR [*Pinty et al.*, 2007] (see also section 4.4). Estimating the uncertainty in the observation within the retrieval maintains the independence of this estimate from other data sets (e.g., in situ), and enables these uncertainties to be validated in addition to the EO measurement, providing confidence in both the observation and its uncertainty.

These retrieval uncertainties, here taken to mean the entire measurement uncertainty as estimated by the retrieval system, can either be validated in the classical validation scheme outlined in section 2.2 or they can be verified with so-called self-collocations, although the latter approach usually covers only the random component of the uncertainty.

In the classical validation scheme, the retrieval (measurement) uncertainties are part of the fundamental consistency check, i.e., equation (1). A more advanced test, based on the same EO versus reference data comparisons, verifies whether the PDF of the differences is compatible with the PDF that combines (1) the retrieval uncertainty of the EO data, (2) the uncertainty on the (in situ) reference data, and (3) the uncertainty due to colocation mismatch or representativeness errors. This is illustrated in Figure 3, which shows an application for total ozone column validation. Here measurement uncertainties are available for both the EO data (magenta) and the ground-based reference data (cyan), but they cannot account for the PDF of the differences between satellite and ground-based instrument (black). Taking into account the collocation uncertainties (Σ, in red), here estimated with a chemical transport model, closes the uncertainty budget (green matches black).

Note that this is an implementation of principles already introduced in section 3.4.4. In some cases, (3) is difficult to quantify due to the typically large variability in surface characterization across the domain of an EO observation (e.g., land surface temperature retrievals).

The self-collocation technique relies on the assumption that the spread in the differences between self-collocations in a region of low intrinsic variability in the target field, should tend toward the measurement uncertainty as the collocation criteria are tightened. Hence, in practice, the spread is quantified for a range of collocation criteria that still provide meaningful numbers of collocated pairs, and the reported (random) measurement uncertainties are compared against the asymptotic behavior of that curve, see, for instance, *Sofieva et al.* [2014] and *Laeng et al.* [2015].

5. Conclusions and Recommendations

This review, and in particular sections 3 and 4, illustrates the wealth of different validation strategies and techniques developed across different EO communities. While each approach has its own assumptions and

limitations (see, e.g., Table 4), several common challenges can be identified, such as the need for fiducial reference measurements, the issue of spatiotemporal scale mismatch, the formulation of accurate requirements, the rigorous treatment of uncertainties, and the use of an appropriate, unambiguous, nomenclature. In this review, we have given particular attention to issues of nomenclature and scale mismatch and to novel techniques that optimize the use of the limited amount of reference data usually available. However, this also highlighted the disconnect between the detailed output of these advanced validation methodologies and the often simplistic and/or ambiguous user requirements against which the validation results need to be checked. For instance, requirements such as those from GCOS [*WMO*, 2011; *Bojinski et al.*, 2014] from space agencies, or from other service providers, typically call for scalar metrics like RMSD, correlation, or anomaly correlation, while state-of-the-art validation methods address the entire PDF.

It is recognized that these requirements are important as such to provide a target for the development of an observing system and thus to steer the process of mission design, system design, or algorithm development. Nevertheless, it is a problem that the used metrics are often not well defined, their origin is not traceable, and their applicability can be cumbersome.

The ultimate goal of a validation exercise is to assess whether a data set is compliant with predefined benchmarks (requirements) that quantify whether a data set is suitable for a particular purpose. It is therefore essential that the metric itself is not ambiguous and allows to fully capture the data quality. The following prerequisites therefore apply for a proper definition of requirements to be used as such validation targets or benchmarks:

1. A clear specification of the spatial and temporal domains for each metric is essential.
2. The mathematical details of the metric calculations have to be provided.
3. It needs to be specified under which conditions a data set would be considered to fulfill the specified requirements.
4. Clear separation between systematic and random error components is required.

Recommendations from the ISSI workshops this paper originates from are therefore as follows:

Refine user requirements. Current definitions of user requirements are very often based on scalar metrics. The motivation for these scalar metrics as well as their actual definition needs to be much more traceable. Thus, more traceability and documentation on the definition and origin of these requirements is needed. A reasonable start was made within CM SAF, e.g., for the requirements of the Hamburg Ocean Atmosphere Parameters and Fluxes from Satellite data (HOAPS).

Beyond scalar metrics. Advanced validation methods like those introduced in section 4 should gain widespread use. These require a thorough understanding of the error characteristics of both the satellite and the reference data. A decomposition of different error components can be achieved using these advanced methods.

Collection of reference data. Traceable approaches for the collection of reference data for EO validation are needed. The standards developed by the CEOS working group on calibration and validation define good standards already, but they are not necessarily available for all ECVs at the moment. Exemplary are the efforts by the GRUAN community, as summarized by *Bodeker et al.* [2016].

Traceable data production. Traceable data production chains are required that allow to trace back the method used for the production including full traceability of the satellite and ancillary data used, including their uncertainties. Quality assurance (QA) frameworks like, e.g., those developed in the FP7 QA4ECV project and further applied in the Copernicus Climate Change Service (C3S) are good examples here.

Sustained fiducial references. The sustained availability of fiducial reference data is not ensured for any of the communities working with EO data. Validation with fiducial references covering the spectrum of the natural variability of a geophysical variable (e.g., seasonal variations, different climate zones) is considered essential for all validation approaches. Collection and long-term maintenance of, e.g., in situ measurement networks, should become more a focus of the agencies and service providers of satellite data products. A change in the programmatic paradigm might be required for this.

Validation tools. Traceable and open source-based tools for the validation of EO data currently do not exist. More effort should be devoted in the research community toward the joint development of such tools as well as the intercomparison of already existing tools for satellite data validation.

While this paper is expected to serve in particular as a reference for current practices and methods in EO data validation across different research communities, it is also a plea for more traceability throughout the entire process for validating EO-based data products.

Appendix A: Selected Advanced Validation Examples

Examples of different validation approaches from different EO communities will be presented in the following, to illustrate the large range of validation strategies employed.

A1. Triple Collocation-Based Approaches

A1.1. Example of TC Error Decomposition for Soil Moisture

As shown by *Stoffelen* [1998] and *Gruber et al.* [2016], TC analyses actually estimate the variance of the true signal which is observed by the three considered data sets, biased with their respective systematic error and then obtain an estimate of the random uncertainties by subtracting these estimated biased signal components from the total variance of the respective data sets. Formally, this can be written using the linear error model:

$$x = \alpha_x + \beta_x \theta + \varepsilon_x \tag{A1}$$

where α_i and β_i denote additive and multiplicative systematic errors of data set i, respectively; ε_i denote random errors in data set i; and θ is the true state of the geophysical variable. The same error model applies to the other data sets in the triplet y and z. If the three data sets have uncorrelated errors and their errors are furthermore uncorrelated with the geophysical variable, then the signal variance (biased with the systematic error of data set x) can be estimated as

$$\beta_x^2 \hat{\sigma}_\theta^2 = \frac{\hat{\sigma}_{xy}\hat{\sigma}_{xz}}{\hat{\sigma}_{yz}} \tag{A2}$$

and the random uncertainties can further be estimated by simply subtracting this biased signal variance from the total variance of the data set:

$$\hat{\sigma}_{\varepsilon_x}^2 = \hat{\sigma}_x^2 - \beta_x^2 \hat{\sigma}_\theta^2 = \hat{\sigma}_x^2 - \frac{\hat{\sigma}_{xy}\hat{\sigma}_{xz}}{\hat{\sigma}_{yz}} \tag{A3}$$

Finally, one can estimate the SNR properties of the data set as the ratio between the estimated biased signal variance and the estimated random uncertainties [*Gruber et al.*, 2016]:

$$\mathrm{SNR}(x) = 10\, log\left(\frac{\beta_x^2 \hat{\sigma}_\theta^2}{\hat{\sigma}_{\varepsilon_x}^2}\right) = -10\, \log\left(\frac{\hat{\sigma}_x^2 \hat{\sigma}_{yz}}{\hat{\sigma}_{xy}\hat{\sigma}_{xz}} - 1\right) \tag{A4}$$

Notice that the SNR is expressed in decibel units in order to linearize its nonlinear behavior for an easier interpretability. Obviously, the above equations can also be rearranged to estimate the same parameters also for data sets y and z.

Separating the measurements into their signal and error components and relating them to each other allows for a meaningful intercomparison of the "quality" of the data sets and also for understanding whether the quality, for example, of retrievals from a certain satellite is driven by a low noise of the sensor or rather by a very strong observed signal. An example of such a decomposition is shown in Figure A1.

A1.2. Example of MTC Error Decomposition for Near-Surface Specific Humidity

Typically, satellite data validation by means of collocation analysis requires the use of ground reference data. Resulting RMSD (E_{sum}) cannot be completely assigned to the satellite retrieval (E_{tot}), as spurious contributions of in situ measurement noise (E_{ins}) and random collocation uncertainty (E_c) might substantially contribute to the overall random uncertainty. In general, triple collocation analysis allows for estimating the magnitude of the stochastic uncertainties in three data sets (see section 3.4.2). However, further decomposition of uncertainties is not possible, because the system of equations is underrepresented with respect to the number of unknowns.

In order to further enhance the uncertainty decomposition, and with that the actual retrieval uncertainty estimation, *Kinzel et al.* [2016] developed the so called multiple triple collocation (MTC) method. This novel

