Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/envsci

Conformity testing of satellite-derived quantitative surface variables

Jean-Luc Widlowski *

European Commission, DG Joint Research Centre (JRC), Institute for Environment and Sustainability, Land Resources Management Unit, Via E. Fermi 2749, 21027 Ispra (VA), Italy

ARTICLE INFO

Article history:
Available online 4 May 2015

Keywords:
Conformity testing
ECV
Quality assurance
GCOS compliance
Environmental regulations
Earth observation
Climate change
COPERNICUS

ABSTRACT

Reliable compliance information of quantitative earth observation (EO) products is a prerequisite for future usages of satellite-derived evidence in (1) regulatory initiatives addressing air quality, development aid, climate risk, agricultural subsidies and the state of the environment among others, (2) liability debates between customers and providers of value-added (quantitative) EO products and services, and (3) auditing efforts and/or contractual negotiations for the operational exploitation of EO data. Irrespective of context, the conformity of an item can only be established with respect to permissible deviations from an agreed reference. The uncertainty of the reference should ideally be smaller than that of the candidate item, and their combined uncertainty should be smaller than the width of the interval defining permissible deviations. While these considerations are an integral part of conformity testing in legal metrology they are not yet included in validation efforts of satellite-derived quantitative surface information. Outside of scientific application contexts, however, the certified compliance of quantitative earth observation products is likely to induce new usages of such information in commercial, judiciary and regulatory contexts. This contribution introduces conformity testing and compares it to validation efforts assessing the value of biophysical EO products with respect to the quality objectives provided by the global climate observing system (GCOS). The findings suggest that, (1) current GCOS quality objectives must be complemented before they may serve as unambiguous requirements for conformity testing of EO products, (2) a consensus on the choice of decision rules must be sought (between data providers and users) since this has a direct impact on what is deemed compliant, and (3) the uncertainty associated with current field validation methods for quantitative biophysical variables is presumably too large to meet the ISO-13528 criteria. The latter thus challenges the eligibility of current field validation methods to provide the reference needed in efforts assessing GCOS compliance of third party EO datasets.

1. Introduction

Satellite remote sensing enables a regular monitoring of our planet's surface from continental to global scales. So-called *retrieval algorithms* convert calibrated satellite measurements into variables describing the chemical and physical properties of the Earth's surface and its constituents. Such earth observation (EO) products are ideally suited to identify, monitor and analyze key environmental variables and

* Tel.: +39 0332 789663.
E-mail addresses: Jean-Luc.Widlowski@jrc.ec.europa.eu, widloje@yahoo.com.
http://dx.doi.org/10.1016/j.envsci.2015.03.018
1462-9011/

processes, as well as potentially hazardous transnational events, be they related to volcanic ash-clouds, atmospheric pollutants, droughts, floods, or other abnormal patterns in seasonal signatures. The quality of quantitative EO information is currently not assessed with possible regulatory applications in mind. In fact, current space law appears rather vague when it comes to the responsibility that service providers must assume to ensure the quality of satellite-derived information products [Ito, 2011]. So far, the use of remote sensing data in legal contexts focuses primarily on its ability to delineate the spatial extent of certain land cover classes (e.g., roads, forests, wetlands, etc.) or to detect temporal changes [De Leeuw et al., 2010; Mayer and Lopez, 2011], rather than it being a source of trustworthy reference data, for example, to challenge infringement procedures or to enforce environmental directives that hinge on the level(s) of specific physical and/or chemical quantities.

Quantitative EO products have the advantage that they can be potentially validated against calibrated in situ measurements rather than through a process of subjective human interpretations. Satellite-derived concentrations of particulate matter (e.g., PM_{10}), ozone, and nitrogen oxides, for example, can be compared against measurements from traditional reference networks in view of supporting European legislation on air quality [WWW-1]. Similarly, satellite-derived quantitative biophysical information can be validated through intensive field campaigns such as to ensure their reliability prior to informing farmers and decision makers about the health and anticipated yields of crops, and – on a larger scale also – the emergence of droughts and food security situations. In addition, quantitative EO products can also be used in the verification, initialization and improvement of numerical prediction models. This is especially relevant for short term climate forecasts where poorly known initial conditions are the main source of uncertainty [Cox and Stephenson, 2007]. Last but not least, the use of quantitative EO products is also likely to increase in commercial applications, among others, for targeted estimates of risks and policy costs by the insurance sector.

At present, the accuracy of quantitative satellite-derived surface information has not yet become part of liability debates. However, given the increasingly prominent role that long term records of quantitative EO products assume in the generation of reference datasets for climate models [Dowell et al., 2013] as well as in efforts to attribute the causes of extreme events, it is likely that the quality of these datasets will come under increasing scrutiny in the future. Similarly, the anticipated uptake of quantitative EO information by the private sector, whether in the context of crop yield forecasts, air quality warnings, or other value-added services [COM-312, 2013], is also likely to bring about a discussion on the responsibility (and liability) of EO service providers (and contributing scientists) as to the quality of these information. Finally, the public (as well as any relevant funding and auditing authorities) should have a means to verify/know that the results of costly EO programmes are trustworthy and comply with predefined quality requirements.

Methods to ensure compliance with quality criteria exist for quite some time already in legal metrology and the manufacturing sector. Logically, any such endeavour must start with an unambiguous definition of the target item/quantity itself. For satellite-derived quantitative surface information, this is often far from trivial due to sensor-specific retrieval algorithms using inherent assumptions and shortcuts. Next, it must be possible to have access to unbiased candidate and reference estimates of that target quantity, as well as, reliable descriptors of the uncertainty associated with these. Again, this may be far from trivial in the EO context, especially if the spatial heterogeneity of the target quantity is large within the nominal field of view of the observing satellite sensor or else changes rapidly in time. In a final step, the compliance of the candidate method/dataset must be evaluated against clearly defined quality requirements. In a regulatory context, the exact specification and wording of the compliance criteria (as well as the procedures to assess these) is typically the result of a negotiation process between (governmental, industrial, private and scientific/expert) stakeholders and may become rather involved. In operational EO contexts, it is the mission requirements issued by the relevant space agencies or, more generally, the quality objectives formulated by international scientific bodies that are used.

The development of reliable reference methods for satellite-derived information products is actively pursued by the working group on calibration and validation (WGCV) within the Committee for Earth Observation Satellites (CEOS). In doing so, CEOS WGCV focuses on a series of so called "essential climate variables" (ECVs) given that many of these quantities are also relevant in contexts other than climate change. The concept of ECV was developed in the 1990s by the Global Climate Observing System (GCOS) in collaboration with user communities and other stakeholders [Bojinski et al., 2014]. Since then, GCOS publishes at regular intervals implementation plans (and satellite supplements) that provide detailed descriptions of the growing number of ECVs as well as the quality objectives that these (satellite-derived) ECV products should ideally satisfy if they are "to be of relevance" to the work of both UNFCCC and IPCC [e.g., GCOS-138, 2010; GCOS-154, 2011]. Over the past few years, the GCOS quality objectives – while intended as high-level programmatic guidance – have become the *de facto* reference criteria for validation efforts of biophysical surface ECVs. While this choice may appear surprising at first, it is a consequence of the general absence of detailed compliance criteria, on the one hand, and the advantages that a GCOS parentage offers with respect to ad hoc quality assurance efforts on the other hand. Perhaps most pertinent from the perspective of the validation community are the facts that the GCOS quality objectives (1) are regularly updated, (2) undergo public consultation, and (3) provide increasingly detailed definitions of the target quantities.

Despite much progress in recent years, trustable and ideally also SI-traceable evidence as to the quality of the retrieved information is still lacking for most satellite-derived quantitative EO products over land. In part, this is due to the complex, multi-stage retrieval process of biophysical ECV estimates from optical remote sensing data (whether acquired by satellites, from observation towers or with hand-held devices in the field). At the same time, it is also clear that EO product compliance cannot be demonstrated conclusively if the quality requirements and decision rules needed to assess

conformity are incomplete or ambiguous. In particular the quality of the reference data is of relevance in this context here, since many field validation efforts do not measure the target biophysical quantity per se but rather infer its value from observations of third party quantities, e.g., [Morisette et al., 2005].

This contribution compares the current GCOS-inspired quality assurance approach for quantitative EO products (over land) with formal conformity testing methods used in legal metrology as well as by the manufacturing sector. More specifically, Section 2 introduces conformity testing and the role that decision rules play when assessing the compliance of items with respect to predefined quality criteria; Section 3 compares these approaches to validation schemes that make use of the GCOS quality criteria for biophysical ECVs. Since the uncertainty of the reference is crucial, Section 4 will focus on eligibility criteria that field validation protocols must satisfy in order to qualify as provider of reference data in conformity testing efforts of third party EO products. Section 5 then highlights some of the consequences of working with imprecise reference methods, while Section 6 summarizes all findings and comments on some of their implications.

2. Conformity testing

Quality objectives are the criteria needed to assess compliance of an entity[1] with respect to customer or regulatory requirements. Conformity testing then is the process that determines whether the estimated target quantity falls within the range of tolerable values or not. Overall, such efforts can be conceptualized into a *decision making* and an *estimation* problem [EPA, 2006]. At best, conformity testing thus accounts for the uncertainty associated with the available data, as well as the risk related to making the wrong acceptance or rejection decision [Eurachem, 2007].

This contribution is neither concerned with the estimation of the target quantity nor with the quantification of the "doubt" that is associated with the validity of this measurement/retrieval process. While the former is the object of much research in the remote sensing community, the latter may be addressed using (1) a *modelling* or *bottom up* approach [JCGM-100, 2008], where a comprehensive mathematical model of the measurement/retrieval process is used to determine the uncertainty contributions of each one of the influencing variables and conditions, and/or (2) a *collaborative study* or *top down* approach [ISO-5725, 1994], where comprehensive intercomparison activities are carried out to obtain estimates of the so called repeatability, trueness and reproducibility standard deviations. In theory both approaches should yield comparable uncertainty estimates. In practice, however, this is not always the case due to (1) incomplete mathematical models (i.e., the presence of unknown effects), and (2) incomplete or unrepresentative variations of all influences during reproducibility assessment [ISO-21748, 2010]. For the purpose of this contribution it is assumed that reliable uncertainty estimates are available.

2.1. Decision rules

Any future use of quantitative EO information within regulatory contexts would benefit from quality requirements that are formulated in a manner such as to enable conformity testing according to established procedures [e.g., JCGM-106, 2012]. In environmental regulations, this typically involves the specification of a *limit value* or else a *tolerance interval* (bounded by two limit values) in order to define the permissible range of the target quantity. A number of sometimes elaborate *decision rules* are then used as a prescription for the acceptance or rejection of an entity with respect to the predefined limit(s). Fig. 1 provides a schematic overview of decision rules commonly used to classify measurement results as being conform (green colour), non-conform (red colour), or inconclusive (orange colour) with respect to a predefined tolerance interval bounded by an upper (T_U) and lower (T_L) limit value.

If the uncertainty associated with the target quantity estimate is negligible (top left panel in Fig. 1) then all data points that fall within the tolerance interval are deemed acceptable while those located outside are not [IEC-115, 2007]. If a descriptor of the uncertainty of the target quantity estimate is available (for example in the form of a coverage interval and associated probability[2]) then conformity can be asserted if the coverage interval (at an appropriate coverage probability) is fully contained within the tolerance interval [ISO-10576, 2003]. Should the coverage interval cross one or both of the limit values then neither conformity nor non-conformity can be attested and more data may be required instead (2 stage assessment in ISO-10576). Finally, if the probability density function (PDF) of the estimator of the target quantity is available – for example through a Monte Carlo propagation scheme [JCGM-101, 2008] – then acceptance decisions can be made by comparing the conformance probability (P_C) of the (possibly non-Gaussian) PDF of the quantity (x) to a predefined minimum required compliance level (C_L) [JCGM-106, 2012] where the conformance probability of the best estimate of x (i.e., $\bar{x}$) is defined as:

$$P_c(\bar{x}) = \int_{T_L}^{T_U} \mathrm{PDF}(x)dx \quad (1)$$

If $P_C \geq C_L$ then the item is accepted as being conform with the requirements (compare with top right panel in Fig. 1). Alternatively, the exceedance probability $P_E = 1 - P_C$, which corresponds to that part of the PDF of x that falls outside of the tolerance interval $[T_L, T_U]$, may be compared against a

[1] An entity can be anything from an individual item, to a group of items, or even a service, provided that it possesses quantifiable characteristics [Källgren et al., 2003].

[2] If the probability density function (PDF) of likely values of the estimated quantity x is symmetric and unimodal, then the widely-used coverage interval $[\bar{x} - U; \bar{x} + U]$ has length 2 U and is centred on the best estimate $\bar{x}$[JCGM-106, 2012]. The "expanded uncertainty" $U(x) = k \cdot u(x)$ is obtained by multiplying the combined standard uncertainty $u(x)$ with a coverage factor k [JCGM-100, 2008]. The relationship between the coverage factor k and the associated coverage probability depends on the PDF of x. For normal PDFs $k = 1$ relates to a 68.27% coverage probability while $k = 2$ relates to 95.45% coverage probability.

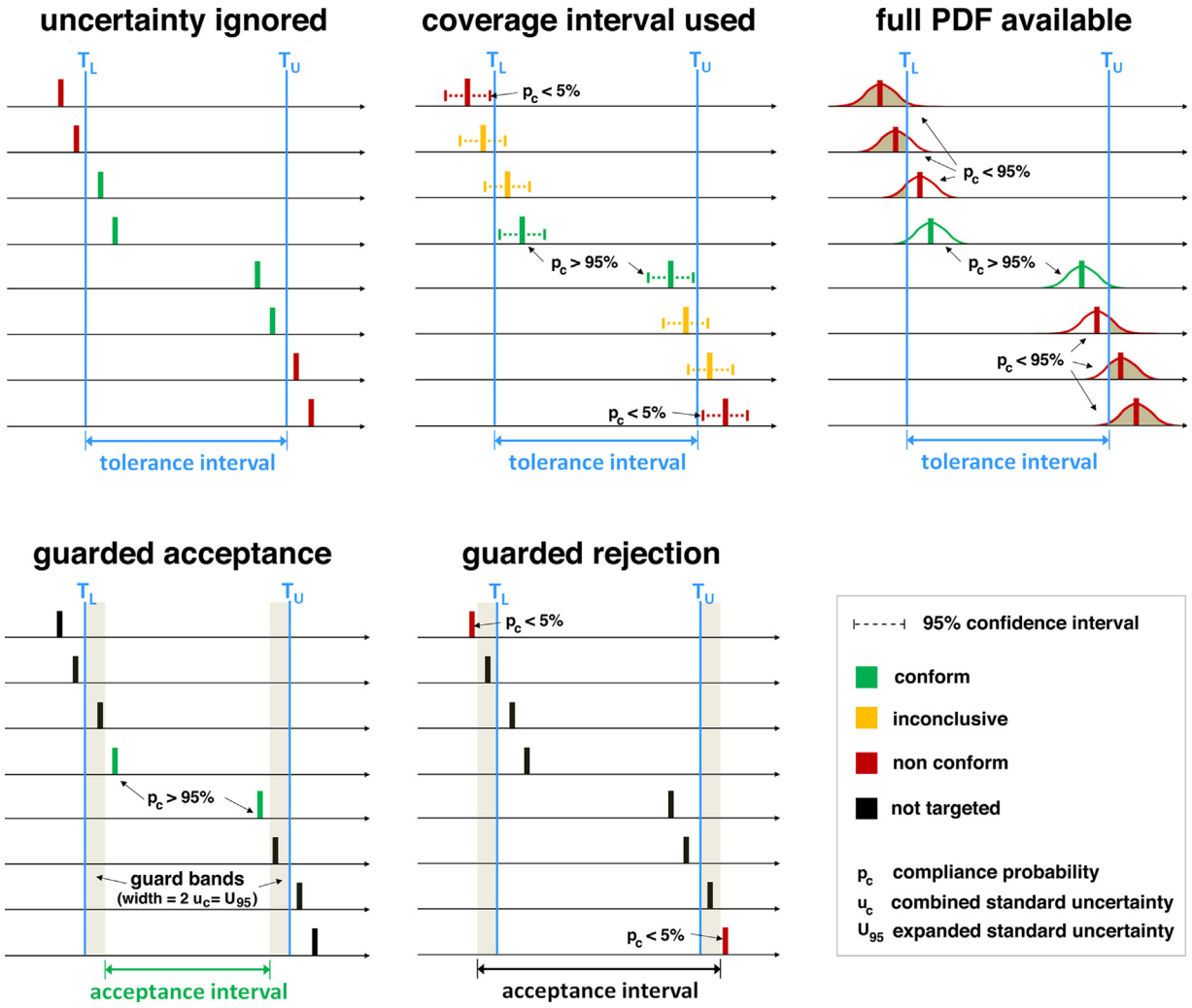

Fig. 1 – Schematic overview of the decision rules used by conformity testing approaches to classify measurement results as being conform (green) or non-conform (red) with respect to a tolerance interval defined by an upper (T_U) and lower (T_L) limit value. The top row displays conformity testing schemes where the uncertainty is (1) negligible [procedure 2 in IEC-115, 2007], (2) described by a coverage factor with an associated uncertainty interval (or equivalent statistics) [ISO-10576, 2003], and (3) characterized by a probability density function (PDF) [JCGM-106, 2012]. The bottom panels present the concept of guard bands that are used to reduce the risk of either accepting a non-conforming entity or else of rejecting a conforming one [e.g., ISO-14253, 1998; ASME-B89731, 2001]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

maximum permissible exceedance level (E_L). If $P_E \leq E_L$ then the item is accepted as being conform to the requirements.

In cases where the uncertainty of the target quantity estimator is assumed constant, so called *guard bands* can be constructed in order to reduce the risk of accepting a non-conforming entity [e.g., ISO-14253, 1998; ASME-B89731, 2001]. The width of the guard band is often set to the expanded uncertainty U having a coverage factor $k = 2$ or else as required by the specifications. A *guarded acceptance* rule (bottom left panel in Fig. 1) is used if clear evidence is needed before accepting an entity as being conform. A *guarded rejection* rule, on the other hand, is employed when clear evidence that a limit has been exceeded is wanted before any negative action is taken [JCGM-106, 2012]. Appendix A of Eurachem (2007) indicates how the size of guard bands can be computed depending on the available uncertainty information.

In essence, the guarded acceptance approach reduces the tolerance interval by a predefined width, g. A shared risk approach is then applied within the resulting *acceptance interval* (i.e., within $\pm|T_U - g|$ if $T_L = -T_U$). Typically, the width of the guard bands is set to a multiple of the combined standard uncertainty, e.g., $g = 2{\cdot}u = U_{95}$. If the width of the guard bands is identical to the coverage interval of items assessed with a coverage interval method then the conformance results of these two decision rules will be identical. By the same argument, the guarded rejection approach will also yield identical non-conformance results if its guard bands are identical to the coverage interval of items tested with a coverage interval method.

Among the various decision rules, a *shared risk* approach (shown in the top left panel of Fig. 1) is highly desirable since it simplifies the acceptance/rejection process. In a shared risk

approach measurement uncertainty is not included, that is, the measurement result is accepted as conforming if it falls within the tolerance interval and rejected otherwise [JCGM-106, 2012]. As its name implies, any consequences of incorrect conformity decision must be shared between consumers and producers alike. However, in order to keep the risks of false rejection and acceptance decisions within reasonable levels a shared risk approach should only be applied if the uncertainty of the measured value is "small enough" as discussed in the next subsection.

2.2. *Maximum permissible error and uncertainty*

In contexts where the focus is on *deviations* from a qualified reference rather than the true value of the target quantity, the above tolerance interval will be centred at zero and fenced in by a *maximum permissible error* (MPE) [e.g., WELMEC, 2006]. This is schematically shown in the upper panel of Fig. 2. The qualified reference is typically provided by a method that was calibrated against high quality reference standards and corrected for systematic biases. However, since the true value of the target quantity is unknowable [JCGM-100, 2008], the *actual* error of the reference method can not be known. The difference between the best estimates obtained with a reference method (R) and a candidate method (Q) thus will always result in an *apparent* error $\varepsilon = Q - R$ (compare with the bottom panel in Fig. 2). The probability distribution of the likely values of ε is given by the convolution of the PDFs describing the likely values of the reference and candidate estimates. If the latter were both Gaussian with a combined standard uncertainty equal to u_R and u_Q, respectively, then the PDF of the likely values of ε would also be Gaussian with a variance given by:

$$u_\varepsilon^2 = u_R^2 + u_Q^2 \tag{2}$$

In some parts of legal metrology, guard bands and acceptance intervals tend to be expressed in terms of a fixed *maximum permissible uncertainty* (MPU) level [e.g., Källgren et al., 2003]. Eurachem (2007) specifies that the "maximum permissible size of the uncertainty should form part of the definition of the [compliance] requirement". By convention the MPU is set to MPE/3 for verification (and MPE/5 for type testing) purposes [e.g., Sommer and Kochsiek, 2002]. In its most relaxed form this empirical threshold thus implies that:

$$2u_\varepsilon = U_{95} \leq \mathrm{MPU} = \frac{\mathrm{MPE}}{3} \tag{3}$$

Another "rule of thumb" that is often used is the 4:1 tolerable uncertainty ratio (TUR) [ASME-B89731, 2001] which can be expressed as:

$$2u_\varepsilon = U_{95} \leq \frac{T_U - T_L}{8} = \frac{\mathrm{MPE}}{4} \tag{4}$$

Yet further rules and approaches exist that in the end appear all somewhat arbitrary [Pendrill, 2007]. In fact, using the measurement capability index, $C_m = \mathrm{MPE}/U_{95}$ it can be shown that the part of the tolerance interval ($T_U - T_L$) which

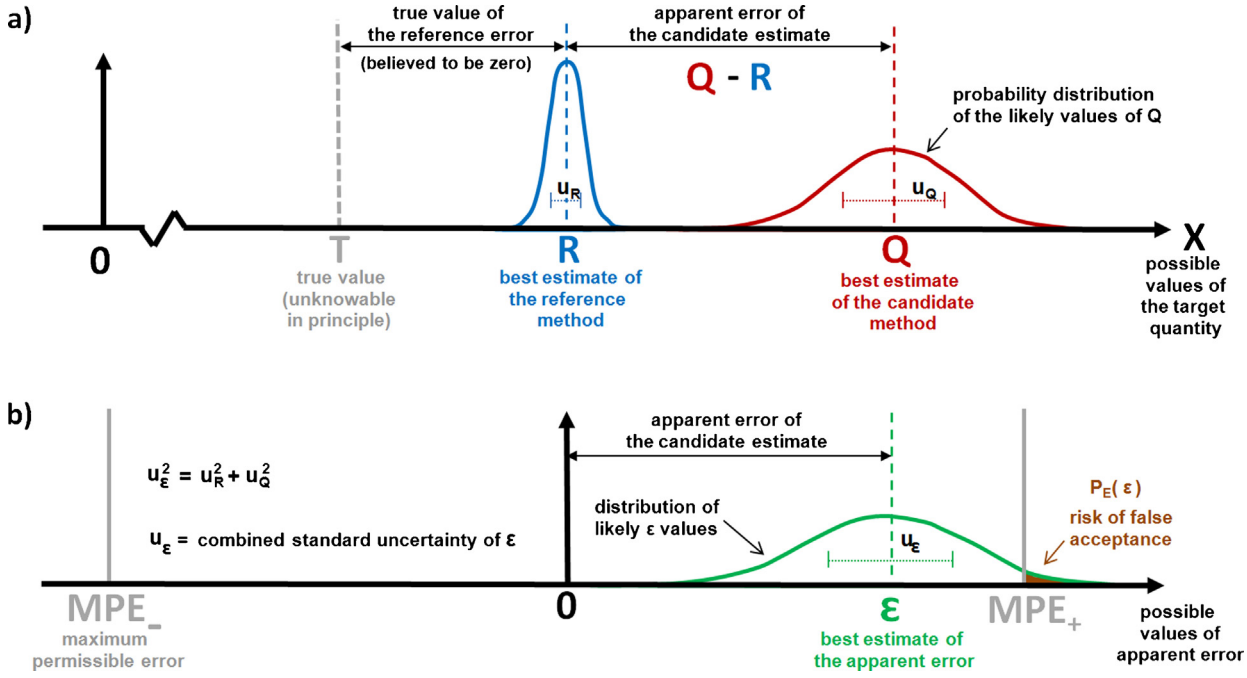

Fig. 2 – Individual estimation effort of the target quantity: the top panel provides a schematic representation of how the unknowable true value (T) of the target quantity (X) relates to the best estimate obtained with a reference (R) and a candidate (Q) method. The uncertainty in the best estimate (of R and Q, respectively) is best described with a probability distribution function (PDF) that may be summarized by a combined standard uncertainty (u_R and u_Q, respectively). The bottom panel shows how the resulting apparent error ($\varepsilon = Q - R$) relates to a predefined tolerance interval (±MPE). Also shown is the combined standard uncertainty (u_ε) associated with this individual apparent error estimate, as well as the probability P_E that ε exceeds ±MPE, i.e., the risk of false acceptances.

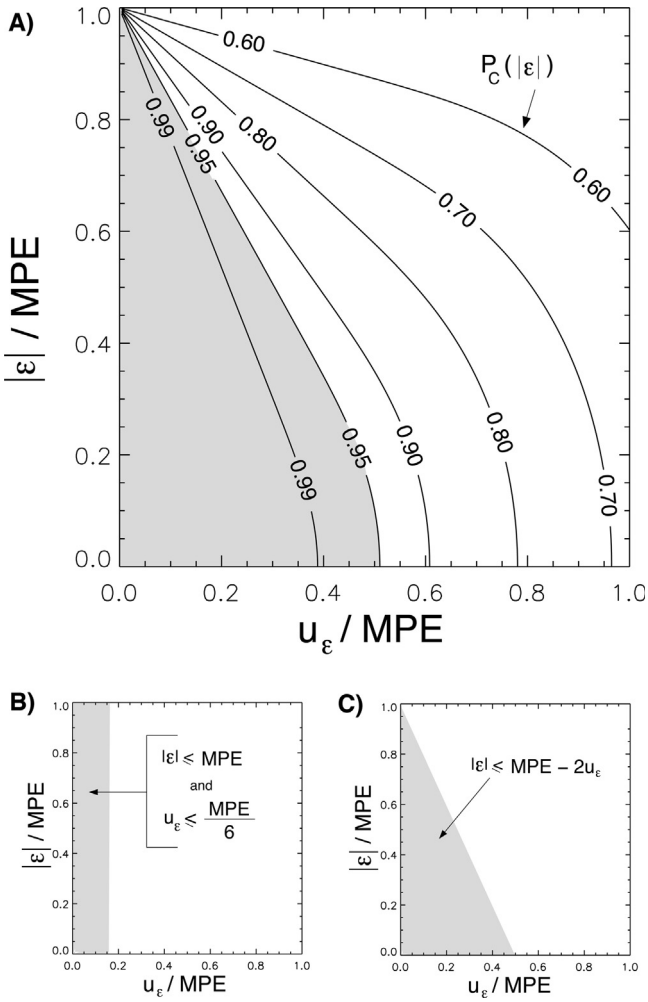

Fig. 3 – Panel A shows contour lines of the probability that a Gaussian PDF falls within the tolerance interval – defined by ±MPE – if the mean (ε) and standard deviation (u_ε) of the PDF are varied from 0 to MPE. More specifically, isolines of the compliance probability $P_C(\varepsilon)$ are mapped as a function of the normalized apparent error (ε/MPE) and its associated combined standard uncertainty (u_ε/MPE) assuming a normal distribution of the likely values of ε. The grey area in panel A indicates those pairs of apparent error and associated combined standard uncertainty that result in $P_C(\varepsilon) \geq 95\%$. The grey area in panel B shows those pairs of ε/MPE and u_ε/MPE that are acceptable in a shared risk approach requiring u_ε to be smaller than MPE/6. The grey area in panel C shows the pairs of ε/MPE and u_ε/MPE that are acceptable in a guarded acceptance approach if the width of the guard band is set to $2u_\varepsilon$.

effectively allows to demonstrate conformity (say with a 95% confidence level) covers only the central 10% of $(T_U - T_L)$ if $C_m = 1$, while it includes the central 79% if $C_m = 4$ and 91% if $C_m = 10$ [JCGM-106, 2012].

Despite their somewhat arbitrary nature, the inequalities in Eqs. (3) and (4) are often used to gauge whether more data or a better estimation method are needed [ISO-10576, 2003], or, whether the MPE values should be increased [OIML, 2009]. WELMEC (2006), for example, recommends that decision making in conformity assessments may follow the *shared risk* principle under the condition that Eq. (3) is satisfied, i.e., the MPU is less than or equal to 1/3 of the MPE. While this removes the need to characterize the PDF of likely values of ε for individual measurements/retrievals it still requires access to the combined standard uncertainty of both the reference and the candidate methods under a large ensemble of test conditions in order to evaluate whether the use of a shared risk approach is actually meaningful.

For a visual comparison of the impact of different decision rules, the grey areas in Fig. 3 highlight those apparent error (ε) and associated uncertainty (u_ε) combinations that – in the case of a Gaussian PDF of likely values – result in a positive outcome when subjected to different types of conformity testing. Panel A, for example, shows how the compliance probability $P_C(\varepsilon)$ varies across a space defined by (normalized[3]) values of the best apparent error estimate ($\varepsilon = Q - R$) and its associated combined standard uncertainty (u_ε). More specifically, the

[3] Many graphs in this contribution appear with axes that are normalized with respect to the MPE. This is purely for ease of representation and does not imply that the normalized (rather than the original) PDFs or populations are Gaussian.

contour lines in panel A relate to the fraction of the Gaussian PDF, defined by $\mathcal{N}(\varepsilon/u_\varepsilon)$, that falls within the tolerance interval defined by ±MPE. The grey shaded area delineates the set of apparent error and associated standard uncertainty characteristics that result in compliance probabilities greater than or equal to 95%.