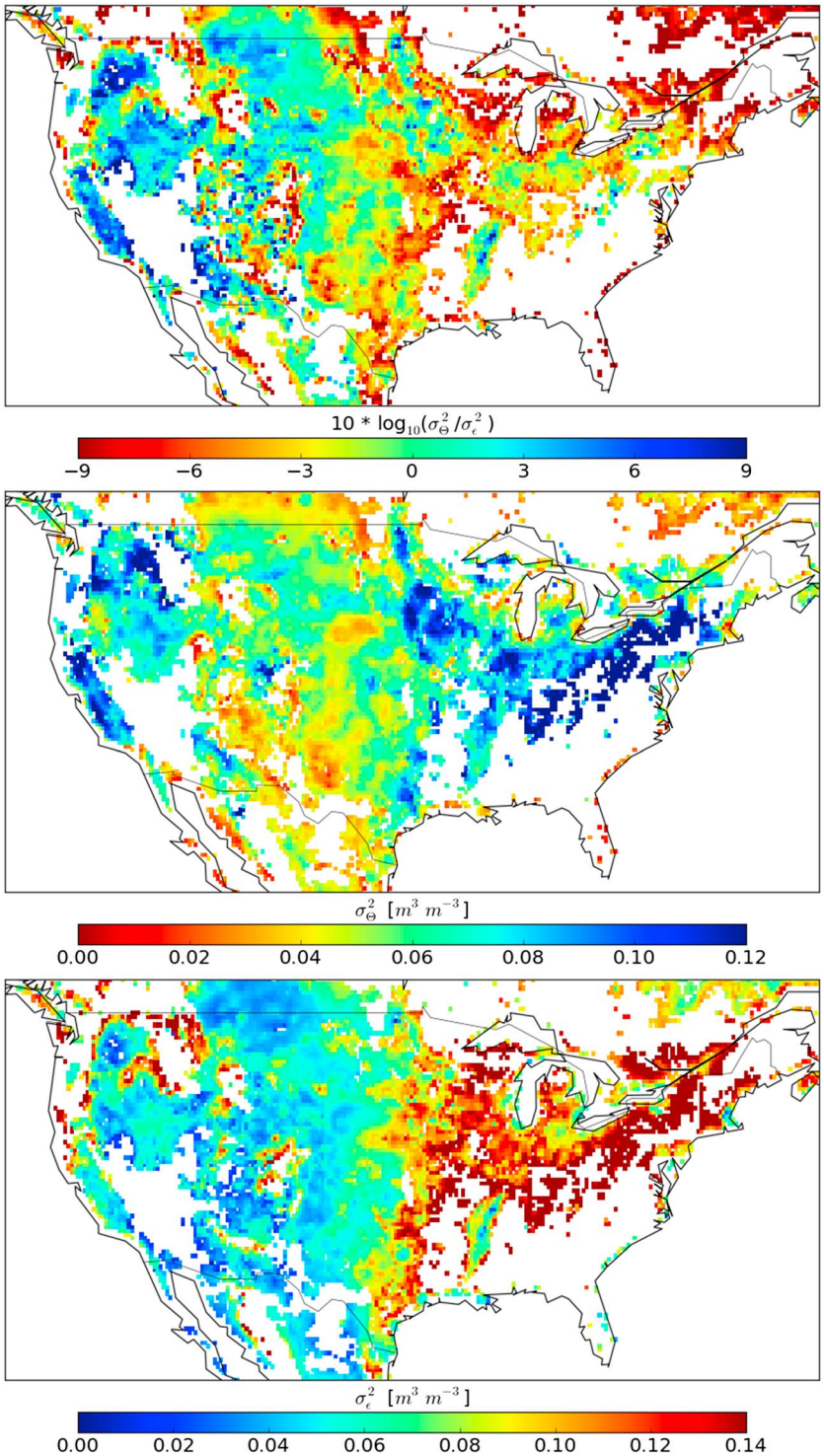

Figure A1. Example of the SNR (dB) of MetOp ASCAT soil moisture retrievals (top) decomposed into its signal (middle) and noise (bottom) components.

technique enables the quantification of E_c, E_{ins}, and E_{tot}. In *Kinzel et al.* [2016] MTC was applied to estimate E_{tot} of the near-surface humidity (q_a) from HOAPS.

For this, selected hourly q_a ship observations of the merged DWD-ICOADS data base (German Meteorological Service (DWD)) [*Woodruff et al.*, 2011] are collocated to respective instantaneous HOAPS estimates. Temporal and spatial decorrelation scales are set to $dt = \pm 180$ min and $dx = \pm 50$ km, resulting from an autocorrelation function analysis. On the basis of these scales, triplets are formed, which consist of (i) two independent ship records and one satellite pixel (i.e., TC1) and (ii) a ship record along with two matching, independent satellite records (i.e., TC2). The resulting TC1 triplets are sorted according to HOAPS q_a magnitudes and subdivided into 20 bins, where each segment includes 5% of all TC1 triplets. TC2 triplets are assigned to these segments, depending on their HOAPS q_a. The following processing steps (that is, equations (A5)–(A9)) are performed separately for each of the 20 segments.

A bias correction with respect to the in situ source is performed, implying that the pairwise differences are solely related to random uncertainty sources.

Generally, the variance of differences between two data sets x and y, V_{xy}, can be quantified as follows:

$$V_{xy} = \mathrm{var}(x) + \mathrm{var}(y) - 2\mathrm{cov}(x, y) \tag{A5}$$

To express each terms shown in equation (A5) as a sum of supposedly random uncertainties, the following error models are assumed for in situ (s) and satellite (sat) data:

$$E_s = E_{ins} \tag{A6a}$$

$$E_{sat} = E_M + E_N \tag{A6b}$$

As each of TC1 and TC2 allow three different combinations of two data sources, equation (A5) can be applied six times in total, using equations (A6a)–(A6b). This automatically gives rise to E_C terms, as all V_{xy} are subject to random collocation uncertainties. The following system of equation results for TC1 equations (A7a)–(A7c) and TC2 equations (A8a)–(A8c):

$$V_{s1,s2} = 2(E_{ins})^2 + (E_C)^2 \tag{A7a}$$

$$V_{s1,sat} = (E_{ins})^2 + (E_M)^2 + (E_N)^2 + (E_C)^2 \tag{A7b}$$

$$V_{s2,sat} = (E_{ins})^2 + (E_M)^2 + (E_N)^2 + (E_C)^2 \tag{A7c}$$

$$V_{s,sat1} = (E_{ins})^2 + (E_M)^2 + (E_N)^2 + (E_C)^2 \tag{A8a}$$

$$V_{s,sat2} = (E_{ins})^2 + (E_M)^2 + (E_N)^2 + (E_C)^2 \tag{A8b}$$

$$V_{sat1,sat2} = 2(E_N)^2 + (E_C)^2 \tag{A8c}$$

To solve equation (A8c) for E_C, E_N (constant) needs to be synthetically quantified [see *Kinzel et al.*, 2016]. Next, E_{ins} can be derived using equation (A7a). An arithmetic mean E_M is subsequently derived from solving equations (A7b)–(A8b). Finally, the random retrieval uncertainty results from

$$E_{tot} = \sqrt{(E_M)^2 + (E_N)^2} \tag{A9}$$

As shown, E_{tot} cannot be solved for using either only TC1 (i.e., equations (A7a)–(A7c)) or TC2 (i.e., equations (A8a)–(A8c)), which demonstrates the advantage of the MTC approach regarding a successful decomposition of all random errors inherent to HOAPS q_a.

The bin-wise random uncertainty decomposition is illustrated in Figure A2. It shows the fairly linear increase of E_{ins} with q_a (yellow), whereas both E_C (black) and E_{tot} (red) maximize over the subtropical q_a regime of 12–17 g kg^{-1}. A thorough discussion of this decomposition pattern and implications regarding the assignment of instantaneous random retrieval uncertainties can be found in *Kinzel et al.* [2016].

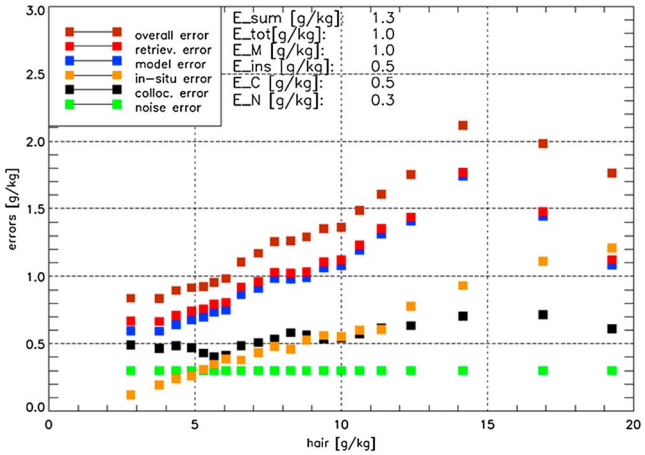

Figure A2. Decomposition of satellite- and MTC-related HOAPS q_a ("hair") error terms, based on MTC matchups between 1995 and 2008, equatorward of 60°N/S. The decomposition is based on 18,005 triplets per TC version per bin. The strings at the top indicate overall arithmetic means of the individual random error contributions. E_{sum} (brown) represents the sum of E_{tot}, E_C, and E_{ins}, that is the noncorrected overall random uncertainty. ©American Meteorological Society. Used with permission.

Assumptions:

1. E_M, E_N, and E_C are satellite independent and do not depend on each other.
2. E_N is Gaussian distributed.
3. E_C only incorporates uncertainties owing to dx and dt and is thus constant for each segment throughout equations (A7a)–(A8c). E_C is neither explicitly assigned to the ship observations nor the satellite records. In fact, it is a result of the methodology.
4. The signal-to-noise ratios of both data sources are alike.

A2. Reducing the Representativeness Errors for Rainfall Estimates

The high spatial and temporal variability of precipitation particularly challenges its global monitoring. Passive microwave satellite (PMW) sensors are therefore still indispensable for precipitation monitoring due to their high global coverage. However, as indirect measurements, they require rigorous validation using more direct in situ surface precipitation data. This example demonstrates a method to improve the spatial representation of surface precipitation reference data along ship tracks in order to validate precipitation from PMW satellite data.

The spatial representativeness error arises when subsampling an area using a point or line measurement, which is summarized in the long-standing point-to-area problem. *Burdanowitz* [2016] quantifies this difference using hypothetical ship tracks placed within a hypothetical PMW satellite pixel in the oceanic area of a scanning S-Pol radar. The S-Pol was deployed on the Caribbean Island of Barbuda as part of the Rain In Cumulus Clouds over the Ocean (RICO) [*Rauber et al.*, 2007] campaign (Figure A3). From the comparison of hypothetical satellite data and hypothetical along-track surface data, the different representation of precipitation manifests in two ways. The first way is whether or not precipitation in the satellite pixel is detected along the track and, second, how the along-track precipitation rate compares to that within the simulated satellite pixel.

The along-track detectability of precipitation strongly depends on the length of the ship track on the one hand, and the observed meteorological conditions on the other hand. Longer ship tracks associated with a higher ship speed increase the chance of detecting a rain shower within the satellite pixel by the reference measurements. The track's orientation might matter if the rain showers tend to orient in elongated patterns. Generally, the along-track detectability increases for widespread homogeneous precipitating areas in contrast to clustered showers. However, without ancillary data of the spatial precipitation distribution within the satellite pixel, the along-track detectability of precipitation cannot be increased.

The quantitative track-area comparison leads to two statistical adjustments applied on the reference data. These adjustments ensure that an along-track precipitation rate reasonably represents the precipitation estimate derived from a PMW satellite sensor. Specifically, the first adjustment uses the precipitation event

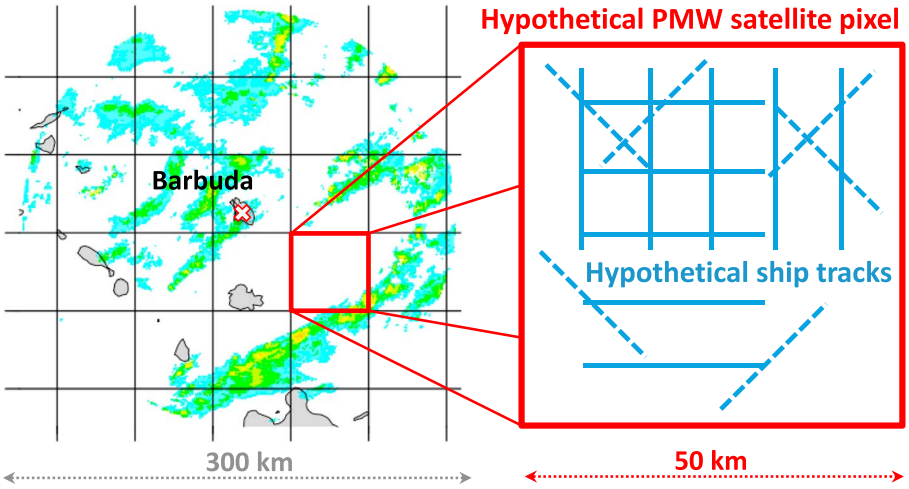

Figure A3. Simulation framework of point-to-area influence on spatial precipitation representation. The surveillance scan from S-Pol radar of RICO contains a red box exemplifying a typical PMW satellite pixel (50 km in diameter). The enlarged box illustrates 16 randomly chosen hypothetical ship tracks of 60 radar pixels in length, corresponding to a typical ship movement of 24 km h^{-1}.

duration along the ship track. This adjustment increases the weight of statistically undersampled short-lasting showers whereas it decreases the weight of statistically oversampled long-lasting showers both along the track. The second adjustment corrects overrepresented and underrepresented precipitation intensities along the track. For this adjustment, a reliable precipitation phase distinction for the surface reference data is required [*Burdanowitz et al.*, 2016]. Both statistical adjustments can strongly reduce the sum of the squared errors for the track area precipitation by more than 80%.

These statistical adjustments were derived for the validation of precipitation from an interim version of HOAPS 3.2 satellite data [*Andersson et al.*, 2010; *Fennig et al.*, 2012] prolonged to 2015 using the Ocean Rainfall And Ice-phase precipitation measurement Network (OceanRAIN; [*Klepp*, 2015]) as surface reference data. The validation procedure follows four major steps, described in *Burdanowitz* [2016]. First, a collocation serves to match both data sets in time and space. Second, an along-track averaging over all OceanRAIN ship measurements collocated to a single HOAPS pixel ensures the optimal spatial coverage for the validation. Third, applying the statistical adjustments to the OceanRAIN along-track averaged precipitation rate makes the reference data better representative of the area of a HOAPS satellite pixel. Fourth, the lower PMW sensor sensitivity in HOAPS requires an exclusion of OceanRAIN precipitation rates below 0.3 mm h^{-1}. As a consequence of the non-Gaussian distribution of precipitation, this cut-off strongly reduces the remaining number of samples to a global number of a few hundreds out of initially 5 million minutes from OceanRAIN. For these remaining samples, HOAPS underestimates the precipitation rate by about 13% with an uncertainty lying in a similar range. However, in a direct comparison of individual cases, HOAPS false detections point at OceanRAIN ship tracks that appear to not capture the precipitating areas within the HOAPS pixel. Solving this remaining aspect of the point-to-area problem would require additional high-resolution sampling of surface precipitation data in order to identify the spatial structure of precipitation within a HOAPS satellite pixel.

This example should serve as general guidance to address the point-to-area problem associated with the validation of precipitation or other non-Gaussian parameters from satellite data using point-like surface reference data. In that respect, future satellite validation applications can profit in a larger scope.

A3. Total Ozone Column Validation: Using an OSSE to Quantify Colocation Mismatch

In the validation of atmospheric data sets, the differences between satellite and ground-based reference measurements can usually not be confronted directly with the reported measurement uncertainties: For all reasonable colocation criteria, i.e., those resulting in a sufficiently large number of comparison pairs, one must also take into account the additional differences due to colocation mismatch, i.e., differences in spatiotemporal sampling and smoothing of the variable and inhomogeneous atmospheric field. A possible approach to quantify these additional uncertainty terms is by performing an Observing System Simulation Experiment (OSSE). This method consists in (1) the creation of appropriate observation operators, which quantify the

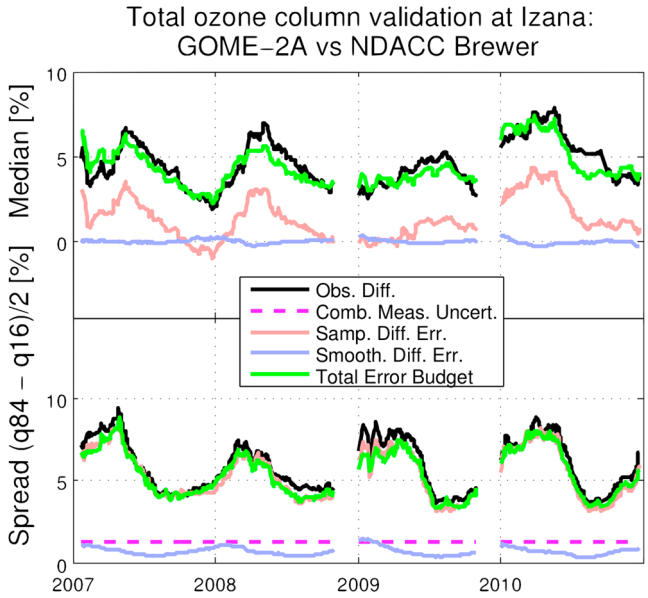

Figure A4. Error budget of comparisons between GODFITv3 GOME-2/MetOp-A satellite total ozone columns and colocated Brewer (daily mean) ground-based measurements from the NDACC station at Izana, Canary Islands. For this illustration, a typical maximum colocation distance of 500 km was imposed. The top panel contains the 3 month running median on the differences, the bottom panel the spread (as an interpercentile). Black lines show the statistics of the observed differences, the colored lines the different components of the error budget, as simulated with an OSSE approach. While the combined measurement uncertainty (magenta) does not account for the observed comparison spread, the error budget can be closed by including the errors due to colocation mismatch (sampling and smoothing differences). Note that the large median error of approximately 3% is a consequence of the mountain top location of the Brewer (direct Sun) instrument, due to which it misses the lower part of the column seen by the satellite measuring down to sea level.

actual 4-D extent of measurement sensitivity of each measurement, (2) application of these observation operators on high-resolution global gridded fields, e.g., from a reanalysis, to simulate the individual measurements, and (3) quantifying the differences between these simulated measurements in an exact copy of the actual validation exercise. When no measurement errors are included in the simulated measurements, this allows an estimate of the differences due solely to colocation mismatch. When also including measurement errors (randomly generated within the measurement uncertainty), one can close the error budget and check to what extent both data sets really agree. This approach was explored for the validation of total ozone column products by *Verhoelst et al.* [2015], and an application of the method on a particular case study of total ozone column validation is shown here in Figure A4.

Figure A4 illustrates that at first sight, the spread on the differences (black line in the bottom panel) is not consistent with combined reported measurement uncertainties (magenta line). However, when taking into account the simulated sampling (orange) and smoothing (blue) differences, the total error budget (green) does agree with the observed spread on the differences. Similarly, the 3 month median on the differences (top panel), which is used as a proxy of systematic errors (on at least a seasonal time scale), is in excellent agreement with the simulation, suggesting negligible systematic errors in the satellite data set.

A4. Indirect Validation of Soil Moisture—SM2RAIN

The validation of coarse resolution satellite soil moisture products is hampered by the unavailability of ground soil moisture observations in many regions, their low density and the low representativeness of point measurements for larger area. Different approaches have been therefore developed to either upscale in situ measurements of soil moisture (see *Crow et al.* [2012], for a review) or downscale coarse resolution soil moisture data [*Peng et al.*, 2017]. The rainfall observation networks are however much more developed than soil moisture networks and allow the assessment of satellite rainfall product on a regular grid basis. Through the SM2RAIN algorithm [*Brocca et al.*, 2014], we are able to obtain rainfall estimates directly from soil moisture observations by inverting the soil water balance equation. Therefore, SM2RAIN provides a method for the indirect validation of satellite soil moisture products through rainfall data.

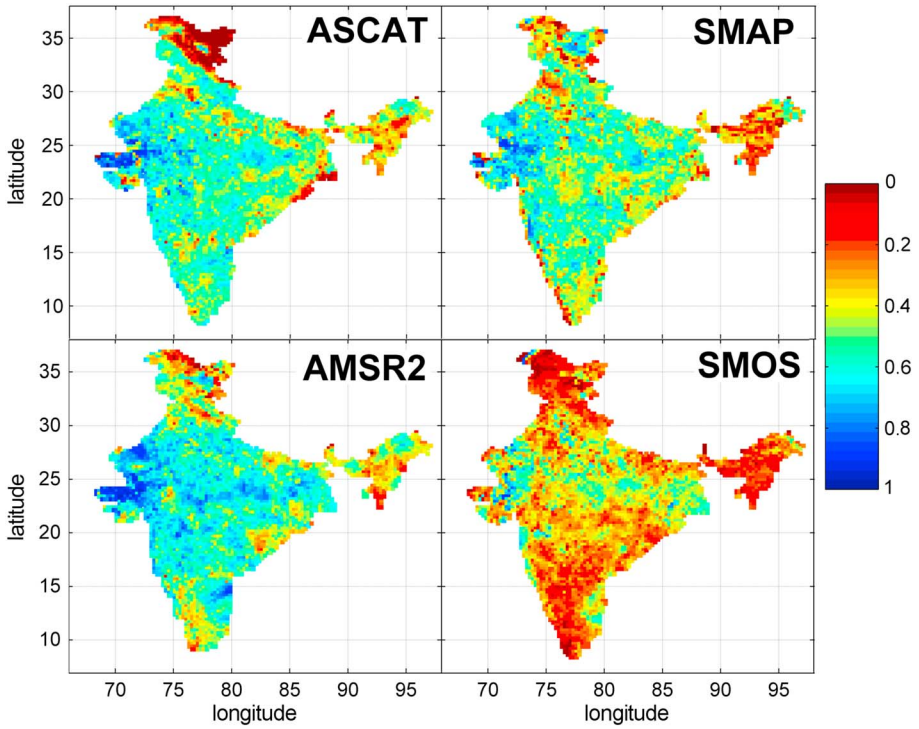

Figure A5. Correlation maps between daily rainfall estimates obtained by ground observations (India Meteorological Department) and satellite soil moisture products, via SM2RAIN, by ASCAT (Advavanced SCATterometer), SMAP (Soil Moisture Active and Passive), SMOS (Soil Moisture Ocean Salinity), and AMSR2 (Advanced Microwave Scanning Radiometer 2) in India. The higher the correlation, the higher the expected quality of satellite soil moisture products. The validation with rainfall data allows to perform the assessment of satellite soil moisture products at the continental scale on a regular grid of 0.25°.

Figure A5 shows an example of indirect validation of four satellite soil moisture products obtained by ASCAT (Advavanced SCATterometer), SMAP (Soil Moisture Active and Passive), SMOS (Soil Moisture Ocean Salinity), and AMSR2 (Advanced Microwave Scanning Radiometer 2) in India in the common period from April to December 2015. Specifically, for each satellite soil moisture product, we calibrated the parameter values of SM2RAIN algorithm in order to reproduce observed rainfall provided by the India Meteorological Department (IMD). The minimization of RMSE between observed and estimated rainfall is used as objective function. Therefore, for each satellite soil moisture product we computed a daily rainfall product that can be compared with ground rainfall data used as benchmark. In Figure A5 the daily Pearson correlation maps for each product are shown. The higher the correlation (blue values), the higher the agreement between observed and estimated rainfall. Indirectly, higher correlations mean better quality of the corresponding satellite soil moisture product. For instance, from Figure A5 it can be inferred that AMSR2 performs the best (median correlation equal to 0.57) followed by ASCAT (0.53) and SMAP (0.49). The lower performance of SMOS (0.31) should be attributed to the influence of Radio Frequency Interferences that, during 2015, affected the northern and southern parts of India.