Within the grey area of panel A the largest possible value that the combined standard uncertainty of the apparent error ($\hat{u}_\varepsilon$) may assume occurs when ε is zero. As can be seen, $\hat{u}_\varepsilon$ amounts to a mere 51.5% of the MPE. If the magnitude of ε is increased, then the width of its associated PDF must decrease in order to satisfy the condition of $P_C(\varepsilon) \geq 0.95$. This is why all contour lines in panel A converge (and the grey area thins out) as $\varepsilon \rightarrow$ MPE. If one were now to impose also the requirement of Eq. (3), i.e., $u_\varepsilon \leq 0.1667 \cdot$MPE, then a large part of the grey area in panel A would be no longer compliant and for values of ε/MPE > 0.73 the standard uncertainty of the apparent error u_ε would have to become smaller than 0.1667 to maintain a 95% compliance.

While panel A was constructed on the basis of a full knowledge of the PDF of likely values of ε, the grey area in panel B highlights those pairs of (ε, u_ε) that would be deemed compliant if a shared risk approach – with the u_ε condition of Eq. (3) – is used as decision rule. Without the restriction on u_ε all (ε, u_ε) pairs in panel B would become compliant. Finally, the grey area in panel C refers to a guarded acceptance approach where the guard bands were set to twice the combined standard uncertainty of the apparent error estimate (i.e., to $2u_\varepsilon$). This is equivalent to a coverage interval approach employing a fixed value of $U = 2u_\varepsilon$. Interestingly, the approach in panel C turns out to be more restrictive than that using the full PDF of likely values of ε (in panel A). By comparing the various grey areas in Fig. 3, it thus becomes apparent that the decision rule itself plays a major role in defining the set of apparent error characteristics that will be deemed compliant.

2.3. *Apparent error populations*

In the context of image-based satellite remote sensing where large amounts of pixel-based information are available, it is convenient to work with apparent error *populations*. Any comparison between (in situ) reference and (satellite-derived) candidate estimates of a (biophysical ECV) target quantity will return sample statistics that – if representative – yield expectations of the (mean and variance) parameters of the underlying apparent error population. Different contexts (e.g., biomes, seasons, latitudes, etc.) may give rise to their own populations. In all generality, any population of apparent errors is characterized by a mean value:

$$\mu_\varepsilon = \int_{-\infty}^{+\infty} \varepsilon \cdot \mathrm{PDF}(\varepsilon) \cdot d\varepsilon \approx \langle \bar{\varepsilon} \rangle = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i \tag{5}$$

and a variance:

$$\sigma_{\bar{\varepsilon}}^2 = \int_{-\infty}^{+\infty} |\varepsilon - \langle \bar{\varepsilon} \rangle|^2 \cdot \mathrm{PDF}(\varepsilon) \cdot d\varepsilon \approx s_{\bar{\varepsilon}}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\varepsilon_i - \langle \bar{\varepsilon} \rangle)^2 \tag{6}$$

where N is the number of individual apparent error estimates ($\varepsilon_i = Q_i - R_i$) in the population sample, $\langle \bar{\varepsilon} \rangle$ is the experimental mean apparent error, and $s_{\bar{\varepsilon}}^2$ is the experimental variance. If $\mu_\varepsilon \neq 0$ then a systematic bias is present that may need correcting. Alternatively, if $\sigma_{\bar{\varepsilon}}^2$ is large then it is likely that the retrieval methods are not equally reliable across the range of (viewing, illumination and surface) conditions encountered. Individual apparent error estimates can be drawn at random from these populations. As such, each one of these individual apparent errors comes with an associated uncertainty (u_{ϵ_i}) that is characteristic for the choice of retrieval method used. As far as conformity testing is concerned, it is both the μ_ε and $\sigma_{\bar{\varepsilon}}^2$ parameters of the apparent error population, as well as, the inclusion or non-inclusion of u_{ϵ_i} that cause differences between the various decision rules commonly used in conformity testing. For ease of reading, $\bar{\varepsilon}$ is used instead of the population mean symbol μ_ε below.

Fig. 4 maps the differences in conformance probabilities between a guarded acceptance and a shared risk approach (panel A), as well as between a guarded acceptance and a PDF-based approach (panel B), across a wide range of apparent error population characteristics. For convenience apparent error populations were assumed Gaussian, i.e., $\mathcal{N}(\bar{\varepsilon}, \sigma_{\bar{\varepsilon}})$ where $\bar{\epsilon}$ is the mean and $\sigma_{\bar{\epsilon}}$ the standard deviation. One million values of ε_i were drawn at random from each such population. In the case of the shared risk approach the rate of ε_i ' s falling outside of the tolerance interval defined by ±MPE was determined, i.e., $P(|\epsilon_i| \leq \mathrm{MPE})$. For the guarded risk approach, $P(|\epsilon_i| \leq \mathrm{MPE} - 2u_\epsilon)$ was computed using the same combined standard uncertainty for all ε_i values, i.e., u_ε = MPE/6. Finally, for the PDF-based approach the compliance probability $P_C(\varepsilon_i)$ was computed – assuming a normal distribution $\mathcal{N}(\varepsilon_i, u_\varepsilon)$ of the likely values of a given ε_i – and then compared against a minimum required compliance level C_L = 95.5%. The various panels in Fig. 4 thus show the difference between these three conformity testing approaches when applied across a space of apparent error populations characterized by values of $\bar{\epsilon}$ and $\sigma_{\bar{\epsilon}}$ that range from zero to MPE. The grey area in panels A and B highlights those $\mathcal{N}(\bar{\varepsilon}, \sigma_{\bar{\varepsilon}})$ where differences in the conformance probability is less or equal to 5%.

The first thing to notice in panels A and B of Fig. 4, is that all contour lines are positive. In other words, the conformance probability is always smaller with a guarded acceptance approach than with a shared risk or a PDF-based decision rule. However, the differences between the shared risk and the guarded acceptance methods (panel A) are far larger than those between the guarded acceptance and the PDF-based approaches (panel B). This is easily noticeable by comparing the extent of the grey (≤5% difference) areas in both panels. In fact, for a population of apparent errors to fall within the grey area in panel A, the maximum permissible value of dispersion $\hat{\sigma}_{\bar{\varepsilon}} = 0.34 \cdot$MPE (when $\bar{\epsilon} = 0$). Since u_ε = MPE/6 has been advocated as a threshold criteria for the use of shared risk approaches [WELMEC, 2006], the results in panel A may thus serve as a means to assess the consequences of such a decision.

While the shared risk approach does not depend on the value of u_ε both the guarded acceptance and the PDF-based methods, however, do. In fact, if $\sigma_{\bar{\epsilon}} = 0$ and the value of $\bar{\epsilon}$ is gradually increased then the conformance probability of the guarded acceptance approach will change from 100% to 0% if the location $\bar{\epsilon}/\mathrm{MPE} = 1 - (2u_\epsilon/\mathrm{MPE})$ is crossed. In panel A of

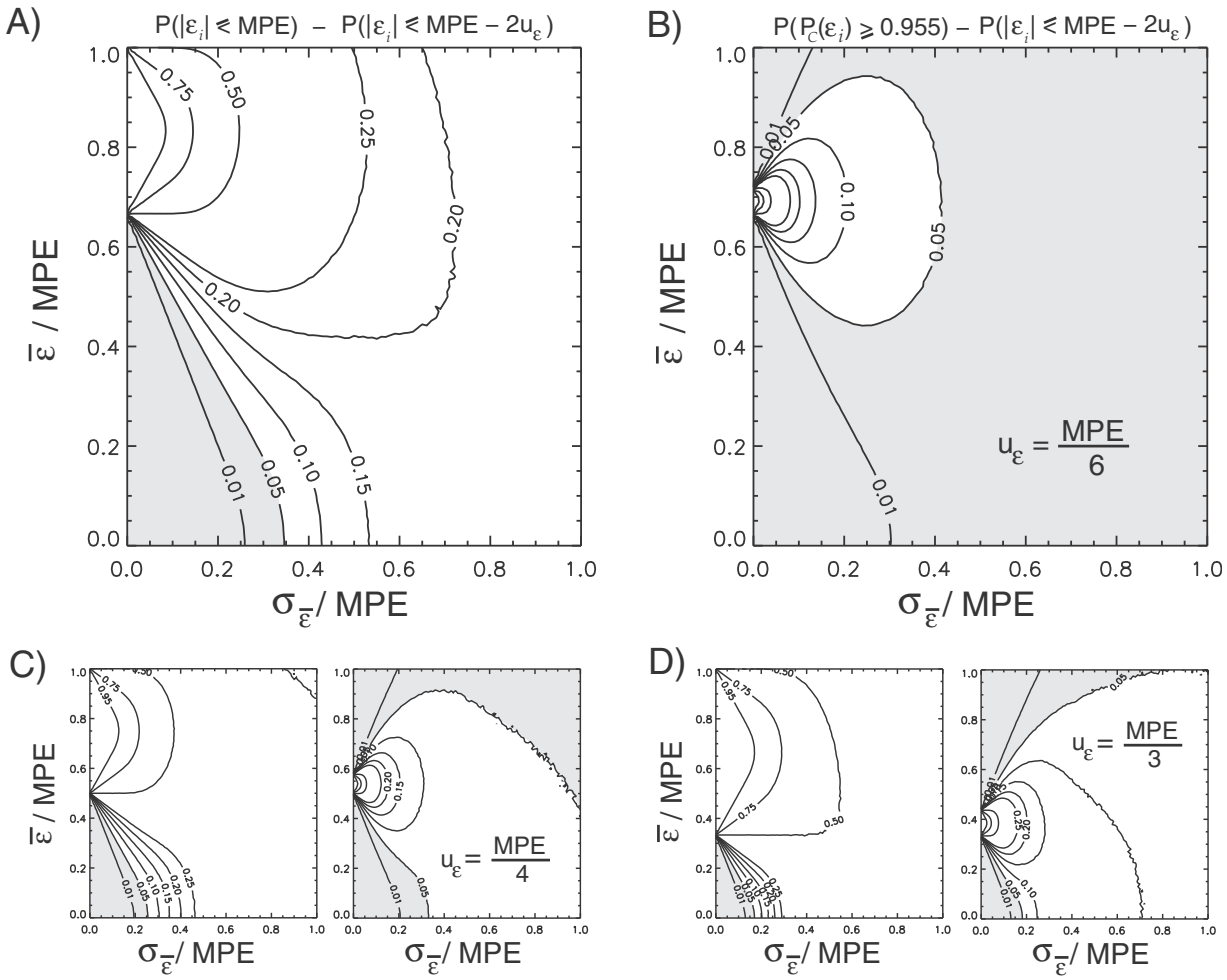

Fig. 4 – Difference in compliance probabilities between a shared risk and guarded acceptance approach (panel A), as well as between a PDF-based and a guarded acceptance approach (panel B), across a wide range of Gaussian apparent error populations. For each such population $\mathbf{N}(\bar{\varepsilon}, \sigma_{\bar{\varepsilon}})$ – where $\bar{\varepsilon}$ is the mean and $\sigma_{\bar{\varepsilon}}$ the standard deviation – one million values of ε_i were drawn at random. In the case of the shared risk approach, $\mathbf{P}(|\epsilon_i| \leq \mathbf{MPE})$ was determined, while for the guarded acceptance approach, $\mathbf{P}(|\varepsilon_i| \leq \mathbf{MPE} - \mathbf{2}u_\varepsilon)$ was computed using a fixed standard uncertainty, i.e., $u_\varepsilon = \mathbf{MPE/6}$. For the PDF-based approach the compliance probability $P_C(\varepsilon_i)$ was computed first – assuming a normal distribution $\mathcal{N}(\varepsilon_i, u_\varepsilon)$ – and then compared against a permissible compliance limit of 95.45%. The grey area highlights those $\mathbf{N}(\bar{\varepsilon}, \sigma_{\bar{\varepsilon}})$ where the compliance difference is less or equal to 5%. Panels C and D show the same plots but for $u_\varepsilon = \mathbf{MPE/4}$ and $u_\varepsilon = \mathbf{MPE/3}$, respectively.

Fig. 4 the value of u_ε was set to MPE/6. Hence, for all values of $\bar{\epsilon}/\mathrm{MPE} \leq 2/3$ the guarded acceptance method will yield $P(|\epsilon_i| \leq \mathrm{MPE} - 2u_\epsilon) = 1$ while for all values $\bar{\epsilon}/\mathrm{MPE} > 2/3$ it will be equal to zero. Since $P(|\epsilon_i| \leq \mathrm{MPE})$ is independent of the value of u_ε the contour lines of the conformance difference between the shared risk and the guarded acceptance methods thus originate from a single point along the ordinate.

If u_ε is set to a value smaller than MPE/6 then the origin of the isolines will move upwards along the ordinate (and vice versa if the value of u_ε is increased). The latter can be seen in panels C and D where u_ε was set to MPE/4 and MPE/3, respectively. Smaller values of u_ε thus will reduce the differences between two decision rules. In general, however, the actual conformance probability will depend both on the value of u_ε and on the properties of the apparent error population.

3. Quality objectives for quantitative biophysical variables

Quality requirements for quantitative EO products depend on the context in which such information is used. In application areas related to weather, water and climate, for example, the World Meteorological Organization (WMO) provides quantitative user-defined requirements addressing (where appropriate) the horizontal and vertical resolution, the observation cycle and timeliness, as well as the stability and uncertainty that observations of physical variables should possess [WWW-2]. More specifically, the WMO requirements distinguish between a minimum (or "threshold") level that is needed for the data to be useful, an ideal (or "goal") level above which further improvements are not necessary, and an

intermediate (or "breakthrough") level which, if achieved, will result in substantial improvements for the targeted application. So far the use of quantitative EO products occurs primarily within scientific application contexts such that available quality objectives are often expressed in a technology-free manner or as aspirational goals that are not necessarily achievable with current observation techniques. This is rather different from regulatory contexts where compliance testing is of the essence. The European air quality directive dealing with particulate matter, for example, stipulates explicitly what measurement technique may serve as the reference method (and by what means other methods must prove that their level of uncertainty is equivalent) [DIRECTIVE, 2008/50/EC]. The directive also indicates the frequency of measurement, the exact definition of the quantity that has to be reported, as well as the precise number of non-compliant cases that may be tolerated in a given interval.

The absence of such detailed quality requirements in remote sensing contexts complicates the validation of satellite-derived quantitative surface products (particularly in a manner that would make this data suitable for inclusion in environmental directives or as evidence in judiciary contexts). At present, the validation community is increasingly using the quality objectives provided by the GCOS as the yard stick to assess the value of quantitative EO products, [e.g., Prieto-Blanco et al., 2009; Verger et al., 2009; Rochdi and Fernandes, 2010; Fang et al., 2012a; Canisius and Fernandes, 2012;Malenovský et al., 2012; Baret et al., 2013; Claverie et al., 2013; D'Odorico et al., 2014; Franch et al., 2014; Posselt et al., 2014; Zibordi et al., 2015]. GCOS has provided valuable guidance to validation efforts by introducing the concept of essential climate variables (ECVs), by attempting to define these quantities in an unambiguous manner, and by regularly updating the accuracy and stability levels that these ECVs should ideally have [GCOS-92, 2004; GCOS-107, 2006; GCOS-138, 2010; GCOS-154, 2011]. The GCOS requirements, however, were not set with a particularly regulatory context or conformity testing framework in mind but rather to provide programmatic high-level guidance [WWW-3]. As such, it is pertinent to revisit these quality objectives and comment on their use in compliance testing efforts for quantitative EO products.

Due to their potential relevance to a variety of regulatory contexts, the focus here will be on quantitative land ECVs that are routinely derived from global satellite observations, namely, (1) the *surface albedo* which controls radiative forcing and thus the planetary energy budget as well as the radiation partitioning between the surface and atmosphere [GTOS-63], (2) the *fraction of absorbed photosynthetically active radiation* (FAPAR) which plays a critical role in the energy balance of ecosystems and the terrestrial part of the carbon cycle [GTOS-65], and (3) the *leaf area index* (LAI) which can be linked to photosynthesis, respiration and rain interception (as well as being correlated with the first two ECVs in this list) [GTOS-66]. These terrestrial ECVs were also chosen because of the complexities associated with their validation, and their relevance with respect to predictions/verifications of crop yields, drought events and reforestation rates with all the consequences that this may have for downstream decision making in the context of international aid, carbon trading and policy formulations.

3.1. GCOS accuracy criteria

One of the activities of GCOS concerns the provision of indicative requirements for both the *accuracy* and *stability* of satellite-derived ECVs. For terrestrial ECVs, the monitoring of "product accuracy" over time enables also conformity testing of "product stability" over invariant test sites. As such, this contribution will focus exclusively on product accuracy – which is defined by GCOS as "closeness between the product values and the true values" where the latter "refer to a locally prevailing reference value" [GCOS-154, 2011]. More specifically, for the terrestrial ECV quantities of interest here, the current GCOS accuracy criteria (Δ) are given as:

- 5% or 0.0025 – whichever is larger – for surface albedo,
- 10% or 0.05 – whichever is larger – for FAPAR,
- 20% or 0.5 – whichever is larger – for LAI.

The tolerable deviation from the reference is thus a constant up to values of 0.05 in the case of the albedo, 0.5 for FAPAR, and 2.5 for LAI. For larger values, the tolerable deviations become an ECV specific fraction of the reference value. The change from a fixed to a relative tolerable deviation occurs at different points along the range of the naturally occurring values for these ECVs, i.e., at 5% of the [0–1] range for albedo, at 50% of the [0–1] range for FAPAR, and at about 15–20% of the range of LAI values. In addition, broadband albedo values fall rarely below 0.05 (even in the absence of snow), while LAI < 2.5 and FAPAR < 0.5 occur rather frequently for most of the land cover classes (except perhaps for broad-leafed and needle-leafed forests) [e.g., Taberner et al., 2010; Cescatti et al., 2012; Fang et al., 2012a; Camacho et al., 2013; Pickett-Heaps et al., 2013]. As a consequence, the path by which these interrelated ECVs may be derived from one another predetermines their actual compliance likelihood.

Conceptually, there are two different ways by which an accuracy criterion – like that of GCOS – could be used in conformity testing. For each one of these, decision rules could then be defined to determine the acceptance and/or rejection of an item. Focusing on the differences between *quasi-instantaneous* estimates of in situ and satellite-derived ECV values one may envisage:

1. **Conformity testing of individual retrievals** (i.e., at the pixel level): this approach implies that the GCOS accuracy criterion Δ, corresponds to a maximum permissible error[4] (i.e., Δ = MPE). In cases where the PDF of the likely values of ε_i is known, then conformity can be declared when the likelihood of ε_i falling within $\pm\Delta$ exceeds the required minimum compliance level, i.e., $P_C(\varepsilon_i) \geq C_L$, where the compliance probability P_C is defined in Eq. (1). If no information on PDFs is available (GUM framework) then only those retrievals may be termed conform that have $|\epsilon_i| \leq \Delta - k \cdot u_{\epsilon_i}$ where, depending on the choice of decision rule either a shared risk ($k = 0$), a coverage interval approach

[4] Relative accuracy criteria (Δ_f) may be converted into absolute MPE values by multiplication with a qualified reference value, i.e., MPE = Δ_fR.

($k > 0$), or a guarded acceptance ($k > 0$ and $u_{\epsilon_i} = u_\epsilon$) approach could be used to determine compliance.

Pixel-based conformity testing has the advantage that compliance information is available for every data entry in a product array. This may be ideal for regulatory contexts or scientific studies requiring high quality information at specific locations and moments in time. However, given the general lack of spatially continuous (in situ) reference data, this conformity testing approach is not likely to be applicable to global/regional EO products unless other satellite-derived reference datasets are available. Furthermore, if the focus is on (1) generating spatially aggregated or temporally averaged products, or (2) studying spatial patterns at a given moment in time or the temporal evolutions at a given location in space, then procedures will have to be defined that allow to deal with non-compliant retrievals within the dataset of interest.

2. **Conformity testing of ensembles of retrievals** (i.e., at the contextual level): this approach implies that the GCOS accuracy criterion (Δ) corresponds to a confidence interval. In cases where the PDF of likely values of ε_i is known, one must thus have $P(P_C(\epsilon_i) \geq C_L) \geq \breve{R}C$ or $'P(P_E(\varepsilon_i) \geq E_L) \leq \hat{R}_{NC}$ to declare compliance. Alternatively, in a GUM framework compliance could be asserted provided that $P(|\varepsilon_i \pm k \cdot u_{\varepsilon_i}| \leq \Delta) \geq \breve{R}C$ or equivalently $P(|\epsilon_i \pm k \cdot u_{\epsilon_i}| > \Delta) \leq \hat{R}_{NC}$, where $\breve{R}_C$ ($\hat{R}_{NC}$) is the minimum (maximum) required rate with which apparent errors may fall within (outside of) the tolerance interval defined by $\pm\Delta$. If $\breve{R}_C \equiv 1 - \hat{R}_{NC} = 1$ then compliance testing of ensembles of retrievals will result in identical outcomes as if the retrievals were all evaluated individually. For other values of $\breve{R}_C$ (or $\hat{R}_{NC}$) the entire set of retrievals (within a particular region, land cover class, temporal window, etc.) will be declared compliant (if its compliance likelihood is larger than $\breve{R}_C$) despite the fact that some of its individual retrievals may be non-compliant.

Ensemble-based conformity testing has the advantage that products may be declared compliant as a whole, or, for specific geographic regions, temporal intervals, land cover types, illumination conditions, etc. While this is clearly of interest to product providers (and also many users), it has the drawback that one may not actually know which ones of the retrievals were compliant and which ones were not. In addition to defining a minimum rate of required compliance $\breve{R}_C$ (or a maximum permissible rate of non-compliance $\hat{R}_{NC}$), this approach also necessitates clear specifications about the contextual categories for which compliance is sought, i.e., whether conformity is needed at the global, regional, country and/or biome level; for specific latitudinal and/or longitudinal zones; for annual, monthly, and/or decadal time steps; for full resolution or coarse-grained EO products, etc.

It is worthwhile to highlight that both of the above approaches require to apply conformity testing to individual retrievals. For the validation of global/regional EO products, ensemble-based conformity testing is the vehicle of choice since it allows – when applied to a representative sample from the global (or contextual) apparent error population – to draw conclusions about the compliance of the global (or contextual) retrievals. Alternatively, if the focus is on quality assurance across small spatial scales, then pixel-based conformity testing may be applied using as reference the maps obtained by correlating high spatial resolution satellite data with quasi-concurrent field estimates [Morisette et al., 2005]. Although GCOS does not state what value of $\breve{R}_C$ should be associated with a given accuracy criterion (Δ), the OSCAR requirements database [WWW-2] of the WMO – which is one of the parent organizations of GCOS – implies, however, that $\breve{R}_C$ could be set to 0.683 if Gaussian distributions of errors were assumed.

3.2. GCOS decision metrics

While GCOS-154 (2011) acknowledges the presence of uncertainty in both the reference and satellite-derived ECV products (e.g., last paragraph on page 7) it does not suggest a means to incorporate these uncertainties into the validation effort per se. Instead it suggests that "A measure such as the *root mean square* or the *mean* has to be chosen to quantify error depending on context". More specifically, the variability of the apparent error "may be quantified by the root mean square (or other measure) of the estimated distribution of errors in product values over a spatial domain, a time interval or a set of similar synoptic situations" [GCOS-154, 2011]. The focus thus is currently on metrics that quantify the distribution of ensembles of apparent error estimates (ε_i) – within some specific contextual class – without, however, making use of any uncertainty information (u_{ϵ_i}) that may be associated with these retrievals. While this appears similar to the idea behind a shared risk approach, it is unlikely that the proposed metrics will yield the same compliance results when applied across a large variety of apparent error populations.

This issue will be demonstrated using the root mean square error (RMSE) of the differences between a number (N) of candidate (Q) and reference (R) estimates of a given EO product:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_i - R_i)^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\epsilon_i)^2} \quad (7)$$

When the number of apparent error estimates becomes very large (i.e., $N \rightarrow \infty$) then the RMSE will tend towards $\sqrt{\bar{\epsilon}^2 + \sigma_\epsilon^2}$. In principle, this makes the RMSE metric particularly attractive for quality assurance efforts since it is able to incorporate information on both the mean apparent error ($\bar{\epsilon}$) and the standard deviation (σ_ϵ) of the underlying population of apparent errors. From a conformity testing point of view, however, it is not clear what cut-off criterion should be used for the RMSE (or any other) metric in order to differentiate between compliant and non-compliant sets of apparent errors in the ECV retrievals.

One obvious choice is to use the GCOS accuracy criterion itself as the cut-off value for metric-based assessments of the compliance of biophysical ECV retrievals, [e.g., Verger et al., 2009]. For bias-free apparent errors distributions with Gaussian characteristics $\mathcal{N}(\bar{\epsilon} = 0, \sigma_\epsilon)$, it may be argued that such a RMSE/MPE ≤ 1 decision rule is equivalent to a shared risk approach with a required minimum compliance rate ($\breve{R}_C$) of 68.3%. This is so because the RMSE will tend towards σ_ϵ if the apparent error population is bias-free ($\bar{\epsilon} = 0$) thus implying that $\pm\sigma_\epsilon$ of the population of apparent errors will be contained within the tolerance interval ($\pm$MPE). Hence a shared risk

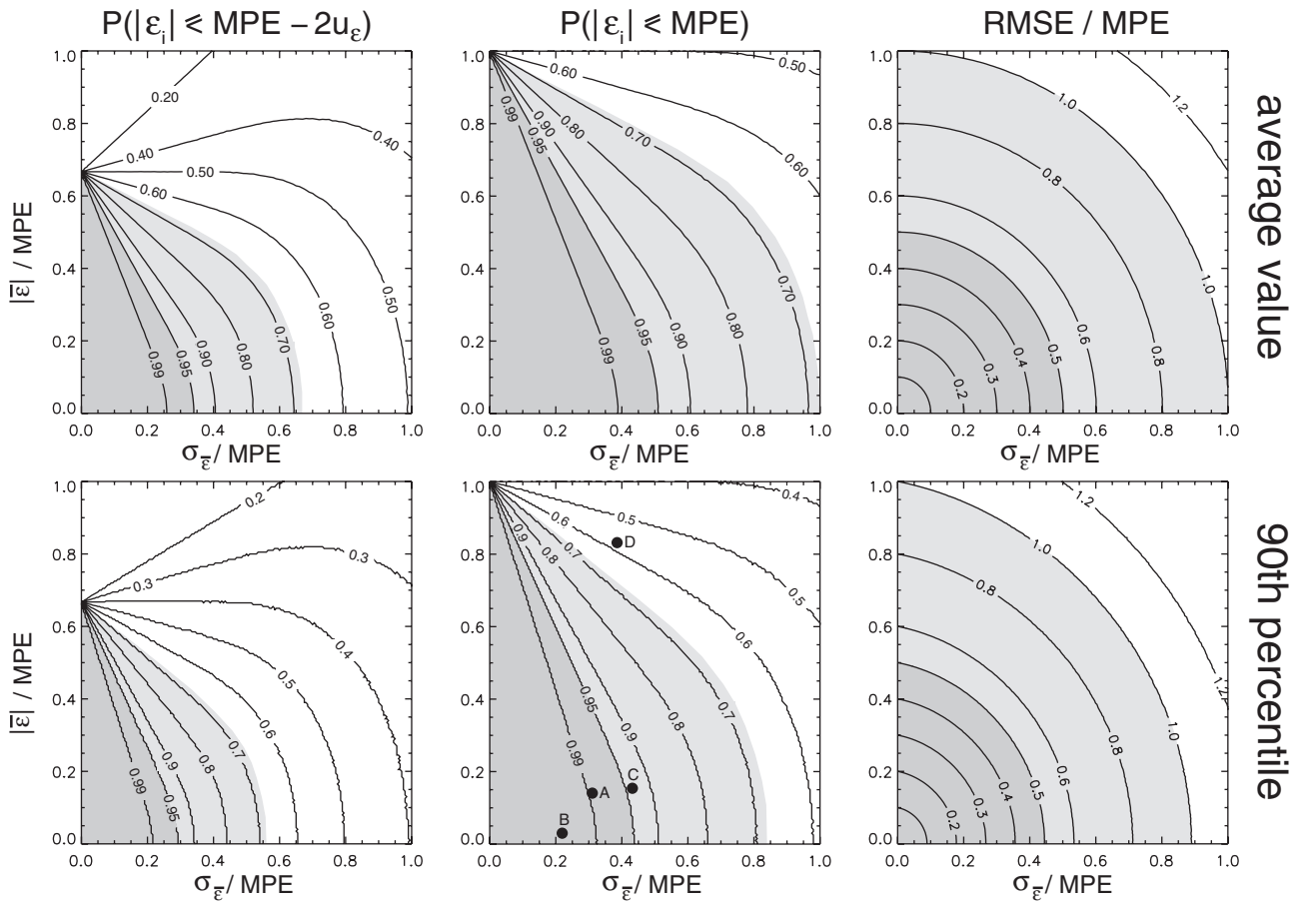

Fig. 5 – Left panels: contour plot of the likelihood of conformity for a guarded acceptance approach requiring 95% confidence that a given apparent error estimate falls within ±MPE, i.e., $P(|\varepsilon_i| \leq \text{MPE} - 2u_\varepsilon)$. Middle panels: contour plot of the likelihood of conformity for a shared risk decision rule that ignores the uncertainty u_ε associated with individual apparent error estimates, i.e., $P(|\varepsilon_i| \leq \text{MPE})$. Right panels: contour plot of the normalized root mean square error (RMSE) metric. The top panels relate to the maximum statistics that can be expected on average (while the bottom panels relate to that which can be expected in 90% of all cases) when 100 values of ε_i are randomly drawn from the underlying Gaussian populations of apparent errors (characterized by the mean $\bar{\varepsilon}$ and the standard deviation $\sigma_{\bar{\varepsilon}}$). Dark grey areas indicate compliance of apparent error populations if the minimum required compliancerate $\breve{R}_C$ is set to 95% (RMSE cut-off value = MPE/2) while the dark and light grey areas together relate to a $\breve{R}_C$ value of 68% (RMSE cut-off value = MPE). The four black disks labelled A to D are specific apparent error populations described in more detail in the text.

approach that chooses $\breve{R}_C \approx 0.683$ will yield identical conformance probabilities as a decision rule that uses the RMSE/MPE ≤ 1 criteria (under the above conditions). By the same logic, the more stringent RMSE/MPE ≤ 0.5 criteria will ensure that $\pm 2\sigma_{\bar{\varepsilon}}$ of the population of apparent errors are contained within ±MPE (if $\bar{\varepsilon} = 0$ and $N \rightarrow \infty$), which for Gaussian PDFs of ε_i is statistically equivalent to a shared risk approach that uses $\breve{R}_C \approx 0.955$. This local equivalence between the RMSE and the shared risk approaches can be used to map out the set of apparent error populations that would be deemed compliant when assessed with conformity testing methods and GCOS recommended metrics.