The example shown here highlights the good potential of indirect methods for the validation of satellite products. Indeed, by using ground measurements, a grid-based validation of satellite soil moisture data is not feasible. By using rainfall observations and SM2RAIN, the validation can be carried out at continental and even global scales.

A5. Consistent Retrieval of Land Surface Variables

An example for the joint retrieval of a consistent set of land surface variables is provided by the Joint Research Centre Two-stream Inversion Package (JRC-TIP) [*Pinty et al.*, 2007, 2008]. The package is built around a one-dimensional representation (called two-stream scheme) [*Pinty et al.*, 2006] of the radiative transfer of the canopy-soil system in the optical domain. Such two-stream radiative transfer schemes are part of the terrestrial components of state-of-art climate and numerical weather prediction models.

In a given spectral band, the two stream model represents the state of the canopy-soil system by the spectrally invariant Leaf Area Index (LAI), the spectrally variable single-scattering albedo and asymmetry factor of the canopy as well as the albedo of the background. From these variables the model simulates the partitioning of the incoming solar energy between the canopy and the soil. In particular, it simulates the fluxes of radiation reflected by, transmitted through, and absorbed by the canopy and that absorbed by the background. The three canopy variables are so-called effective variables, meaning that their values are selected such that observed fluxes are accurately simulated. By contrast, the background albedo is a true variable.

The JRC-TIP can use any combination of available fluxes in multiple spectral bands as observational constraints. The inversion adapts the model's state variables to minimize a misfit function that quantifies the difference between observations and their model-simulated counterparts and the deviation from prior information on the state variables. Once the state is retrieved, in a second step, the model is used to simulate a consistent set of radiant fluxes. This means, that JRC-TIP delivers a set of state variables and fluxes that is consistent with the physics of the radiative transfer model. In particular, this ensures conservation of energy. In the misfit function, the model data misfit is weighted in inverse proportion to the data uncertainty, i.e., the sum of the observational uncertainty and the uncertainty due to model error, while the deviation from the prior is weighed in inverse proportion to the prior uncertainty. This allow approximation of (by the inverse of the misfit function's second derivative) the uncertainty in the retrieved (or posterior) model state. The posterior state uncertainty is then (by a linearization of the model) mapped forward to the posterior flux uncertainty. JRC-TIP thus delivers a full PDF description in the combined state and flux space.

Typical input available from space are albedos in the visible and the near-infrared broad bands. In this configuration the package has been applied to input albedos derived from MODIS [*Pinty et al.*, 2007, 2008, 2011a, 2011b], MISR [*Pinty et al.*, 2007, 2008], and Globalbedo [*Disney et al.*, 2016]. Applications to MODIS and Globalbedo were at global scale in 1 km resolution.

The validation of JRC-TIP products is conducted at four levels (see *Kaminski et al.* [2017], for details):

1. Assessment of the quality of the two-stream model against benchmarks provided by comprehensive three-dimensional Monte Carlo model simulations [*Pinty et al.*, 2001; *Widlowski et al.*, 2007, 2011].
2. Comparison of retrieved variables against in situ observations. Owing to the consistency of the retrieved state variables and fluxes, we can use observations for any of them to validate the entire set of derived products. *Pinty et al.* [2011c] demonstrates this procedure for observations of direct transmission collected over a 400 m transect in a mixed deciduous beech forest in eastern Germany. The observations provide only limited spatial sampling of the MODIS 1 km cell over which the retrieval is performed. Also, the temporal coverage is limited so that the observations from several years were transformed to a standard year through a statical procedure called Random Forest Algorithm.
3. Comparison of the variables retrieved from different albedo input, as demonstrated over selected sites for inputs from MODIS and MISR by *Pinty et al.* [2007, 2008].
4. Comparison against products derived with alternative retrieval approaches [see, e.g., *Disney et al.*, 2016].

A6. Temporal Stability Assessments

A6.1. Stability of the METEOSAT Multidecadal Albedo Record

A multidecadal data product of surface albedo derived from geostationary satellite observations is available from EUMETSAT [*Govaerts et al.*, 2008; *Lattanzio et al.*, 2015]. Observations from geostationary satellites have enabled the retrieval of surface albedo since 1982 and therefore make it an important information source for climate studies. Details on the Meteosat surface albedo (MSA) retrieval scheme and data product validation are available in the literature [*Loew and Govaerts*, 2010; *Lattanzio et al.*, 2013, 2015]. A unique feature of the MSA product is that it provides quantitative uncertainty information on a per pixel basis. Details how this uncertainty information is calculated is provided in *Govaerts and Lattanzio* [2007].

Long-term time series of the MSA product are available for two spatial domains. The first one (0DEG) has its nadir location at approximately 0° longitude above the equator and covers mainly Africa and Europe, while the so-called Indian ocean coverage (IODC) is centered at either 57°E or 63°E, dependent on the satellite location and covers mainly large parts of Africa, the Arabian peninsula as well as large parts of southern and central Asia.

Results of a temporal stability analysis result based on a number of 1627 sites of the SAVS database [*Loew et al.*, 2016] are provided in Figure A6 for both the 0DEG and IODC coverage. A total of 331 of these sites were

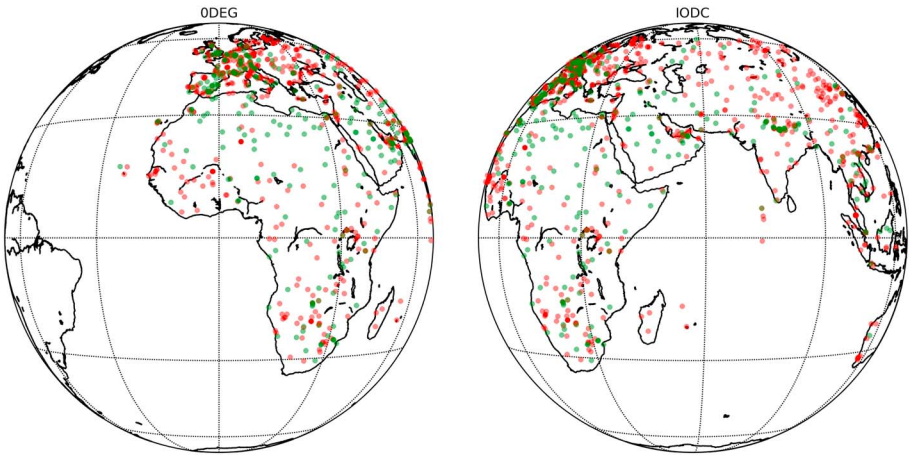

Figure A6. Results of temporal stability analysis for SAVS sites covered by the MSA data set for the 0DEG and the IODC coverage. GCOS criteria is met (green), GCOS criteria is not met (red).

found to meet either the absolute or relative GCOS criteria [*WMO*, 2011] which corresponds to roughly 20% of the investigated sites. These are indicated in green in Figure A6. They cover a wide range of latitudes and are located in both, the 0DEG and IODC Meteosat footprints.

The temporal stability was estimated by calculating the probability that the statistical estimator for the temporal stability $\hat{\beta}$ is smaller than a predefined threshold. A weighted least squares regression approach was used for the estimation of $\hat{\beta}$, which allows the uncertainties of each data sample to be explicitly taken into account. It was found that the quantitative uncertainty information in the MSA data product significantly helps to obtain more robust estimators for the long-term stability ($\hat{\beta}$).

A6.2. Sea Surface Temperature

Stability in the SST data record is critical for climate applications assessing long-term trends in surface temperatures, with GCOS stability requirements for SST as an Essential Climate Variable (ECV) of <0.03 K per decade over spatial scales of 100 km [*World Meteorological Organization*, 2016]. The Along-Track Scanning Radiometer Reprocessing for Climate (ARC) and European Space Agency Sea Surface Temperature (SST) Climate Change Initiative (CCI) projects both provide EO SST data sets for climate, constructed using observations from series of satellite instruments [*Merchant et al.*, 2014, 2012]. Several things are important for achieving a stable SST record from EO data: first, harmonization between satellite instruments in a series; second, correction for sensor drift; and third, time-of-day adjustment to ensure stable sampling within the diurnal cycle [*Merchant et al.*, 2014]. EO SST data set stability can be assessed with reference to in situ data, but globally this is challenging and not possible at spatial scales of 100 km, as in situ observation networks are often not designed to meet climate stability requirements, e.g., data coverage, length of record, and accuracy of measurements [*Kennedy*, 2013]. Multidecadal stability for both the ARC and SST CCI data sets is assessed in the Pacific only, using tropical moored buoy SST measurements at 1 m depth. The target stability for ARC SSTs of 0.005 K yr^{-1}, was met in the tropics from 1994 to 2010 [*Merchant et al.*, 2012], and the L4 analysis product from the SST CCI project, containing blended data from several satellite instrument records had a stability of ($0.1 < \text{trend} < 0.32$ mK yr^{-1}) over the 1995–2010 period [*Rayner et al.*, 2014].

A6.3. Water Vapor

Water vapor is the most important greenhouse gas, is playing a dominant role in feedback mechanism related to climate change, and thus is a key variable for climate analysis. Its relevance is reflected in a large number of available data records (see, e.g., www.gewex-vap.org/). In the following exemplary results from the analysis of the stability of water vapor, data records are shown. The first example for the assessment of the temporal stability is taken from the GEWEX water vapor assessment (G-VAP). Details on G-VAP can be found at www.gewex-vap.org and in *Schröder et al.* [2016a, 2016b].

Within G-VAP the trend estimation largely follows the work of *Weatherhead et al.* [1998] and *Mieruch et al.* [2014]. The estimation of the trend uncertainty considers the noise as computed from the anomaly and autocorrelation, also following *Weatherhead et al.* [1998]. Trend estimation is applied as a tool to identify stability issues in the data records; climate change is not discussed.

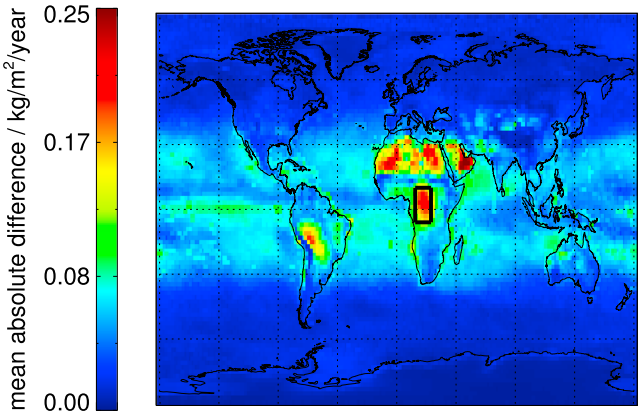

Figure A7. Mean absolute difference of trend estimates using 11 data records. The 11 data records include NCEP reanalysis CFSR, ECMWF reanalyses ERA-Interim and ERA-20C, Japanese reanalysis JRA55, the NASA reanalyses MERRA and MERRA-2 (MERRA MA, MERRA-M2), HIRS based product from NASA (NNHIRS), and the combined satellite, ground-based, and in situ data record NVAPM from CSU [updated from *Schröder et al.*, 2016b].

Also, the relation between saturation vapor pressure and temperature, expressed by the Clausius-Clapeyron equation, can be used to estimate the expected change in TCWV given the change in temperature. Assuming constant relative humidity and pressure, the change in saturation vapor pressure depends on the air temperature and the change of the air temperature. Given a change in air temperature of 1 K, the expected change in mixing ration is 6% at 300 K and 7.5% at 275 K. This relation can be computed as regression of TCWV in percent and SST in Kelvin. Prior to the regression, the SST and TCWV anomaly time series are smoothed with a 12 month low-pass filter as in *Mears et al.* [2007]. A series of additional assumptions need to be applied when relating changes in SST to changes in TCWV. These are described in, e.g., *Mieruch et al.* [2014]. A comparison with theoretical expectations are a valuable sanity check but the underlying assumptions need to be kept in mind.

The Penalized Maximal *F* test [*Wang*, 2008a, 2008b] is utilized to detect break points, because it can be applied to time series of (deseasonalized) anomalies and of anomaly differences and because it does not require supervision. Here break points are considered if the associated level of significance is 0.05 or smaller. Then, the null hypothesis of a break free time series needs to be rejected. For each break point detected, the PMF test returns the size of the shift in the model fitted to the time series at the break point. Then, the step size relative to the variability, further called break size, is computed. Homogeneity, i.e., the presence of breakpoints is assessed on the basis of anomaly differences. As references, HOAPS-3.2 observations (over ocean) and ERA-Interim reanalysis (over land) have been used.

Prior to analyses the data records have been regridded onto a common grid over the common period 1988–2008 [*Schröder et al.*, 2016a]. The utilized data records are listed in the caption of Figure A7.

Results of the trend estimation are shown in Figure A7. The mean absolute difference among the trend estimates from 11 data records exhibits small differences in the extratropics, medium differences in parts of the subtropics and tropics and regional maxima over South America, Central Africa, Sahara, and the Arabian Peninsula. Among others, the homogeneity was analyzed using regional anomaly difference time series over Central Africa (marked with a black box in Figure A7). As shown in *Schröder et al.* [2016b] various break points are observed which differ among data records, in sign and break size and as a function of time. The latter indicates that ERA-Interim, used here as reference to compute anomalies, is (typically) not causing the break points. Also, the break points frequently coincide with changes in the observing system. It is emphasized that the time, the sign, and the break size of break points are typically a function of region and data record and are typically not evident on global scale. It is obviously important to verify the homogeneity and the stability on global and all regional scales [*Schröder et al.*, 2016a, 2016b].

More generally, *Schröder et al.* [2016b] concluded that on regional and global ice-free ocean scales the TCWV trend estimates are often significantly different among the different data records and that these differences were typically found to be caused by breakpoints or series of breakpoints. Not surprisingly all data records exhibit regression values outside the theoretically expected range, except the HOAPS and REMSS data records.

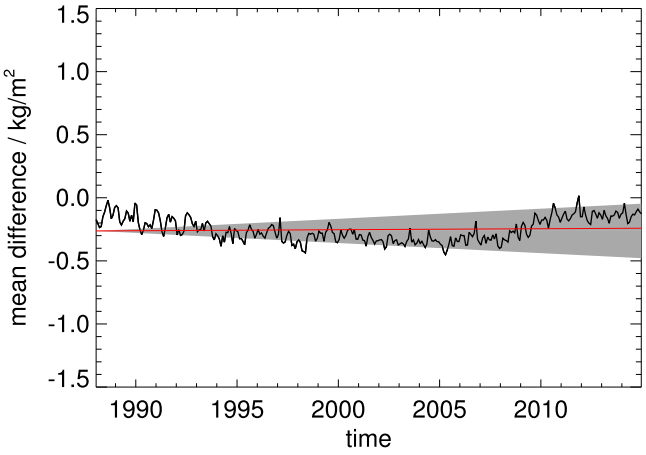

Figure A8. Time series of mean difference between TCWV from HOAPS-4 and REMSS v7. The red line shows the slope from linear regression and the requirement (0.08 kg/m^2) is used to define the grey-shaded area.

The second example on the examination of stability comes from EUMETSAT's CM SAF. Within CM SAF the product requirements, the metric and the reference data records, are defined in an external review process. More specifically, the CM SAF long-term data records are compared to a variety of reference data records, mainly because a data record with global coverage, known uncertainties, and bias-free at all scales is not available. In order to allow a unique conclusion regarding the requirements, a single data record is explicitly defined for compliance testing to requirements. For TCWV the microwave-based product from REMSS, in version 7, is used, among others, to verify stability. The (optimal) stability requirement has been derived as follows: As described above the functional relationship between saturation water vapor pressure and temperature expressed by the Clausius-Clapeyron equation leads to the expectation that TCWV increases by approximately 7% per 1 K increase of the global mean temperature. With a warming rate of 0.2°C/decade this is equivalent to an increase of 1.4%/decade. Using a factor of 0.2 [*Ohring et al.*, 2005] yields a stability requirement of 0.078 kg/m^2 (0.28%), which has been rounded to 0.08 kg/m^2.

The HOAPS-4 data record has been regridded to the REMSS grid. Then, the mean difference has been computed as global monthly averages of all valid, collocated pairs using the cosine of the latitude as weight. The slope is the output of a linear regression which does not consider uncertainties of the input nor outliers. The probability that the slope, here the stability, is smaller than the requirement is estimated based on the uncertainty of the slope and the requirement. Figure A8 shows the time series of the mean difference, the linear regression and as shaded area the requirement. The actual stability is 0.008 $\pm$ 0.007 kg/m^2/decade. The probability that the slope is smaller than the requirement is >99%, and thus, the requirement is fulfilled.

Acknowledgments
Sadly, this review paper has become one of the last scientific achievements of Prof. Dr. Alexander Loew. His co-authors, the ISSI team which he lead with such enthusiasm, are grateful for having had the opportunity to work so closely with him, and they will remember him for his sharp mind and warm heart. We acknowledge the support from the International Space Science Institute (ISSI). This publication is an outcome of the ISSI's Team on "EO Validation Across Scales." This work has received funding from the EC FP7 Quality Assurance for Essential Climate Variables (QA4ECV) (grant agreement 607405) project for the preparation of the Copernicus Climate Service, and the EC Horizon 2020 project GAIA-CLIM (grant agreement 640276). The contents of this publication are the sole responsibility of authors and can in no way be taken to reflect the views of the European Commission. Parts of the work were also supported through the ESA Climate Change Initiative (CMUG and GHG-CCI) and the European Commission's eartH2Observe project (European Union's Seventh Framework Program, grant agreement 603608). R.V.D. has been financially supported by the German Federal Ministry for Education and Research (BMBF) via the Young Investigators Group CoSy-CC^2 (grant 01LN1306A). J.K. has been funded by the German Research Foundation (DFG), Research Unit 1740 (Atlantic Freshwater Cycle). M.S. acknowledges the financial support by the EUMETSAT member states through CM SAF. Free access to the microwave-based TCWV products from REMSS is acknowledged as well.

References

Adams, J., N. Gobron, J.-L. Widlowski, and C. Mio (2016), A model-based framework for the quality assessment of surface albedo in situ measurement protocols, *J. Quant. Spectrosc. Radiat. Transfer*, *180*, 126–146, doi:10.1016/j.jqsrt.2016.04.005.

Andersson, A., K. Fennig, C. Klepp, S. Bakan, H. Graßl, and J. Schulz (2010), The Hamburg ocean atmosphere parameters and fluxes from satellite data—HOAPS-3, *Earth Syst. Sci. Data*, *2*(2), 215–234, doi:10.5194/essd-2-215-2010.

Bell, S. (2001), Measurement good practice guide no. 11 (issue 2): A beginner's guide to uncertainty of measurement, NPL, Teddington, U. K. [Available at https://www.wmo.int/pages/prog/gcos/documents/gruanmanuals/UK_NPL/mgpg11.pdf.]

Bell, W. (Ed.) (2015), Post-launch characterisation of satellite instruments, paper presented at Seminar on Use of Satellite Observations in Numerical Weather Prediction, ECMWF, Shinfield Park, Reading, 8–12 Sept., 2014. [Available at https://www.ecmwf.int/en/elibrary/8048-post-launch-characterisation-satellite-instruments.]

Birman, C., F. Karbou, and J.-F. Mahfouf (2015), Daily rainfall detection and estimation over land using microwave surface emissivities, *J. Appl. Meteorol. Climatol.*, *54*(4), 880–895, doi:10.1175/JAMC-D-14-0192.1.

Bodeker, G. E., et al. (2016), Reference upper-air observations for climate: From concept to reality, *Bull. Am. Meteorol. Soc.*, *97*(1), 123–135, doi:10.1175/BAMS-D-14-00072.1.

Boersma, K. F., H. J. Eskes, and E. J. Brinksma (2004), Error analysis for tropospheric NO_2 retrieval from space, *J. Geophys. Res.*, *109*, D04311, doi:10.1029/2003JD003962.

Bojinski, S., M. Verstraete, T. C. Peterson, C. Richter, A. Simmons, and M. Zemp (2014), The concept of essential climate variables in support of climate research, applications, and policy, *Bull. Am. Meteorol. Soc.*, *95*(9), 1431–1443, doi:10.1175/BAMS-D-13-00047.1.

Brocca, L., T. Moramarco, F. Melone, and W. Wagner (2013), A new method for rainfall estimation through soil moisture observations, *Geophys. Res. Lett.*, *40*, 853–858, doi:10.1002/grl.50173.

Brocca, L., L. Ciabatta, C. Massari, T. Moramarco, S. Hahn, S. Hasenauer, R. Kidd, W. Dorigo, W. Wagner, and V. Levizzani (2014), Soil as a natural rain gauge: Estimating global rainfall from satellite soil moisture data, *J. Geophys. Res. Atmos.*, *119*, 5128–5141, doi:10.1002/2014JD021489.

Bulgin, C. E., O. Embury, G. Corlett, and C. J. Merchant (2016), Independent uncertainty estimates for coefficient based sea surface temperature retrieval from the along-track scanning radiometer instruments, *Remote Sens. Environ.*, *178*, 213–222, doi:10.1016/j.rse.2016.02.022.

Burdanowitz, J. (2016), Point-to-area validation of passive microwave satellite precipitation with shipboard disdrometers, PhD thesis, Universität Hamburg, Hamburg, Germany.

Burdanowitz, J., C. Klepp, and S. Bakan (2016), An automatic precipitation-phase distinction algorithm for optical disdrometer data over the global ocean, *Atmos. Meas. Tech.*, *9*(4), 1637–1652, doi:10.5194/amt-9-1637-2016.

Calbet, X., R. Kivi, S. Tjemkes, F. Montagner, and R. Stuhlmann (2011), Matching radiative transfer models and radiosonde data from the EPS/METOP Sodankyla campaign to IASI measurements, *Atmos. Meas. Tech.*, *4*(6), 1177–1189, doi:10.5194/amt-4-1177-2011.

Calbet, X., N. Peinado-Galan, P. Ripodas, T. Trent, R. Dirksen, and M. Sommer (2016), Consistency between GRUAN sondes, LBLRTM and IASI, *Atmos. Meas. Tech. Discuss.*, *2016*, 1–18, doi:10.5194/amt-2016-344.