To do so, it was assumed that all apparent error populations are Gaussian in nature $\mathcal{N}(\bar{\varepsilon}, \sigma_{\bar{\varepsilon}})$ and that their mean apparent error $\bar{\varepsilon}$ and standard deviation $\sigma_{\bar{\varepsilon}}$ span the range $0 \leq \bar{\varepsilon}/\text{MPE} \leq 1$ and $0 \leq \sigma_{\bar{\varepsilon}}/\text{MPE} \leq 1$ in steps of 0.01 each. A bootstrapping approach was then used to draw ten thousand instances of 100 values of ε_i from each single apparent error population. This resulted in 10,000 different values of (1) the RMSE statistic, (2) the $P(|\epsilon_i| \leq \text{MPE})$ compliance rate for the shared risk approach, and, (3) the $P(|\epsilon_i| \leq \text{MPE} - 2u_\epsilon)$ compliance rate for the guarded acceptance method requiring a 95% probability that ε_i actually falls within the tolerance interval ±MPE. For the latter approach, the combined standard uncertainty of individual apparent error estimates u_{ϵ_i} was fixed at MPE/6 in accordance with Eq. (3). In a final step, the average and the 90th percentile of the thus obtained 10,000 RMSE and compliance statistics were chosen.

Fig. 5 thus shows contour plots of the average value (top panels) as well as the 90th percentile (bottom panels) of the distribution of $P(|\epsilon_i| \leq \text{MPE-2u}_\epsilon)$, $P(|\epsilon_i| \leq \text{MPE})$ and RMSE/MPE statistics (in the left, middle and right panels respectively) in a space defined by $\bar{\varepsilon}$/MPE (on the ordinate) and $\sigma_{\bar{\varepsilon}}$/MPE (on the abscissa). For the purpose of comparing the impact of these

approaches, all apparent error populations having RMSE/MPE ≤ 0.5 or $P(|\epsilon_i| \leq \text{MPE} - k \cdot u_\varepsilon) \geq 0.955$ are shaded dark grey in Fig. 5, while those having RMSE/MPE ≤ 1 or $P(|\epsilon_i| \leq \text{MPE} - k \cdot u_\varepsilon) \geq 0.683$ are contained within the light and dark grey areas (where $k = 0$ relates to the shared risk approach while $k = 2$ identifies the guarded acceptance decision rule).

The first thing to note about the panels in Fig. 5 is that the shapes of the contour lines are not identical among the three methods. For the conformity testing approaches in the left and middle panels they originate from a single point along the ordinate whereas for RMSE testing they are radially symmetric around the origin. As a consequence, the extent of the grey (compliance) area in Fig. 5 changes with the method that is used to assert conformity. More specifically, the extent of the combined dark and light grey area increases from the left to the right panel, that is, from the guarded risk, to the shared risk, and then to the RMSE based approach. In fact, had one also included a graph where the mean apparent error ($\bar{\epsilon}$) was used as the metric of choice then all of the $\mathcal{N}(\bar{\epsilon}, \sigma_\varepsilon)$ in the depicted population space would have been found compliant (since $0 \leq \bar{\epsilon}/\text{MPE} \leq 1$ in Fig. 5).

Equally interesting in Fig. 5 is the fact that, if the compliance criterion is refined to a minimum compliance rate of 95.5% or a RMSE cut-off value equal to MPE/2 (dark grey area), then the shared risk and also the guarded acceptance approach both yield different results from the RMSE approach. Overall the extent (but not location) of the dark grey area containing the compliant apparent error populations is somewhat more similar under those conditions between the RMSE approach and the $2u_\varepsilon$ guarded acceptance method. This is different from the $\breve{R}_C \approx 0.683$ scenario (light and dark grey area) where the compliant apparent error populations of the RMSE approach and the shared risk method were most similar.

One final aspect of Fig. 5 is that the light and dark grey areas, containing the compliant apparent error populations for a given decision rule, decrease between the corresponding top and bottom panels. This is not surprising, since the top panels provide information on the value that one can expect on average when computing the RMSE or the non-compliance likelihood on the basis of 100 randomly collected apparent error estimates, while the bottom panels depict the largest value that these statistics are likely to assume in 90 out of 100 such efforts. The bottom panels thus offer a more robust means to determine which apparent error population $\mathcal{N}(\bar{\epsilon}, \sigma_\varepsilon)$ is likely to be labelled compliant when subjected to different sampling efforts.

For demonstration purposes four black disks (marked A to D) were added to the panel showing the 90th percentile of the non-compliance likelihood for a shared risk approach. These four items relate to the apparent error distributions obtained by Claverie et al. (2013) in one of the most comprehensive recent studies on the quality of LAI retrievals over croplands (panels a to d of their Fig. 8). One can see that item D lies outside of the light grey $\breve{R}_C \approx 0.683$ compliance region in Fig. 5 and thus is rather likely to be non-conform with respect to the requirements (i.e., MPE = 0.5). On the other hand, items A and B fall both within the (dark grey) $\breve{R}_C \approx 0.955$ region and thus seem likely to be conform to the quality requirements. Finally, item C – which lies just outside of the dark grey region if the 90th percentile statistics are used – will fall within the dark grey zone if the average required compliance likelihood is used instead (top middle graph). This dependency on the decision rule is also relevant for item D, which can no longer be excluded from being conform if the RMSE/MPE < 1 criteria is being used (bottom right panel), or, item A which will fall out of the dark grey $\breve{R}_C \approx 0.955$ region if a guarded acceptance approach with u_ε = MPE/6 is used (bottom left panel).

While it is known that the GCOS accuracy recommendations were not formulated with a regulatory context in mind, the overall usage of satellite-derived quantitative EO products in commercial, judiciary and regulatory contexts would certainly benefit if widely accepted conformity testing procedures were adopted for their validation. In this case, one has to decide first whether the focus of such efforts should be the rejection or the acceptance of EO products, and second by what decision rule(s) and associated criteria this should be done. Currently, the most reliable conformity testing approaches, that do not require access to a PDF of likely values, include the coverage interval method and the guarded acceptance approach because they make use of the uncertainty associated with the reference and candidate methods, i.e., u_ε. In fact, even the decision to adopt a shared risk approach should be based on an analysis of the magnitude of u_ε as well as the cost associated with erroneous declarations of compliance.

During the past decade or so, international bodies concerned with the calibration and validation of EO data have increasingly stressed the relevance of measurement and retrieval uncertainties. This has recently led to the adoption of the ''Quality Assurance for Earth Observation'' (QA4EO) framework [WWW-4] by CEOS, and its subsequent inclusion in the work program of the Group on Earth Observation [GEO-WP, 2012]. Increasingly, publications estimate the uncertainty of the random error component that is associated with global (or context specific) satellite-based ECV retrievals [e.g., Fang et al., 2012b; D'Odorico et al., 2014]. Some efforts have also been geared towards uncertainty assessments of in situ ECV retrievals [e.g., Richardson et al., 2011]. The latter is particularly relevant since it is the magnitude of u_R that determines whether an in situ method is actually eligible to serve as reference in compliance testing efforts.

4. Eligibility criteria for field protocols

Apparent error populations with characteristics that fall within the acceptance region of a given conformity testing approach can only testify as to the compliance of the candidate retrieval method if the reference method used in their generation actually delivers adequate proxies for the true values of the target quantity. In other words, field validation methods should be validated against primary reference datasets prior to any usage as working standards in conformity testing efforts of third party (EO) products. Ideally, such benchmarking efforts should include a large variety of (environmental) conditions in order to provide information on the systematic and random error structures of the in situ retrieval scheme as well as the combined standard uncertainty of the bias-corrected field estimates. These information will

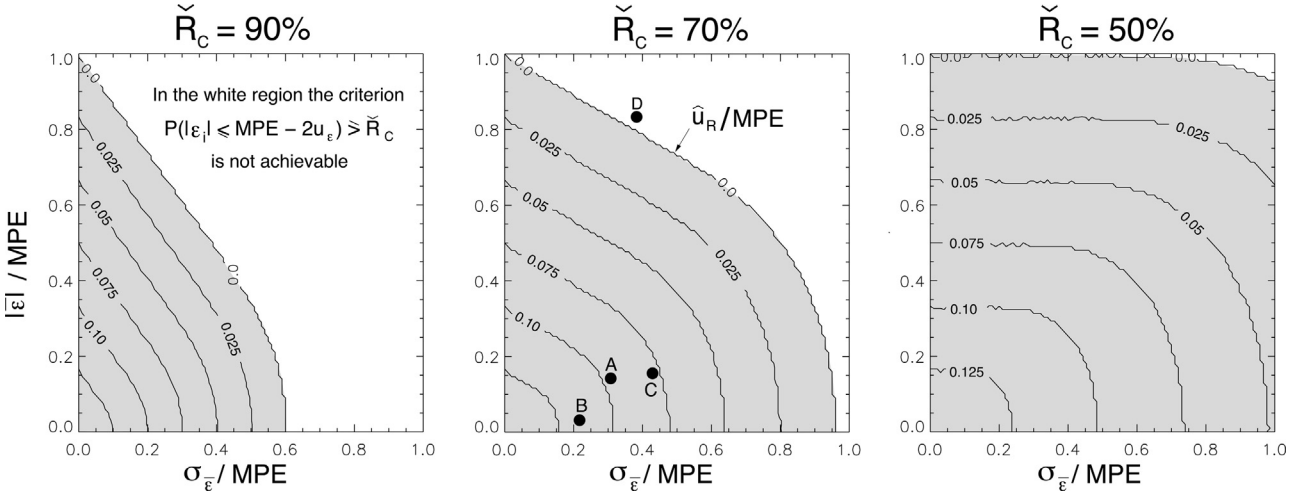

Fig. 6 – Maximum value that the combined standard uncertainty of the reference method ($\hat{u}_R$) may have – according to ISO13528 – in order to determine compliance of candidate methods giving rise to (Gaussian) apparent error populations. More specifically, for each apparent error population $\mathcal{N}(\bar{\varepsilon}, \sigma_{\bar{\varepsilon}})$ – characterized by its mean $\bar{\varepsilon}$ and standard deviation $\sigma_{\bar{\varepsilon}}$ – one million values of ε_i were drawn at random and the likelihood of ε_i's falling outside of the tolerance interval defined by ±MPE was computed using a guarded acceptance approach, i.e., $P(|\epsilon_i| \leq MPE - 2u_\epsilon)$ where the uncertainty of the individual apparent error estimate was varied in the range $0 \leq u_\varepsilon \leq MPE$. The left, middle and right panels show contour lines of $\hat{u}_R$/MPE derived – using Eq. (8) – from the largest value of u_ε allowing to satisfy the minimum permissible compliance rate ($\breve{R}_C$) of 90%, 70% or 50% for a given population of apparent errors. The grey area delineates the region of apparent error populations $\mathcal{N}(\bar{\varepsilon}, \sigma_{\bar{\varepsilon}})$ where the compliance criterion $P(|\epsilon_i| \leq MPE - 2u_\epsilon) \geq \breve{R}_C$ is achievable. The four black dots labelled A to D are described in more detail in the text.

help to (1) adjust the first and second moments of the apparent error distribution (derived from the candidate and in situ methods) to what would have been obtained if the candidate method had been compared against the method yielding the primary reference dataset instead, and (2) assess whether the combined standard uncertainty of the in situ retrieval method is actually small enough to serve as reference in conformity testing efforts of third party datasets. While the first of these is relatively straightforward if a suitable primary reference dataset is available, the second is subject to the width of the prescribed tolerance interval and will be dealt with in more detail below.

4.1. *Qualified reference methods*

For a proper evaluation of a candidate estimation method, the combined standard uncertainty of the reference method u_R should only be a small part of the overall uncertainty (u_ε) associated with the apparent error between the two methods. This concept is well known in analytical chemistry where reference materials play an important role in (1) method validation/calibration efforts, (2) the estimation of measurement uncertainty, and (3) quality control and quality assurance activities [Eurachem, 2002]. In the context of proficiency testing, ISO-13528 formalises this intuitive knowledge by noting that ideally the standard uncertainty of the reference (u_R) should not exceed 30% of what is considered a tolerable range of deviations ($\tilde{\sigma}$) from some (well characterized, accurate and highly precise) primary reference method, [ISO-13528, 2005]:

$$u_R \leq 0.3 \cdot \tilde{\sigma} \tag{8}$$

where $\tilde{\sigma}$ is referred to as the *proficiency standard deviation* in the context[5] of ISO-13528 and corresponds to u_ε in this discussion on ECV retrievals. Pendrill (2014), who writes on measurement uncertainty in decision making and conformity testing, states that the criterion in Eq. (8) is a commonly used limit to ensure that measurement quality variations are small. To the best of our knowledge, however, Eq. (8) has not yet been applied to field validation protocols (of LAI, FAPAR and surface albedo) in order to determine their suitability to assess the compliance of satellite derived estimates (of these same quantities) with respect to predefined quality criteria (like those given by GCOS). This is partly due to the difficulties associated with experimental quantifications of the various uncertainty terms that contribute to u_R (especially when in-direct measurement techniques are used).

Should the criterion in Eq. (8) not be met then the uncertainty of the reference dataset is deemed non-negligible. In that case the risk exists that, the evaluation of third party datasets (i.e., quantitative EO products) will be marred by spurious warnings and erroneous conclusions due to the level of uncertainties associated with the reference data (rather than because of any flaws in the candidate products) [ISO-13528, 2005]. Looking for a reference method capable of satisfying Eq. (8) is thus highly advisable. This is particularly

[5] ISO-13528 provides guidelines on how to assess the proficiency of laboratory methods to deliver measurement results that fall within predefined quality objectives.

relevant in contexts where a single (“best practice”) in situ retrieval scheme is to be applied across a large number of test sites and/or environmental conditions.

Fig. 6 provides information on the largest value that the combined standard uncertainty of a reference method ($\widehat{u}_R$) may have – according to Eq. (8) – such that a ($k = 2$) guarded risk approach will on average satisfy the required minimum compliance rate ($\breve{R}_C$) when applied to a given apparent error population. More specifically, Fig. 6 displays the values of $\widehat{u}_R$/MPE for a large ensemble of Gaussian apparent error populations such that $\breve{R}_C$ is equal to 90% (left panel), 70% (middle panel) and 50% (right panel). For each $\mathcal{N}(\bar{\varepsilon}, \sigma_{\bar{\varepsilon}})$ – characterized by the mean $\bar{\varepsilon}$ and standard deviation $\sigma_{\bar{\varepsilon}}$ – one million values of ε_i were drawn at random and the likelihood with which these ε_i fell within the acceptance interval was computed using a guarded acceptance approach, i.e., as $P(|\epsilon_i| \leq \text{MPE} - 2u_\epsilon)$. This step was carried out by varying the combined standard uncertainty of the individual apparent error estimate in the range $0 \leq u_\varepsilon \leq \text{MPE}$. For each apparent error population the largest value of u_ε still capable of achieving the predefined $\breve{R}_C$ was then determined. Finally, using Eq. (8) this was converted to the corresponding value of $\widehat{u}_R$/MPE.