Compernolle, S., J.-C. Lambert, and S. Niemeijer (2016), Prototype QA/Validation service for atmospheric ECV precursors: Detailed processing model—Version 2, QA4ECV Rep., Belgian Inst. for Space Aeronomy, Bruxelles.

Crow, W. T., D. G. Miralles, and M. H. Cosh (2010), A quasi-global evaluation system for satellite-based surface soil moisture retrievals, *IEEE Trans. Geosci. Remote Sens.*, *48*(6), 2516–2527, doi:10.1109/TGRS.2010.2040481.

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*, RG2002, doi:10.1029/2011RG000372.

Cunnold, D. M., W. P. Chu, R. A. Barnes, M. P. McCormick, and R. E. Veiga (1989), Validation of SAGE II ozone measurements, *J. Geophys. Res.*, *94*(D6), 8447–8460, doi:10.1029/JD094iD06p08447.

Damadeo, R. P., J. M. Zawodny, and L. W. Thomason (2014), Reevaluation of stratospheric ozone trends from SAGE II data using a simultaneous temporal and spatial analysis, *Atmos. Chem. Phys.*, *14*(24), 13,455–13,470, doi:10.5194/acp-14-13455-2014.

Deza, J. I., M. Barreiro, and C. Masoller (2013), Inferring interdependencies in climate networks constructed at inter-annual, intra-season and longer time scales, *Eur. Phys. J. Spec. Top.*, *222*(2), 511–523, doi:10.1140/epjst/e2013-01856-5.

Dirksen, R. J., M. Sommer, F. J. Immler, D. F. Hurst, R. Kivi, and H. Vömel (2014), Reference quality upper-air measurements: GRUAN data processing for the Vaisala RS92 radiosonde, *Atmos. Meas. Tech.*, *7*(12), 4463–4490, doi:10.5194/amt-7-4463-2014.

Disney, M., J.-P. Muller, S. Kharbouche, T. Kaminski, M. Voßbeck, P. Lewis, and B. Pinty (2016), A new global fAPAR and LAI dataset derived from optimal albedo estimates: Comparison with MODIS products, *Remote Sens.*, *8*(4), 275, doi:10.3390/rs8040275.

Domingues, M. O., O. Mendes Jr., and A. Mendes da Costa (2005), On wavelet techniques in atmospheric sciences, *Adv. Space Res.*, *35*, 831–842, doi:10.1016/j.asr.2005.02.097.

Donges, J. F., Y. Zou, N. Marwan, and J. Kurths (2009), The backbone of the climate network, *Europhys. Lett.*, *87*(4), 48007.

Donges, J. F., I. Petrova, A. Loew, N. Marwan, and J. Kurths (2015), How complex climate networks complement eigen techniques for the statistical analysis of climatological data, *Clim. Dyn.*, *45*, 2407–2424.

Donlon, C. J., P. J. Minnett, N. Fox, and W. Wimmer (2014), *Optical Radiometry for Ocean Climate Measurements, Experimental Methods in the Physical Sciences*, vol. 47, chap. 5.2, pp. 557–603, Strategies for the Laboratory and Field Deployment of Ship-Borne Fiducial Reference Thermal Infrared Radiometers in Support of Satellite-Derived Sea Surface Temperature Climate Data Records, Elsevier, Amsterdam.

Donner, R. V., Y. Zou, J. F. Donges, N. Marwan, and J. Kurths (2010), Recurrence networks—A novel paradigm for nonlinear time series analysis, *New J. Phys.*, *12*(3), 33025.

Dorigo, W., et al. (2017), Esa CCI soil moisture for improved Earth system understanding: State-of-the art and future directions, *Remote Sens. Environ.*, doi:10.1016/j.rse.2017.07.001, in review.

Elsayed, M. (2010), An overview of wavelet analysis and its application to ocean wind waves, *J. Coastal Res.*, *26*, 535–540, doi:10.2112/04-0274.1.

Entekhabi, D., and I. Rodriguez-Iturbe (1994), Analytical framework for the characterization of the space-time variability of soil moisture, *Adv. Water Resour.*, *17*, 35–45, doi:10.1016/0309-1708(94)90022-1.

Eyring, V., et al. (2016), Esmvaltool (v1.0)—A community diagnostic and performance metrics tool for routine evaluation of earth system models in CMIP, *Geosci. Model Dev.*, *9*(5), 1747–1802, doi:10.5194/gmd-9-1747-2016.

Fennig, K., A. Andersson, S. Bakan, C. Klepp, and M. Schröder (2012), Hamburg ocean atmosphere parameters and fluxes from satellite data- HOAPS-3.2—Monthly means/6-hourly composites, *Satellite Application Facility on Climate Monitoring (CM SAF)*, German Weather Service (DWD), Offenbach, doi:10.5676/EUM_SAF_CM/HOAPS/V001.

Ghent, D., I. Trigo, A. Pires, O. Sardou, J. Bruniquel, M. Martin, F. Goettsche, C. Prigent, C. Jimenez, and J. Remedios (2016), Globtemperature product user guide, Univ. of Leicester, U. K.

Gobron, N., B. Pinty, F. Melin, M. Taberner, M. M. Verstraete, M. Robustelli, and J.-L. Widlowski (2007), Evaluation of the MERIS/ENVISAT FAPAR product, *Adv. Space Res.*, *39*, 105–115.

Govaerts, Y. M., and A. Lattanzio (2007), Retrieval error estimation of surface albedo derived from geostationary large band satellite observations: Application to Meteosat-2 and meteosat-7 data, *J. Geophys. Res.*, *112*, D05102, doi:10.1029/2006JD007313.

Govaerts, Y. M., A. Lattanzio, M. Taberner, and B. Pinty (2008), Generating global surface albedo products from multiple geostationary satellites, *Remote Sens. Environ.*, *112*(6), 2804–2816, doi:10.1016/j.rse.2008.01.012.

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

Gruber, A., C.-H. Su, S. Zwieback, W. Crow, W. Dorigo, and W. Wagner (2016), Recent advances in (soil moisture) triple collocation analysis, *Int. J. Appl. Earth Obs. Geoinf.*, *45*, 200–211, doi:10.1016/j.jag.2015.09.002.

Guan, L., and H. Kawamura (2004), Merging satellite infrared and microwave SSTS: Methodology and evaluation of the new SST, *J. Oceanogr.*, *60*, 905–912, doi:10.1007/s10872-004-5782-x.

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, *J. Hydrol.*, *377*(1), 80–91.

Hamed, K. H. (2011), The distribution of Kendall's tau for testing the significance of cross-correlation in persistent data, *Hydrol. Sci. J.*, *56*(5), 841–853, doi:10.1080/02626667.2011.586948.

Hamed, K. H. (2016), The distribution of Spearmans rho trend statistic for persistent hydrologic data, *Hydrol. Sci. J.*, *61*(1), 214–223, doi:10.1080/02626667.2014.968573.

Hamming, R. W. (1950), Error detecting and error correcting codes, *Bell Syst. Technol. J.*, *29*(2), 147–160.