The first thing to notice in Fig. 6 is that a large number of apparent error populations (white area) cannot be found compliant using a $P(|\epsilon_i| \leq \text{MPE} - 2u_\epsilon) \geq \breve{R}_C$ criterion if the maximum permissible combined standard uncertainty of the reference method has to adhere to the criterion in Eq. (8) and $\breve{R}_C$ is 90%. The extent of the white region in the panels of Fig. 6 decreases as the value of $\breve{R}_C$ becomes smaller. Furthermore, if $\sigma_{\bar{\varepsilon}} = 0$ then the switch from compliance to non-compliance occurs at one specific value of u_ε (and thus also for one specific value of $\widehat{u}_R$). As such, all contour lines in Fig. 6 start at the same location on the ordinate irrespective of the value of $\breve{R}_C$. Furthermore, the $\widehat{u}_R/\text{MPE} = 0$ isolines in Fig. 6 coincide – within numerical precision – with those from a shared risk approach yielding a conformity likelihood of 0.9, 0.7 and 0.5, respectively (compare with the top middle panel of Fig. 5). Similarly, the $\widehat{u}_R/\text{MPE} = 0.05$ isolines for $\breve{R}_C = 90\%$, 70% and 50% in Fig. 6 coincide with the 0.9, 0.7 and 0.5 compliance isolines depicted in the top left panel of Fig. 5 for a $k = 2$ guarded acceptance approach having $u_\varepsilon = \text{MPE}/6$. Last but not least, it may be of interest to know that a guarded risk approach with $k = 1$ instead of $k = 2$, i.e., one for which the compliance condition is $P(|\epsilon_i| \leq \text{MPE} - u_\epsilon) \geq \breve{R}_C$, would result in identically shaped contour lines than those currently in Fig. 6 but where the isolines relate to twice the current $\widehat{u}_R$/MPE values.

Perhaps the most pertinent aspect of Fig. 6 is that the values for $\widehat{u}_R$/MPE are all rather small. From panel A in Fig. 3 it is known that the maximum value of u_ϵ occurs if $\varepsilon = 0$ and that this $\widehat{u}_\epsilon$ amounts to about 0.51·MPE if a guarded acceptance approach with $k = 2$ is to signal compliance. As a consequence, the largest value of $\widehat{u}_R$ will always occur for apparent error populations having $\bar{\varepsilon} = 0$ and also $\sigma_{\bar{\varepsilon}} = 0$ in Fig. 6. For Gaussian PDFs the maximum value of $\hat{u}_R$ is thus a constant and its value amounts to $0.3 \cdot \hat{u}_\varepsilon \approx 15\%$ of the MPE. This is about three times larger than the value of $\hat{u}_R$ that would still permit the use of a shared risk approach (according to the ISO-13528 criteria) if the rule of thumb described in Eq. (3) were used, i.e., $u_R \leq 0.3 \cdot \text{MPE}/6 = 0.05 \cdot \text{MPE}$.

For demonstration purposes the same four items (labelled A to D) as in Fig. 5 were added to the central panel of Fig. 6. One can see that item D lies outside of the region where the $P(|\varepsilon_i| \leq \text{MPE} - 2u_\varepsilon) \geq 0.7$ criterion is achievable on average (even if $\widehat{u}_R = 0$). For item C the value of $\widehat{u}_R$ is about 0.08·MPE, while for item A it is 0.095·MPE and for item B one has $\widehat{u}_R \approx 0.115 \cdot \text{MPE}$. If all three items are to be evaluated with the same reference dataset then the combined standard uncertainty of that reference method should not exceed 8% of the MPE according to the ISO-13528 criterion described in Eq. (8). Given that $\widehat{u}_\varepsilon^2 = \widehat{u}_R^2 + \widehat{u}_Q^{*2}$, this requirement in turn implies that the maximum combined standard uncertainty of the (A, B, C) candidate methods $\widehat{u}_Q^*$ should not exceed 25.4% of the MPE if a guarded acceptance approach with $k = 2$ is to yield compliance. It is not possible, from the data available in Claverie et al. (2013) to infer u_R or to determine the u_Q of items A, B and C and thus any compliance as suggested by Fig. 5 cannot be confirmed.

4.2. *Uncertainty limits for GCOS reference methods*

On the basis of these considerations and the GCOS accuracy criteria listed in section 3.1, one can derive the maximum value of u_R that in situ retrieval methods should have in order to qualify – according to ISO-13528 – as reference in conformity testing efforts aiming to assess the GCOS compliance of satellite-derived biophysical ECV estimates. Table 1 provides the qualifying value of the combined standard uncertainty of the reference method ($\widehat{u}_R$) under the assumption that (1) the PDF of the likely values of ε_i is Gaussian, (2) the GCOS accuracy criteria for LAI, FAPAR and surface albedo delimit the tolerance interval, i.e., MPE = Δ, and (3) a minimum conformance probability of $P_C(\varepsilon_i) \geq 0.955$ (left half of Table 1) and $P_C(\varepsilon_i) \geq 0.683$ (right half of Table 1) are required to assert compliance. The latter probabilities correspond to $\pm 2u_{\epsilon_i}$ and $\pm u_{\epsilon_i}$, respectively, of the Gaussian PDF of likely values of ε_i being contained within the tolerance interval defined by the GCOS accuracy criterion.

More specifically, the first column in Table 1 specifies four arbitrary values of the best estimate of the apparent error (ε_i) normalized by the maximum permissible error taken as the GCOS accuracy criterion here. The second and sixth column then list the largest corresponding value of the combined standard uncertainty of the apparent error ($\widehat{u}_\epsilon$) – again normalized by the MPE – that still permits to achieve a compliance probability of 95.5% and 68.3%, respectively. For example, if $\varepsilon_i/\text{MPE} = 0$ (first entry in Table 1) then the value of $\widehat{u}_\epsilon/\text{MPE}$ cannot be larger than 0.51 if $P_C(\varepsilon_i) \geq 0.955$ but may rise to 1.0 if $P_C(\varepsilon_i) \geq 0.683$. The various remaining columns in Table 1 then list $\widehat{u}_R$ as well as its complement for the candidate method, i.e., $\widehat{u}_Q^* = \sqrt{\widehat{u}_\epsilon^2 - \widehat{u}_R^2}$ in absolute and relative terms for the three ECVs of interest. The relative magnitudes of $\widehat{u}_R$ and $\widehat{u}_Q^*$ were computed with respect to the ECV magnitudes where the GCOS accuracy criterion changes from an absolute to a relative requirement, i.e., for LAI = 2.5, FAPAR = 0.5 and albedo = 0.05.

From Table 1 it becomes apparent that the maximum permissible combined standard uncertainty ($\widehat{u}_R$) which an in situ ECV estimate should ideally have – in order to serve as benchmark in efforts evaluating the compliance of third party

Table 1 – Overview of the maximum combined uncertainty that in situ reference methods (R) for biophysical ECVs should have according to ISO-13528 in order to qualify for conformity testing of the apparent error ($\varepsilon_i = Q - R$) of some candidate method (Q) at a particular test site. The largest combined uncertainty of the reference method ($\widehat{u}_R$) is provided under the assumption that (1) the PDF of the likely values of ε is Gaussian, (2) the GCOS-154 (2011) accuracy criterion for LAI, FAPAR or surface albedo delimits the tolerance interval, i.e., Δ = MPE, and (3) a compliance probability P_C = 95.5% (left half of table) or P_C = 68.3% (right half of table) is imposed. The percentage values of $\widehat{u}_R$ relate to LAI = 2.5, FAPAR = 0.5 and albedo = 0.05 where the GCOS accuracy criterion change from absolute (0.5, 0.05, 0.0025) to relative (20%, 10%, 5%), respectively. Also indicated is the maximum combined standard uncertainty that the candidate method may have under those conditions, i.e., $\widehat{u}_Q^* = \sqrt{\widehat{u}_\epsilon^2 - \widehat{u}_R^2}$.

$\frac{\epsilon_i}{MPE}$	$P_C(\varepsilon_i)$ = 95.5%				$P_C(\varepsilon_i)$ = 68.3%			
	$\frac{\widehat{u}_\epsilon}{MPE}$	$\widehat{u}_R$ and $(\widehat{u}_Q^*)$			$\frac{\widehat{u}_\epsilon}{MPE}$	$\widehat{u}_R$ and $(\widehat{u}_Q^*)$		
		LAI	FAPAR	Albedo		LAI	FAPAR	Albedo
0.0	0.51	0.076 (0.242) or 3.1% (9.7%)	0.008 (0.025) or 1.5% (5.1%)	0.0004 (0.0012) or 0.8% (2.4%)	1.00	0.150 (0.477) or 6.0% (19.1%)	0.015 (0.048) or 3.0% (9.5%)	0.0008 (0.0024) or 1.5% (4.7%)
0.3	0.30	0.045 (0.143) or 1.8% (5.7%)	0.004 (0.013) or 0.9% (2.5%)	0.0002 (0.0007) or 0.4% (1.4%)	0.95	0.142 (0.453) or 5.7% (18.1%)	0.014 (0.045) or 2.8% (9.1%)	0.0007 (0.0023) or 1.4% (4.5%)
0.6	0.17	0.026 (0.083) or 1.0% (3.3%)	0.003 (0.010) or 0.5% (1.9%)	0.0001 (0.0004) or 0.3% (0.8%)	0.76	0.114 (0.362) or 4.6% (14.5%)	0.011 (0.036) or 2.3% (7.2%)	0.0006 (0.0018) or 1.1% (3.6%)
0.9	0.04	0.006 (0.019) or 0.2% (0.8%)	0.001 (0.004) or 0.1% (0.7%)	3×10^{-5} (0.0001) or 0.1% (0.2%)	0.21	0.032 (0.100) or 1.3% (4.0%)	0.003 (0.010) or 0.6% (2.0%)	0.0002 (0.0005) or 0.3% (1.0%)

biophysical ECV products with respect to GCOS accuracy criteria – is rather small. Even if $\varepsilon_i = 0$ and the compliance probability is set to only 68.3%, the value of $\widehat{u}_R$ is still only a mere 6% for LAI, 3% for FAPAR and 1.5% for albedo. The complementary value of the candidate method $\widehat{u}_Q^*$ is about 3.18 times that of the reference, which in the above example amounts to 19.1% for LAI, 9.5% for FAPAR and 4.7% albedo. Obviously, as ε_i deviates from zero (and/or the minimum required compliance level is increased to 95.5%), then the value of $\widehat{u}_{\epsilon_i}$ and hence also $\widehat{u}_R$ and $\widehat{u}_Q^*$ must decrease to maintain the predefined minimum compliance level.

4.3. *Primary reference datasets*

Experimental benchmarking efforts that aim at assessing whether a given field measurement method suffers from a systematic bias and/or is capable of satisfying the uncertainty criteria laid out in Table 1 should have access to another, yet more precise, *primary* reference dataset. The standard uncertainty of this primary reference dataset/method should – by the same ISO 13528 logic – ideally be smaller than $0.3 \cdot \widehat{u}_R$. For the biophysical variables of interest here, this then would translate into a maximum combined standard uncertainty of the primary reference method that is no larger than 1.8% for LAI, 0.9% for FAPAR and 0.45% for albedo if the minimum compliance level is 68.2% (and half of that if C_L = 95.5%) assuming furthermore that the apparent error was free from bias.

Since FAPAR can only be estimated by indirect means and canopy albedo measurements have to be up-scaled to the test site/pixel scale it is not clear how any primary reference data with the above uncertainties could be generated by experimental means for these ECV quantities. In fact, the WMO expects only an achievable uncertainty (at the 95% level) for hourly radiation totals of 3% for high quality instruments (secondary standards) and 5% for instruments operating in networks [Table 7.5 in WMO-2008]. While this corresponds to a standard uncertainty of 1.5% for the secondary standards some of the currently available high quality albedometers come with an expanded calibration uncertainty (k = 2) of 1.2–1.7%, which is, however, much higher than the 0.45% required above for a primary reference datasets for albedo.

Often it is assumed that primary reference datasets for LAI can be acquired through labour intensive *direct* measurement methods in the field. Estimates of the relative standard error of such direct LAI estimations vary from a couple of percent [Fig. 1 of Breda, 2003] to 10–12% [Morrison, 1991 and Kalácska et al. (2005)] and even 25–30% [Table 2 of Eriksson et al., 2005]. Mussche et al. (2001) distinguish between the standard error of the mean (which in their study was of the order of 1.8–2.6% for LAI values in the range between 1.54 and 5.52 for oak/beech stands and 2.4–9.3% for LAI values in the range between 1.64 and 4.53 for ash) and the standard error of a single measurement (which was about five times larger than the standard error of the mean). Again these uncertainty estimates tend to be larger than what would be needed for primary reference datasets according to the criteria laid out in the previous subsection.

Even intensive field measurement campaigns thus seem unlikely to provide rigorous evidence as to the eligibility of field methods to deliver the reference data in efforts assessing GCOS compliance of third party biophysical EO products. An alternative approach to quantify the uncertainty of satellite-derived and/or in situ estimates of quantitative biophysical variables could be the simulation of mission-specific EO data as well as typical field measurements using highly realistic virtual 3D plant environments like those available at WWW-5. The simulated EO and in situ data is then subjected to the same processing steps as any real data such that the resulting ECV estimates can be compared against the true value of the target quantity as it exists within the virtual plant environment, [e.g., Disney et al., 2006, 2011; Widlowski, 2010; Hovi and Korpela, 2014; Côté et al., 2015]. Obviously such a MOdel-based

Quality Assurance (MOQA) framework necessitates the prior verification of the 3D Monte Carlo modelling tools, preferably through the use of established proficiency or conformity testing procedures [e.g., Widlowski et al., 2013, 2015]. Within the limits of their simulation paradigm, Monte Carlo radiative transfer models have been recognized by GCOS as a means of providing method-specific uncertainty information [GCOS-154, 2011] and this at quasi-arbitrary precision levels.

5. Working with imprecise reference methods

Most field validation campaigns nowadays do not measure the biophysical target quantity per se but rather infer its value from observations of correlated variables collected at a series of sampling locations within the area of interest. Up-scaling to the spatial domain covered by the satellite-derived ECV product awaiting validation is then carried out with the help of quasi-concurrently acquired high spatial resolution satellite imagery [Morisette et al., 2005]. Inevitably, such a complex and multi-stage estimation process will give rise to (random and systematic) contributions to the combined standard uncertainty (u_R) of the in situ retrieval method. An important but rarely dealt with aspect is thus what the consequences of working with imprecise reference data are.

Focusing on the regulatory context again, where reference methods must have combined standard uncertainties that are sufficiently small (with respect to some legally binding tolerance level/interval) such as (1) to enable unambiguous conformity testing, (2) to assess the *equivalence* of alternative methods [ECWG-GDE, 2010], and (3) to evaluate the *proficiency* of candidate methods in meeting predefined uncertainty criteria [ISO-13528, 2005]. With the relevance of u_R in the context of conformity testing already addressed, the focus here will thus be on the outcome of tests for the equivalence and proficiency of (retrieval) methods.

5.1. Equivalence of retrieval methods

Regression analysis is perhaps the preferred way to show how satellite-derived estimates compare against in situ reference data. Underlying such efforts is the concept that both populations are (ideally) linearly related, i.e., $Q_i = a + b \cdot R_i$, and that their equivalence can be demonstrated if (1) the slope b and intercept a are insignificantly different from unity and zero, respectively, and (2) the expanded (relative) uncertainty of the candidate method is smaller or equal to that defined by the data quality objective [ECWG-GDE, 2010]. Standard regression procedures, like the ordinary least squares (OLS) method, typically assume that the independent variable (R_i) is known exactly. To account for cases where both the dependent (Q_i) and the independent variables contain errors, so called *errors-in-variable* models have been developed [Gillard, 2010].

In the context of validating quantitative EO products, the estimates of both the satellite and in situ retrieved methods are random variables, that is, $\tilde{Q}_i = Q_i + e_i$ and $\tilde{R}_i = R_i + d_i$, where Q_i and R_i are the true values of the candidate and reference methods, respectively, and d and e are the random error terms that are associated with the retrieved quantities, $\tilde{Q}_i$ and $\tilde{R}_i$, respectively. In the population limit, these errors can be assumed to have zero expectation while their variances are $\sigma_e^2 = \langle e_i^2 \rangle$ and $\sigma_d^2 = \langle d_i^2 \rangle$. Furthermore, one can assume that d and e do not exhibit correlations; neither with each other nor with the true values of the reference and candidate methods. Under those conditions the OLS estimator of the regression slope $\tilde{b}$ is asymptotically linked to the true slope, i.e., $\tilde{b} = b \cdot \lambda$ where the *reliability* $\lambda = (\sigma_{\tilde{R}}^2 - \sigma_d^2)/\sigma_{\tilde{R}}^2$ is confined to the range between zero and one since $\sigma_R^2 = \sigma_{\tilde{R}}^2 - \sigma_d^2$ must be strictly positive. As a consequence, the OLS coefficient $\tilde{b}$ will be biased towards zero and the bias of the regression slope, i.e., $\tilde{b} - b = -(1-\lambda)b$, will be positive if the true slope is negative and negative if the true slope is positive.

Following on from Huang et al. (2006) – who used the errors-in-variables approach to correct regression slopes between in situ and satellite-derived LAI estimates – one can show that the normalized bias of the regression slope $(b - \tilde{b})/b$ is equal to the ratio of the random to total variance in the reference method $(\sigma_d^2/\sigma_{\tilde{R}}^2)$:

$$\frac{b - \tilde{b}}{b} = \frac{\sigma_d^2}{\sigma_{\tilde{R}}^2} \tag{9}$$

While estimates of $\sigma_{\tilde{R}}^2$ typically fall around 1.5 for in situ LAI efforts [Fang et al., 2012b; Camacho et al., 2013; Claverie et al., 2013] reliable information on the typical magnitudes of σ_d^2 is harder to come by. Francq and Govaerts (2014) indicate that estimates of σ_d^2 may be obtained with replicated data. For the purpose of illustration, the regression data provided in Fig. 4 of Huang et al. (2006) suggests that $\sigma_d^2/\sigma_{\tilde{R}}^2 \approx 0.32$ for LAI. An impact of this order would be highly relevant since the value of the regression slope is used to (correct for systematic bias and to) estimate the uncertainty associated with the candidate method. The latter then is compared against predefined criteria to establish equivalence [ECWG-GDE, 2010].

5.2. Proficiency of retrieval methods

The uncertainty of the reference method features in a variety of metrics aiming to identify potential issues with the candidate method. The χ^2 metric used in Pinty et al. (2001), for example, served to identify whether the simulations of candidate models were noticeably different from that of reference models. Another metric, the *erreur normalisé* $E_n = |Q - R|/2u_\varepsilon$, was used by Gerboles et al. (2011) and Widlowski et al. (2013) to assess whether the expanded uncertainty of the candidate method may have been improperly characterized. Large values of E_n may, however, also be caused by an undetected systematic bias contribution. In fact, Eurolab (2006) specifies that a bias is considered significant if the magnitude of the absolute deviation between the candidate and reference estimates exceeds twice the standard uncertainty of this deviation. For the discussion here, the focus will be on the z′ score which is recommended by ISO-13528 for the evaluation of candidate laboratories in the context of proficiency testing:

$$z' = \frac{|Q_i - R_i|}{\sqrt{\tilde{\sigma}^2 + u_{\tilde{R}}^2}} \tag{10}$$

where $\tilde{\sigma}$ is the prescribed proficiency standard deviation that the candidate method must meet. The usage of the z′ score is

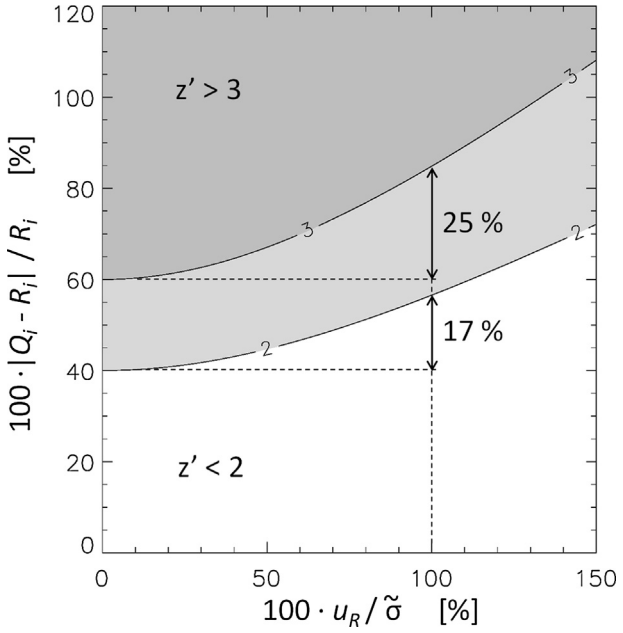

Fig. 7 – Isolines of z′ = 2 and z′ = 3 generated under the assumption that the tolerable uncertainty level $\tilde{\sigma}$ amounts to 20% of the reference value R (this is equivalent to the GCOS accuracy criteria for LAI ≥ 2.5). z′ values were generated for different levels of uncertainty associated with the reference solution (u_R is varied from 0 to 150% of $\tilde{\sigma}$) as well as different levels of absolute bias of the candidate method ($|Q_i - R_i|$ is changed from 0 to 120% of R_i). If $u_R \approx \tilde{\sigma}$ then the relative bias between the candidate and reference solution must be 17% or 25% higher than if the reference solution is known exactly in order to cross the z′ = 2 and z′ = 3 trigger values, respectively.

only meaningful if the uncertainty of the reference method adheres to Eq. (8), i.e., $u_R \leq 0.3 \cdot \tilde{\sigma}$.

What then is the consequence when Eq. (8) is no longer valid? For this one must know that the output of, χ^2 z' and E_n metrics is typically compared against (one or more) trigger values to determine whether the candidate dataset is substantially different from the reference. In the case of the χ^2 and E_n metrics all outcomes larger than unity raise a warning flag. In the case of the z′ score, ISO-13528 indicates that outcomes of $2 < z' < 3$ and $z' \geq 3$ are equivalent to "warning" and "action" signals, respectively. In particular the latter is typically raised only when "results are so exceptional as to merit investigation and corrective action"[ISO-13528, 2005].

If the standard uncertainty of the reference method (u_R) increases then the difference between the reference (R) and the candidate (Q) estimates has to become increasingly larger before the metric will trigger its warning sign. This is documented in Fig. 7, which shows a contour plot of the z′ score for different levels of biases existing between the candidate and reference methods (along the ordinate) as well as for different levels of uncertainty associate with the reference quantity (along the abscissa). As the uncertainty of the reference method is increased, then the z′ = 2 and z′ = 3 isolines curve upward because larger deviations between the candidate and reference solutions (along the ordinate) are now needed before the z′ scores cross the trigger values of 2 and 3.

To put this into context, Fig. 7 includes the findings of Fernandes et al. (2001) who computed the theoretical error budget for overstorey LAI using a retrieval equation that is often used for in-direct LAI estimations in the field. Their analysis found that the LAI errors at the plot-level were close to 20% for needle leaf stands which is quasi-identical to the GCOS accuracy requirement for satellite-derived LAI products. Hence if one assumes $u_R/\tilde{\sigma} = 1$ then this implies that the bias between the candidate and reference solutions (along the ordinate in Fig. 7) must reach 57% before the z′ = 2 isoline is crossed (and 85% before the z′ = 3 isoline is crossed). Compared to the case where the reference value is known exactly (i.e., 0 on the ordinate in Fig. 7), the bias between the candidate and reference values thus must be 17% (and 25%) larger before the trigger values of z′ = 2 (and z′ = 3) are reached. If one were to assume that $u_R/\tilde{\sigma} = 1.5$ then the corresponding increases in bias are of the order of 32% and 48%, respectively. The use of reference methods with excessively large combined uncertainties (in the sense of ISO-13528) thus increases the risk for users of quantitativeEO products (in the sense that compliance may be certified even though it should not have been).

6. Concluding remarks

This contribution introduced conformity testing as a means to assess the value of satellite-derived quantitative surface variables (e.g., LAI, FAPAR and albedo). Quality requirements in this context are typically expressed in terms of a permissible interval of deviations from a qualified reference. Shared risk approaches seem to offer a convenient means to assess compliance here, because they only verify whether the apparent errors (between the reference and candidate retrievals) fall within the tolerance interval or not. Since they do not account for uncertainties in the measurement or estimation process, shared risk approaches cannot provide confidence statements as to the reliability of their outcome. In fact, the usage of shared risk approaches is only recommended in legal metrology if it can be demonstrated that the uncertainty of the apparent error (ε) is negligible with respect to the permissible interval of deviations.

Due to the inherent nature of the measurement process, it is the uncertainty of both the reference and the candidate items (i.e., EO products here) that affect the reliability with which the apparent errors of a given quantitative EO product can be estimated. Guarded risk approaches (as well as PDF-based decision rules) account for the uncertainty associated with the reference and candidate items. As such they permit statements expressing the confidence supporting the outcome of the conformity testing effort. The guarded acceptance approach, for example, does this by reducing the width of the tolerance interval in a manner that depends on the combined uncertainty of the apparent error existing between the reference (R) and candidate (Q) items, i.e., $u_\epsilon^2 = u_R^2 + u_Q^2$.

Conformity testing of quantitative EO products may be carried out at the level of individual retrievals (i.e., pixels) or perhaps more meaningful at the level of ensembles of

retrievals. The latter give rise to apparent error populations for which a minimum required compliance rate (with respect to the prescribed tolerance interval) has to be established. It was shown in this contribution that the shared risk approach performed rather different from the guarded acceptance and PDF-based decision rules when applied across a large variety of apparent error populations. This was largely due to the magnitude of u_ε which was used to define both the width of the guard band and the spread of the PDF of likely ε_i values. Obviously, the smaller the combined uncertainty of the apparent error is, the better becomes the resulting agreement between different decision rules. When the shared risk and guarded acceptance approaches were compared with the RMSE metric, considerable differences in the type of apparent error populations that would be deemed compliant were noted. This was especially the case if the minimum compliance rate was set to 95%. In general, the likelihood of a positive outcome in conformity testing is thus affected by (1) the choice of decision rule, (2) the width of tolerance interval, and for ensemble based testing (3) the level of the minimum required compliance rate.

Guarded risk approaches are among the most reliable decision rules for conformity testing. The width of their guard bands is typically set to $2 \cdot u_\varepsilon$ such as to reduce the risk of false acceptances to about 5%. In legal metrology it is often recommended that the value of $2 \cdot u_\varepsilon$ does not exceed one third of the predefined tolerance interval. At the same time, the uncertainty of the reference should ideally be no more than 30% of u_ε. In the context of biophysical ECVs, D'Odorico et al. (2014) recently estimated that the random error of FAPAR products lies at 10–20% while Fang et al. (2012a) using the same estimation technique indicate a range of 25–35% for LAI. If these figures are representative estimates of u_Q then they alone are already close to or even exceeding the rule of thumb that $u_\varepsilon \leq$ MPE/6. However, more pertinent still is the condition that the uncertainty associated with (indirect) in situ retrieval methods of biophysical ECVs is typically larger than the recommend $u_R/u_\varepsilon \leq 0.3$ of ISO-13528. Fernandes et al. (2001), for example, indicate a value of about 20% for u_R in the context of in situ LAI estimates, while Fensholt et al. (2004) quote 15% for FAPAR estimates. These figures, if representative, thus place a doubt on the eligibility of current field validation schemes to provide suitable sources of reference data in conformity testing efforts of satellite-derived biophysical ECVs on the basis of the current GCOS quality criteria.

As an alternative, validated 3D Monte Carlo radiative transfer models may be used to assess the compliance of satellite-derived ECV products (as well as the physical consistency between different ECV quantities retrieved from the same EO data stream). By ingesting thus simulated 'satellite data' into the actual processing chain used in the ground segment of space agencies an independent means is obtained to evaluate the quality of the retrieval algorithm under controlled conditions (and this irrespective of ECV definitions and sensor-specific acquisition and compositing schemes). Such a MOdel-based Quality Assurance (MOQA) framework is currently discussed by CEOS WGCV and also by GCOS as a viable means to achieve uncertainty characterizations of in situ ECV retrieval schemes. Particularly interesting here, is that the MOQA scheme provides a financially attractive means to increase the statistical robustness of quality assurance efforts for quantitative EO products; simply by increasing the number of virtual plant environments to twice, triple or even ten times that available for direct comparison efforts in the real world (where the number of globally available direct validation sites typically falls between 30 and 50, [e.g., Camacho et al., 2013; Fang et al., 2012b].

Despite these advantages, it must be understood that a MOQA framework can only yield uncertainty estimates that reflect the range of variability contained within the virtual systems under study. As such the degree of realism that is embedded in model-based approaches requires an independent verification (just like field validation schemes require such a verification to test their eligibility as reference in compliance testing efforts). In this context, field validation efforts are essential (1) to confirm the biases and differences estimated with a MOQA setup between different (satellite and in situ) retrieval schemes, (2) to monitor eventual changes in the performance of satellite retrieval algorithms due to causes other than those addressed within the MOQA framework, and (3) to contribute to the determination of the actual regression slopes needed to compute the true uncertainty of the satellite products and potentially also to correct them from biases. At the same time, the MOQA framework may help to quantify the various uncertainty components of different in situ measurement protocols, such as to permit the validation community to select the most adequate procedure for a given environment or tolerance interval. Quality assessments of quantitative EO variables over land are thus likely to gain from validation efforts that combine field and model-based quality assurance techniques.

While it may always be argued that the decision rules and uncertainty criteria presented here are too stringent to be of any practical use in an EO context, the mere fact that they have been developed in legal metrology testifies to their value in situations where conformance truly matters. Any weakening of quality requirements and/or conformity testing will ultimately have to be borne by the users of satellite-derived quantitative surface products. Whether this is appropriate or not depends on the cost associated with false compliance information and hence the application context. However, if the usage of quantitative EO products is to be encouraged outside of its classical science application context then the availability of trustworthy and user-oriented compliance information may be the best means to achieve this.

Acknowledgements

This study would not have been possible without the support of F. Raes and A. Belward, heading the Climate Risk Monitoring unit and the Land Resources Management unit, respectively, at the Institute for Environment and Sustainability of the European Commission's Joint Research Centre in Ispra, Italy. A special thank you also goes to D. Buzica and M. Gerboles who helped to clarify the author's understanding of both equivalence and compliance testing in the context of European ambient air quality directives.