Herrnegger, M., H. P. Nachtnebel, and K Schulz (2015), From runoff to rainfall: Inverse rainfall-runoff modelling in a high temporal resolution, *Hydrol. Earth Syst. Sci.*, *19*(11), 4619–4639, doi:10.5194/hess-19-4619-2015.
Hollmann, R., et al. (2013), The ESA climate change initiative: Satellite data records for essential climate variables, *Bull. Am. Meteorol. Soc.*, *94*(10), 1541–1552, doi:10.1175/BAMS-D-11-00254.1.
Hosoda, K., and H. Kawamura (2004), Examination of the merged sea surface temperature using wavelet analysis, *J. Oceanogr.*, *60*, 843–852, doi:10.1007/s10872-004-5777-7.
Hubert, D., et al. (2016), Ground-based assessment of the bias and long-term stability of 14 limb and occultation ozone profile data records, *Atmos. Meas. Tech.*, *9*(6), 2497–2534, doi:10.5194/amt-9-2497-2016.
Immler, F. J., J. Dykema, T. Gardiner, D. N. Whiteman, P. W. Thorne, and H. Vömel (2010), Reference Quality upper-air measurements: Guidance for developing GRUAN data products, *Atmos. Meas. Tech.*, *3*(5), 1217–1231, doi:10.5194/amt-3-1217-2010.
Joint Committee for Guides in Metrology (JCGM) (2008), Evaluation of measurement data—Guide to the expression of uncertainty in measurement, *Tech. Rep.*, BIPM, Sèvres, France.
Joint Committee for Guides in Metrology (JCGM) (2009), Evaluation of measurement data—An introduction to the 'Guide to the expression of uncertainty in measurement' and related documents, *Tech. Rep.*, BIPM, Sèvres, France.
Joint Committee for Guides in Metrology (JCGM) (2011), Evaluation of measurement data—Supplement 2 to the "Guide to the expression of uncertainty in measurement": Extension to any number of output quantities, *Tech. Rep.*, BIPM, Sèvres, France.
Joint Committee for Guides in Metrology (JCGM) (2012), International vocabulary of metrology—Basic and general concepts and associated terms, BIPM, Sèvres, France.
Joo, S., J. Eyre, and R. Marriot (2013), The impact of MetOp and other satellite data within the Met office global NWP system using an adjoint-based sensitivity method, *Mon. Weather Rev.*, *141*(10), 3331–3342, doi:10.1175/MWR-D-12-00232.1.
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, *Int. J. Remote Sens.*, *21*(17), 3383–3390, doi:10.1080/014311600750020000.
Justice, C. O., et al. (2013), Land and cryosphere products from Suomi NPP VIIRS: Overview and status, *J. Geophys. Res. Atmos.*, *118*, 9753–9765, doi:10.1002/jgrd.50771.
Kaminski, T., and P.-P. Mathieu (2017), Reviews and syntheses: Flying the satellite into your model: On the role of observation operators in constraining models of the Earth system and the carbon cycle, *Biogeosciences*, *14*(9), 2343–2357, doi:10.5194/bg-14-2343-2017.
Kaminski, T., P. J. Rayner, M. Voßbeck, M. Scholze, and E. Koffi (2012a), Observing the continental-scale carbon balance: Assessment of sampling complementarity and redundancy in a terrestrial assimilation system by means of quantitative network design, *Atmos. Chem. Phys.*, *12*(16), 7867–7879, doi:10.5194/acp-12-7867-2012.
Kaminski, T., W. Knorr, M. Scholze, N. Gobron, B. Pinty, R. Giering, and P.-P. Mathieu (2012b), Consistent assimilation of MERIS FAPAR and atmospheric CO_2 into a terrestrial vegetation model and interactive mission benefit analysis, *Biogeosciences*, *9*(8), 3173–3184, doi:10.5194/bg-9-3173-2012.
Kaminski, T., B. Pinty, M. Voßbeck, M. Lopatka, N. Gobron, and M. Robustelli (2017), Consistent retrieval of land surface radiation products from EO, including traceable uncertainty estimates, *Biogeosciences*, *14*(9), 2527–2541, doi:10.5194/bg-14-2527-2017.
Kennedy, J. J. (2013), A review of uncertainty in in situ measurements and data sets of sea surface temperature, *Rev. Geophys.*, *52*, 1–32, doi:10.1002/2013RG000434.
Keppens, A., et al. (2015), Round-robin evaluation of nadir ozone proFIle retrievals: Methodology and application to MetOp-A GOME-2, *Atmos. Meas. Tech.*, *8*, 2093–2120, doi:10.5194/amt-8-2093-2015.
Kerr, Y. H. et al. (2010), The SMOS mission: New tool for monitoring key elements of the global water cycle, *Proc. IEEE*, *98*(5), 666–687, doi:10.1109/JPROC.2010.2043032.
Kinzel, J., K. Fennig, M. Schröder, A. Andersson, K. Bumke, and R. Hollmann (2016), Decomposition of random errors inherent to HOAPS-3.2 near-surface humidity estimates using multiple triple collocation analysis, *J. Atmos. Oceanic Technol.*, *33*, 1455–1471, doi:10.1175/JTECH-D-15-0122.1.
Kirchner, J. W. (2009), Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward, *Water Resour. Res.*, *45*, W02429, doi:10.1029/2008WR006912.
Klepp, C. (2015), The oceanic shipboard precipitation measurement network for surface validation—OceanRAIN, *Atmos. Res.*, *163*, 74–90.
Knorr, W. (2000), Annual and interannual CO_2 exchanges of the terrestrial biosphere: Process-based simulations and uncertainties, *Global Ecol. Biogeogr.*, *9*(3), 225–252.
Knorr, W., T. Kaminski, M. Scholze, N. Gobron, B. Pinty, R. Giering, and P.-P. Mathieu (2010), Carbon cycle data assimilation with a generic phenology model, *J. Geophys. Res.*, *115*, G04017, doi:10.1029/2009JG001119.
Kobayashi, S., P. Poli, and V. O. John (2017), Characterisation of special sensor microwave water vapor profiler (SSM/T-2) radiances using radiative transfer simulations from global atmospheric reanalyses, *Adv. Space Res.*, *59*(4), 917–935, doi:10.1016/j.asr.2016.11.017.
Koenker, R. (2005), *Quantile Regression, Econometric Society Monographs*, vol. 38, Cambridge Univ. Press, Cambridge, U. K.
Kraskov, A., H. Stögbauer, and P. Grassberger (2004), Estimating mutual information, *Phys. Rev. E*, *69*, 66138, doi:10.1103/PhysRevE.69.066138.
Laeng, A., et al. (2015), Validation of MIPAS IMK/IAA methane profiles, *Atmos. Meas. Tech.*, *8*, 5251–5261, doi:10.5194/amt-8-5251-2015.
Lattanzio, A., J. Schulz, J. Matthews, A. Okuyama, B. Theodore, J. J. Bates, K. R. Knapp, Y. Kosaka, and L. SchÃller (2013), Land surface albedo from geostationary satelites: A multiagency collaboration within SCOPE-CM, *Bull. Am. Meteorol. Soc.*, *94*(2), 205–214, doi:10.1175/BAMS-D-11-00230.1.
Lattanzio, A., F. Fell, R. Bennartz, I. F. Trigo, and J. Schulz (2015), Quality assessment and improvement of the EUMETSAT Meteosat surface albedo climate data record, *Atmos. Meas. Tech.*, *8*(10), 4561–4571, doi:10.5194/amt-8-4561-2015.
Lauer, A., et al. (2017), Benchmarking CMIP5 models with a subset of ESA CCI phase 2 data using the {ESMValTool}, *Remote Sens. Environ.*, doi:10.1016/j.rse.2017.01.007.
Lehmann, E. L., and J. P. Romano (2005), *Testing Statistical Hypotheses*, 3rd ed., Springer, New York.
Lewis, P., J. Gmez-Dans, T. Kaminski, J. Settle, T. Quaife, N. Gobron, J. Styles, and M. Berger (2012), An earth observation land data assimilation system (EO-LDAS), *Remote Sens. Environ.*, *120*, 219–235, doi:10.1016/j.rse.2011.12.027. the Sentinel Missions - New Opportunities for Science.
Loew, A. (2014), Terrestrial satellite records for climate studies: How long is long enough? A test case for the Sahel, *Theor. Appl. Climatol.*, *115*(3), 427–440, doi:10.1007/s00704-013-0880-6.
Loew, A., and Y. Govaerts (2010), Towards multidecadal consistent METEOSAT surface albedo time series, *Remote Sens.*, *2*(4), 957–967, doi:10.3390/rs2040957.
Loew, A., and F. Schlenz (2011), A dynamic approach for evaluating coarse scale satellite soil moisture products, *Hydrol. Earth Syst. Sci.*, *15*(1), 75–90.

Loew, A., R. Bennartz, F. Fell, A. Lattanzio, M. Doutriaux-Boucher, and J. Schulz (2016), A database of global reference sites to support validation of satellite surface Albedo datasets (savs 1.0), *Earth Syst. Sci. Data*, *8*(2), 425–438, doi:10.5194/essd-8-425-2016.

Lu, Q., and W. Bell (2014), Characterizing channel center frequencies in AMSU-A and MSU microwave sounding instruments, *J. Atmos. Oceanic Technol.*, *31*(8), 1713–1732, doi:10.1175/JTECH-D-13-00136.1.

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

Mears, C. A., B. D. Santer, F. J. Wentz, K. E. Taylor, and M. F. Wehner (2007), Relationship between temperature and precipitable water changes over tropical oceans, *Geophys. Res. Lett.*, *34*, L24709, doi:10.1029/2007GL031936.

Merchant, C. J. et al. (2012), A twenty-year independent record of sea surface temperature for climate from along-track scanning radiometers, *J. Geophys. Res.*, *117*, C12013, doi:10.1029/2012JC008400.

Merchant, C. J., et al. (2014), Sea surface temperature datasets for climate applications from phase 1 of the European Space Agency Climate Change Initiative (SST CCI), *Geosci. Data J.*, *1*, 179–191.

Mieruch, S., M. Schröder, S. Noël, and J. Schulz (2014), Comparison of decadal global water vapor changes derived from independent satellite time series, *J. Geophys. Res. Atmos.*, *119*, 12,489–12,499, doi:10.1002/2014JD021588.

Morin, E. (2011), To know what we cannot know: Global mapping of minimal detectable absolute trends in annual precipitation, *Water Resour. Res.*, *47*, W07505, doi:10.1029/2010WR009798.

Nappo, C. J., et al. (1982), "The Workshop on the Representativeness of Meteorological-Observations", June 1981, Boulder, Colorado, *Bull. Am. Meteorol. Soc.*, *63*, 761–764.

Newman, S. M., A. M. Larar, W. L. Smith, I. V. Ptashnik, R. L. Jones, M. I. Mead, H. Revercomb, D. C. Tobin, J. K. Taylor, and J. P. Taylor (2012), The Joint Airborne IASI Validation Experiment: An evaluation of instrument and algorithms, *J. Quant. Spectrosc. Radiat. Transfer*, *113*, 1372–1390, doi:10.1016/j.jqsrt.2012.02.030.

Ohring, G., B. Wielicki, R. Spencer, B. Emery, and R. Datla (2005), Satellite instrument calibration for measuring global climate change: Report of a Workshop, *Bull. Am. Meteorol. Soc.*, *86*, 1303–1313, doi:10.1175/BAMS-86-9-1303.

Otto, J., et al. (2016), Uncertainty: Lessons learned for climate services, *Bull. Am. Meteorol. Soc.*, *97*(12), ES265–ES269, doi:10.1175/BAMS-D-16-0173.1.

Paluš, M., D. Hartman, J. Hlinka, and M. Vejmelka (2011), Discerning connectivity from dynamics in climate networks, *Nonlinear Processes Geophys.*, *18*, 751–763.

Peng, J., A. Loew, O. Merlin, and N. E. C. Verhoest (2017), A review of spatial downscaling of satellite remotely sensed soil moisture, *Rev. Geophys.*, *55*, doi:10.1002/2016RG000543.

Pinty, B. et al. (2001), Radiation Transfer Model Intercomparison (RAMI) exercise, *J. Geophys. Res.*, *106*, 11,937–11,956, doi:10.1029/2000JD900493.

Pinty, B., T. Lavergne, R. E. Dickinson, J.-L. Widlowski, N. Gobron, and M. M. Verstraete (2006), Simplifying the interaction of land surfaces with radiation for relating remote sensing products to climate models, *J. Geophys. Res.*, *111*, D02116, doi:10.1029/2005JD005952.

Pinty, B., T. Lavergne, M. Voßbeck, T. Kaminski, O. Aussedat, R. Giering, N. Gobron, M. Taberner, M. M. Verstraete, and J.-L. Widlowski (2007), Retrieving surface parameters for climate models from MODIS-MISR albedo products, *J. Geophys. Res.*, *112*, D10116, doi:10.1029/2005JD005952.

Pinty, B., T. Lavergne, T. Kaminski, O. Aussedat, R. Giering, N. Gobron, M. Taberner, M. M. Verstraete, M. Voßbeck, and J.-L. Widlowski (2008), Partitioning the solar radiant fluxes in forest canopies in the presence of snow, *J. Geophys. Res.*, *113*, D04104, doi:10.1029/2007JD009096.

Pinty, B., I. Andredakis, M. Clerici, T. Kaminski, M. Taberner, M. M. Verstraete, N. Gobron, S. Plummer, and J.-L. Widlowski (2011a), Exploiting the modis albedos with the two-stream inversion package (JRC-TIP): 1. Effective leaf area index, vegetation, and soil properties, *J. Geophys. Res.*, *116*, D09105, doi:10.1029/2010JD015372.

Pinty, B., M. Clerici, I. Andredakis, T. Kaminski, M. Taberner, M. M. Verstraete, N. Gobron, S. Plummer, and J.-L. Widlowski (2011b), Exploiting the modis albedos with the two-stream inversion package (JRC-TIP): 2. Fractions of transmitted and absorbed fluxes in the vegetation and soil layers, *J. Geophys. Res.*, *116*, D09106, doi:10.1029/2010JD015373.

Pinty, B., M. Jung, T. Kaminski, T. Lavergne, M. Mund, S. Plummer, E. Thomas, and J.-L. Widlowski (2011c), Evaluation of the JRC-TIP 0.01 products over a mid-latitude deciduous forest site, *Remote Sens. Environ.*, *115*, 3567–3581, doi:10.1016/j.rse.2011.08.018.

Praveen, K., and E. Foufoula-Georgiou (1997), Wavelet analysis for geophysical applications, *Rev. Geophys.*, *35*, 385–412, doi:10.1029/97RG00427.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992), *Numerical Recipes in C — The Art of Scientific Computing*, 2nd ed., Cambrige Univ. Press, Cambridge, U. K.

Radebach, A., R. V. Donner, J. Runge, J. F. Donges, and J. Kurths (2013), Disentangling different types of El Niño episodes by evolving climate network analysis, *Phys. Rev. E*, *88*(5), 52807, doi:10.1103/PhysRevE.88.052807.

Rauber, R. M. et al. (2007), Rain in shallow cumulus over the ocean: The RICO campaign, *Bull. Am. Meteorol. Soc.*, *88*(12), 1912–1928, doi:10.1175/BAMS-88-12-1912.

Rayner, P., M. Scholze, W. Knorr, T. Kaminski, R. Giering, and H. Widmann (2005), Two decades of terrestrial Carbon fluxes from a Carbon Cycle Data Assimilation System (CCDAS), *Global Biogeochem. Cycles*, *19*, GB2026, doi:10.1029/2004GB002254.

Rayner, N., J. Kennedy, C. Atkinson, T. Graham, E. Fiedler, A. McLaren, and G. Corlett, (2014), ESA SST CCI climate assessment report. [Available at http://www.esa-sst-cci.org/PUG/pdf/SST_CCI-CAR-UKMO-201-Issue_1-signed.pdf.]

Rayner, P., A. M. Michalak, and F. Chevallier (2016), Fundamentals of data assimilation, *Geosci. Model Dev. Discuss.*, *2016*, 1–21, doi:10.5194/gmd-2016-148.

Rehfeld, K., and J. Kurths (2014), Similarity estimators for irregular and age-uncertain time series, *Clim. Past*, *10*(1), 107–122, doi:10.5194/cp-10-107-2014.

Rodgers, C. D. (1976), Retrieval of atmospheric temperature and composition from remote measurement of thermal radiation, *Rev. Geophys.*, *14*(4), 609–624.

Rodgers, C. D. (2000), *Inverse Methods for Atmospheric Sounding, Series on Atmospheric, Oceanic and Planetary Physics*, vol. 2, World Sci., Singapore.

Rodgers, C. D., and B. J. Connor (2003), Intercomparison of remote sounding instruments, *J. Geophys. Res.*, *108*(D3), 4116, doi:10.1029/2002JD002299.

Roman, J., R. Knuteson, and S. Ackerman (2014), Time-to-detect trends in precipitable water vapor with varying measurement error, *J. Clim.*, *27*(21), 8259–8275, doi:10.1175/JCLI-D-13-00736.1.

Roman, J., R. Knuteson, T. August, T. Hultberg, S. Ackerman, and H. Revercomb (2016), A global assessment of NASA AIRS v6 and EUMETSAT IASI v6 precipitable water vapor using ground-based GPS SuomiNet stations, *J. Geophys. Res. Atmos.*, *121*, 8925–8948, doi:10.1002/2016JD024806.

Roman, M. O. et al. (2009), The MODIS (collection v005) BRDF/ALBEDO product: Assessment of spatial representativeness over forested landscapes, *Remote Sens. Environ.*, *113*(11), 2476–2498, doi:10.1016/j.rse.2009.07.009.

Scholze, M., T. Kaminski, W. Knorr, S. Blessing, M. Voßbeck, J. P. Grant, and K. Scipal (2016), Simultaneous assimilation of SMOS soil moisture and atmospheric CO_2 in-situ observations to constrain the global terrestrial carbon cycle, *Remote Sens. Environ.*, *180*, 334–345.

Schröder, M., M. Lockhoff, J. M. Forsythe, H. Q. Cronk, T. H. V. Haar, and R. Bennartz (2016a), The GEWEX water vapor assessment: Results from intercomparison, trend, and homogeneity analysis of total column water vapor, *J. Appl. Meteorol. Climatol.*, *55*(7), 1633–1649, doi:10.1175/JAMC-D-15-0304.1.

Schröder, M., et al. (2016b), GEWEX water vapor assessment (G-VAP)—Final Report, World Climate Research Programme, GEWEX, Wash. [Available at http://gewex-vap.org.]

Sofieva, V. F., et al. (2014), Validation of GOMOS ozone precision estimates in the stratosphere, *Atmos. Meas. Tech.*, *7*(7), 2147–2158, doi:10.5194/amt-7-2147-2014.

Stevenson, S., B. Fox-Kemper, M. Jochum, B. Rajagopalan, and S. Yeager (2010), ENSO model validation using wavelet probability analysis, *J. Clim.*, *23*, 5540–5547, doi:10.1175/2010JCLI3609.1.

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

Su, C.-H., D. Ryu, A. W. Western, and W. Wagner (2013), De-noising of passive and active microwave satellite soil moisture time series, *Geophys. Res. Lett.*, *40*, 3624–3639, doi:10.1002/grl.50695.

Su, C.-H., D. Ryu, W. T. Crow, and A. W. Western (2014a), Beyond triple collocation: Applications to soil moisture monitoring, *J. Geophys. Res. Atmos.*, *119*, 6419–6439, doi:10.1002/2013JD021043.

Su, C.-H., D. Ryu, W. T. Crow, and A. W. Western (2014b), Stand-alone error characterisation of microwave satellite soil moisture using a Fourier method, *Remote Sens. Environ.*, *154*, 115–126, doi:10.1016/j.rse.2014.08.014.

Taylor, K. E. (2001), Summarizing multiple aspects of model performance in a single diagram, *J. Geophys. Res.*, *106*, 7183–7192, doi:10.1029/2000JD900719.

Tian, Y., Y. Liu, K. R. Arsenault, and A. Behrangi (2014), A new approach to satellite-based estimation of precipitation over snow cover, *Int. J. Remote Sens.*, *35*(13), 4940–4951, doi:10.1080/01431161.2014.930208.

Tiao, G. C., G. C. Reinsel, D. Xu, J. H. Pedrick, X. Zhu, A. J. Miller, J. J. DeLuisi, C. L. Mateer, and D. J. Wuebbles (1990), Effects of autocorrelation and temporal sampling schemes on estimates of trend and spatial correlation, *J. Geophys. Res.*, *95*(D12), 20,507–20,517, doi:10.1029/JD095iD12p20507.

Tobin, D. C., H. E. Revercomb, R. O. Knuteson, B. M. Lesht, L. L. Strow, S. E. Hannon, W. F. Feltz, L. A. Moy, E. J. Fetzer, and T. S. Cress (2006), Atmospheric radiation measurement site atmospheric state best estimates for atmospheric infrared sounder temperature and water vapor retrieval validation, *J. Geophys. Res.*, *111*, D09S14, doi:10.1029/2005JD006103.

Torrence, C., and G. P. Compo (1998), A practical guide to wavelet analysis, *Bull. Am. Meteorol. Soc.*, *79*, 61–78, doi:10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2.

Tsonis, A. A., and P. J. Roebber (2004), The architecture of the climate network, *Phys. A*, *333*, 497–504.

Turk, F. J., R. Sikhakolli, P. Kirstetter, and S. L. Durden (2015), Exploiting over-land Oceansat-2 scatterometer observations to capture short-period time-integrated precipitation, *J. Hydrometeorol.*, *16*(6), 2519–2535, doi:10.1175/JHM-D-15-0046.1.

Tuttle, S. E., and G. D. Salvucci (2014), A new approach for validating satellite estimates of soil moisture using large-scale precipitation: Comparing AMSR-E products, *Remote Sens. Environ.*, *142*, 207–222, doi:10.1016/j.rse.2013.12.002.

Venema, V. K. C., et al. (2012), Benchmarking homogenization algorithms for monthly data, *Clim. Past*, *8*(1), 89–115, doi:10.5194/cp-8-89-2012.

Verbesselt, J., R. Hyndman, G. Newnham, and D. Culvenor (2010), Detecting trend and seasonal changes in satellite image time series, *Remote Sens. Environ.*, *114*(1), 106–115, doi:10.1016/j.rse.2009.08.014.

Verhoelst, T., J. Granville, F. Hendrick, U. Köhler, C. Lerot, J.-P. Pommereau, A. Redondas, M. Van Roozendael, and J.-C. Lambert (2015), Metrology of ground-based satellite validation: Co-location mismatch and smoothing issues of total ozone comparisons, *Atmos. Meas. Tech.*, *8*(12), 5039–5062, doi:10.5194/amt-8-5039-2015.

von Storch, H., and F. W. Zwiers (2003), *Statistical Analysis in Climate Research*, Cambridge Univ. Press, Cambridge, U. K.

Wanders, N., M. Pan, and E. F. Wood (2015), Correction of real-time satellite precipitation with multi-sensor satellite observations of land surface variables, *Remote Sens. Environ.*, *160*, 206–221, doi:10.1016/j.rse.2015.01.016.

Wang, X. L. (2008a), Accounting for autocorrelation in detecting mean shifts in climate data series using the penalized maximal T or F test, *J. Appl. Meteorol. Climatol.*, *47*(9), 2423–2444, doi:10.1175/2008JAMC1741.1.

Wang, X. L. (2008b), Penalized maximal F test for detecting undocumented mean shift without trend change, *J. Atmos. Oceanic Technol.*, *25*(3), 368–384, doi:10.1175/2007JTECHA982.1.

Weatherhead, E. C., et al. (1998), Factors affecting the detection of trends: Statistical considerations and applications to environmental data, *J. Geophys. Res.*., *103*, 17,149–17,161, doi:10.1029/98JD00995.

Weiss, M., et al. (2014), On Line Validation Exercise (OLIVE): A web based service for the validation of medium resolution land products. Application to FAPAR Products, *Remote Sens.*, *6*, 4190–4216.

Widlowski, J.-L., et al. (2007), Third Radiation Transfer Model Intercomparison (RAMI) exercise: Documenting progress in canopy reflectance models, *J. Geophys. Res.*, *112*, D09111, doi:10.1029/2006JD007821.

Widlowski, J.-L., et al. (2011), RAMI4PILPS: An intercomparison of formulations for the partitioning of solar radiation in land surface models, *J. Geophys. Res.*, *116*, G02019, doi:10.1029/2010JG001511.

World Meteorological Organization (WMO) (2011), Systematic observation requirements for satellite-based data products for climate modelling: Supplemental details to the satellite-based component of the implementation plan for the global observing system for climate in support of the UNFCCC (2010 Update). GCOS-154, Geneva, Switzerland.

World Meteorological Organization (WMO) (2016), The global observing system for climate: Implementation needs, Geneva, Switzerland.

Woodruff, S. D., H. F. Diaz, J. D. Elms, and S. J. Worley (2011), COADS release 2 data and metadata enhancements for improvements of marine surface flux fields, *Phys. Chem. Earth*, *23*, 517–526, doi:10.1016/S0079-1946(98)00064-0.

Yamasaki, K., A. Gozolchiani, and S. Havlin (2009), Climate networks based on phase synchronization analysis track El-Niño, *Prog. Theor. Phys. Suppl.*, *179*, 178–188, doi:10.1143/PTPS.179.178.

Yoo, C. (2002), A ground validation problem of remotely sensed soil moisture data, *Stochastic Environ. Res. Risk Assess.*, *16*, 175–187, doi:10.1007/s00477-002-0092-6.

Yost, F. R. (2016), Sharing the data: The information policies of NOAA and EUMETSAT, *IFLA J.*, *42*(1), 5–15, doi:10.1177/0340035215611135.

Zeng, Y., et al. (2015), Analysis of current validation practices in Europe for space-based climate data records of essential climate variables, *Int. J. Appl. Earth Observ. Geoinform.*, *42*, 150–161, doi:10.1016/j.jag.2015.06.006.

Zhang, X., and F. W. Zwiers (2004), Comment on applicability of prewhitening to eliminate the influence of serial correlation on the Mann-Kendall test by Sheng Yue and Chun Yuan Wang, *Water Resour. Res.*, *40*, W03805, doi:10.1029/2003WR002073.