REFERENCES

ASME-B89731, 2001. Guidelines for Decision Rules: Considering Measurement Uncertainty in Determining Conformance to Specifications. American Society of Mechanical Engineers, ASME B89.7.3.1:2001, New York.

Baret, F., Weiss, M., Lacaze, R., Camacho, F., Makhmara, H., Pacholcyzk, P., Smets, B., 2013. GEOV1: LAI and FAPAR essential climate variables and FCOVER global time series capitalizing over existing products. Part1: principles of development and production. Remote Sens. Environ. 137, 299–309.

Bojinski, S., Verstraete, M., Peterson, T.C., Richter, C., Simmons, A., Zemp, M., 2014, September. The concept of essential climate variables in support of climate research, applications, and policy. Bull. Am. Meteorol. Soc. 431–443.

Breda, N.J.J., 2003. Ground-based measurements of leaf area index: a review of methods, instruments and current controversies. J. Exp. Bot. 54 (392) 2403–2417.

Camacho, F., Cernicharo, J., Lacaze, R., Baret, F., Weiss, M., 2013. GEOV1: LAI, FAPAR essential climate variables and FCOVER global time series capitalizing over existing products. Part 2: validation and intercomparison with reference products. Remote Sens. Environ. 137, 310–329.

Canisius, F., Fernandes, R., 2012. Evaluation of the information content of medium resolution imaging spectrometer (MERIS) data for regional leaf area index assessment. Remote Sens. Environ. 119, 301–314.

Cescatti, A., Marcolla, B., Vannan, S.K.S., Yun Pan, J., Román, M.O., Yang, X., Ciais, P., Cook, R.B., Law, B.E., Matteucci, G., Migliavacca, M., Moors, E., Richardson, A.D., Seufert, G., Schaaf, C.B., 2012. Intercomparison of MODIS albedo retrievals and in situ measurements across the global FLUXNET network. Remote Sens. Environ. 121, 323–334.

Claverie, M., Vermote, E., Weiss, M., Baret, F., Hagolle, O., Demarrez, V., 2013. Validation of coarse spatial resolution LAI and FAPAR time series over cropland in southwest France. Remote Sens. Environ. 139, 216–230.

COM-312, 2013. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL Establishing the Copernicus Programme and repealing Regulation (EU) No 911/2010, COM(2013) 312 final/2. European Commission. Available at: http://www.copernicus.eu/pages-principales/library/policy-documents/

Côté, J.-F., Fournier, R.A., Verstraete, M.M., 2015. Canopy architectural models in support of methods using hemispherical photography. In: Fournier, R.A., Hall, R. (Eds.), Hemispherical Photography for Forest Science: Theory, Methods and Applications. Springer. (in press).

Cox, P., Stephenson, D., 2007. A changing climate for prediction. Science 317, 207–208.

De Leeuw, J., Georgiadou, Y., Kerle, N., De Gier, A., Inoue, Y., Ferwerda, J., Smies, M., Narantuya, D., 2010. The function of remote sensing in support of environmental policy. Remote Sens. 2, 1731–1750.

Disney, M.I., Lewis, P., Saich, P., 2006. 3D modelling of forest canopy structure for remote sensing simulations in the optical and microwave domains. Remote Sens. Environ. 100, 114–132.

Disney, M.I., Lewis, P., Gomez-Dans, J., Roy, D., Wooster, M.J., Lajas, D., 2011. 3D radiative transfer modelling of fire impacts on a two-layer savanna system. Remote Sens. Environ. 115, 1866–1881.

Directive, 2008. 2008/50/EC of The European Parliament and of The Council of 21 May 2008 on ambient air quality and cleaner air for Europe, Published in the Official Journal of the European Union, L152/1 on 11/06/2008. .

D'Odorico, P., Gonsamo, A., Pinty, B., Gobron, N., Coops, N., Mendez, E., Schaepman, M.E., 2014. Intercomparison of fraction of absorbed photosynthetically active radiation products derived from satellite data over Europe. Remote Sens. Environ. 142, 141–154.

Dowell, M., Lecomte, P., Husband, R., Schulz, J., Mohr, T., Tahara, Y., Eckman, R., Lindstrom, E., Wooldridge, C., Hilding, S., Bates, J., Ryan, B., Lafeuille, J., Bojinski, S., 2013. Strategy Towards an Architecture for Climate Monitoring from Space. 39Available from www.ceos.org, www.wmo.int/sat, and www.cgms-info.org.

ECWG-GDE, 2010. Guide to the Demonstration of Equivalence of Ambient Air Monitoring Methods, Report by an EC Working Group on Guidance for the Demonstration of Equivalence. Available at: http://ec.europa.eu/environment/air/quality/legislation/pdf/equivalence.pdf.

EPA, 2006. Guidance on Systematic Planning Using the Data Quality Objectives Process. United States Environmental Protection Agency, EPA QA/G-4, 111http://www.epa.gov/quality/qs-docs/g4-final.pdf (accessed January 2014).

Eriksson, H., Eklundh, L., Hall, K., Lindroth, A., 2005. Estimating LAI in deciduous forest stands. Agric. Forest Meteorol. 129, 27–37.

Eurachem, 2002. The Selection and Use of Reference materials: A Basic Guide for Laboratories and Accreditation Bodies, EEE/RM/062rev3. Available at http://www.eurachem.org/images/stories/Guides/pdf/EEE-RM-062rev3.pdf.

Eurachem, 2007. Use of Uncertainty Information in Compliance Assessment, EURACHEM/CITAC Guide, 1st ed.15Available at http://www.eurachem.org/guides/Interpretation_with_expanded_uncertainty_2007_v1.pdf.

Eurolab, 2006. Guide to the Evaluation of Measurement Uncertainty for Quantitative Test Results, Technical Report No. 1/2006. European Federation of National Associations of Measurement, Testing and Analytical Laboratories. .

Fang, H., Wei, S., Jiang, C., Scipal, K., 2012a. Theoretical uncertainty analysis of global MODIS, CYCLOPES, and GLOBCARBON LAI products using a triple collocation method. Remote Sens. Environ. 124, 610–621.

Fang, H., Wei, S., Liang, S., 2012b. Validation of MODIS and CYCLOPES LAI products using global field measurement data. Remote Sens. Environ. 119, 43–54.

Fensholt, R., Sandholt, I., Schultz Rasmussen, M., 2004. Evaluation of MODIS LAI, fAPAR and the relation between fAPAR and NDVI in a semi-arid environment using in situ measurements. Remote Sens. Environ. 91 (3–4) 490–507.

Fernandes, R.A., White, H.P., Leblanc, S.G., Pavlic, G., Mc-Nairn, H., Chen, J.M., Hall, R.J., 2001, August. Examination of error propagation in relationships between leaf area index and spectral vegetation indexes from Landsat TM and ETM. In: Proc. 23rd Can. Remote Sensing Symp., Quebec City, QC, Canada, pp. 41–51.

Franch, B., Vermote, E.F., Claverie, M., 2014. Intercomparison of Landsat albedo retrieval techniques and evaluation against in situ measurements across the US SURFRAD network. Remote Sens. Environ. 152, 627–637.

Francq, B.G., Govaerts, B.B., 2014. Measurement methods comparison with errors-in-variables regressions. From horizontal to vertical OLS regression, review and new perspectives. Chemom. Intell. Lab. Syst. 134, 123–139.

GCOS-92, 2004. Implementation Plan for the Global Observing System for Climate in Support of the UNFCCC. GCOS-92 (WMO/TD No. 1219) http://www.wmo.int/pages/prog/gcos/Publications/gcos-92_GIP.pdf.

GCOS-107, 2006. Supplemental Details to the Satellite-Based Component of the Implementation Plan for the Global Observing System for Climate in Support of the UNFCCC, GCOS–107 (WMO/TD No. 1338) http://www.wmo.int/pages/prog/gcos/Publications/gcos-107.pdf.

GCOS-138, 2010. Implementation Plan for the Global Observing System for Climate in Support of the UNFCCC (2010 Update).GCOS-138 (GOOS-184, GTOS-76, WMO-TD/No. 1523) http://www.wmo.int/pages/prog/gcos/Publications/gcos-138.pdf (accessed March 2013).

GCOS-154, 2011. Systematic Observation Requirements for Satellite-based Data Products for Climate. 2011 Update: 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. http://www.wmo.int/pages/prog/gcos/documents/SatelliteSupplement2011Update.pdf.

GEO-WP, 2012. Group on Earth Observation (GEO) Workplan 2012–2015, Revision 3. http://www.earthobservations.org/documents/work plan/geo_wp1215_rev3_140123.pdf (accessed 04.11.14).

Gerboles, M., Buzica, D., Brown, R.J.C., Yardley, R.E., Hanus-Illnar, A., Salfinger, M., Vallant, B., Adriaenssens, E., Claeys, N., Roekens, E., Sega, K., Jurasovi?c, J., Rychlik, S., Rabinak, E., Tanet, G., Passarella, R., Pedroni, V., Karlsson, V., Alleman, L., Pfeffer, U., Gladtke, D., Olschewski, A., O'Leary, B., O'Dwyer, D., amd Pockeviciute, M., Biel-Cwikowska, J., Tursic, J., 2011. Interlaboratory comparison exercise for the determination of As, Cd, Ni and PM10 in Europe. Atmos. Environ. 45, 3488–3499.

Gillard, J., 2010. An overview of linear structural models in errors in variables regression. REVSTAT – Stat. J. 8 (1) 57–80.

GTOS-63, Assessment of the Status of the Development of the Standards for the Terrestrial Essential Climate Variables, T08: Albedo, Global Terrestrial Observing System, http://www.fao.org/gtos/doc/ecvs/t08/t08.pdf (accessed 05.06.13)

GTOS-65, Assessment of the Status of the Development of the Standards for the Terrestrial Essential Climate Variables, T10: FAPAR, Global Terrestrial Observing System, http://www.fao.org/gtos/doc/ECVs/T10/T10.pdf (accessed 05.06.13)

GTOS-66, Assessment of the Status of the Development of the Standards for the Terrestrial Essential Climate Variables, T11: LAI, Global Terrestrial Observing System, http://www.fao.org/gtos/doc/ECVs/T11/T11.pdf (accessed 05.06.13)

Hovi, A., Korpela, I., 2014. Real and simulated waveform-recording LiDAR data in juvenile boreal forest vegetation. Remote Sens. Environ. 140, 665–678.

Huang, D., Yang, W., Tanş, B., Rautiainen, M., Zhang, P., Hu, J., Shabanov, N.V., Linder, S., Knyazikhin, Y., Myneni, R.B., 2006. The Importance of Measurement Errors for Deriving Accurate Reference Leaf Area Index Maps for Validation of Moderate-Resolution Satellite LAI Products. IEEE Transactions on Geoscience and remote sensing 44 (7) 1866–1871.

IEC-115, 2007. Application of Uncertainty of Measurement to Conformity Assessment Activities in the Electrotechnical Sector, International Electronic Commission, IEC Guide 115, Edition 1.0. .

ISO-5725, 1994. Accuracy (trueness and precision) of Measurement Methods and Results (Part 1–6 Together with Corrections). International Organisation for Standards, Geneva, Switzerland.

ISO-14253-1, 1998. Geometrical Product Specification (GPS) – Inspection by Measurement of Workpieces and Measuring Equipment – Part 1: Decision Rules for Providing Conformance or Non-Conformance with Specifications, ISO 14253-1:1998(E). International Organisation for Standards, Geneva, Switzerland.

ISO-10576-1, 2003. Statistical Methods – Guidelines for the Evaluation of Conformity with Specified Requirements – Part 1: General Principles. ISO 10576-1:2003(E) International Organisation for Standards, Geneva, Switzerland.

ISO-13528, 2005. Statistical Methods for Use in Proficiency Testing by Interlaboratory Comparisons, ISO 13528:2005(E). International Organisation for Standards, Geneva, Switzerland.

ISO-21748, 2010. Guidance for the Use of Repeatability, Reproducibility and Trueness Estimates in Measurement Uncertainty Estimation, ISO 21748,:2010(E). International Organisation for Standards, Geneva, Switzerland.

Ito, A., 2011. Legal Aspects of Satellite Remote Sensing. Martinus Nijhoff Publishers, 356, http://dx.doi.org/10.1163/ej.9789004190320.i-354.12.

JCGM-100, 2008. Evaluation of Measurement data – Guide to the Expression of Uncertainty in Measurement (GUM 1995 With Minor Corrections), Joint Committee for Guides in Metrology, JCGM 100:2008. Available from http://www.bipm.org/

JCGM-101, 2008. Evaluation of Measurement Data — Supplement 1 to the Guide to the Expression of Uncertainty in Measurement — Propagation of Distributions Using a Monte Carlo Method, JCGM 101:2008. Available from http://www.bipm.org/

JCGM-106, 2012. Evaluation of Measurement Data – The role of Measurement Uncertainty in Conformity Assessment, JCGM 106:2012. Available from http://www.bipm.org/

Kalácska, M., Calvo-Alvarado, J.C., Sánchez-Azofeifa, G.A., 2005. Calibration and assessment of seasonal changes in leaf area index of a tropical dry forest in different stages of succession. Tree Physiol. 25, 733–744.

Källgren, H., Lauwaars, M., Magnusson, B., Pendrill, L., Taylor, P., 2003. Role of measurement uncertainty in conformity assessment in legal metrology and trade. Accredit. Qual. Assur. 8, 541–547.

OIML-TC3, 2009. The Role of Measurement Uncertainty in Conformity Assessment Decisions in Legal Metrology, OIML/TC 3/SC 5/N1. Organisation International de Métrologie Légale, 52.

Malenovský, Z., Rott, H., Cihlar, J., Schaepman, M.E., García-Santos, G., Fernandes, R., Berger, M., 2012. Sentinels for science: potential of Sentinel-1, -2, and -3 missions for scientific observations of ocean, cryosphere, and land. Remote Sens. Environ. 120, 91–101.

Mayer, A.L., Lopez, R.D., 2011. Use of remote sensing to support forest and wetlands policies in the USA. Remote Sens. 3, 1211–1233.

Morisette, J.T., Baret, F., Privette, J.L., Myneni, R.B., Nickeson, J., Garrigue, S., Shabanov, N., Weiss, M., Fernandes, R., Leblanc, S., Kalacska, M., Sánchez-Azofeifa, G.A., Chubey, M., Rivard, B., Stenberg, P., Rautiainen, M., Voipio, P., Manninen, T., Pilant, A., Lewis, T., Iames, J., Colombo, R., Meroni, M., Busetto, L., Cohen, W., Turner, D., Warner, E.D., Petersen, G.W., Seufert, G., Cook, R., 2005. Validation of global moderat resolution LAI products: a framework proposed within the CEOS Land Product Validation subgroup. IEEE Trans. Geosci. Remote Sens. 44, 1804–1817.

Morrison, I.K., 1991. Effect of trap dimensions on litter fall collected in an Acer saccharum stand in northern Ontario. Can. J. Forest Res. 21, 939–941.

Mussche, S., Samson, R., Nachtergale, L., De Schrijver, A., Lemeur, R., Lust, N., 2001. A comparison of optical and direct methods for monitoring the seasonal dynamics of leaf area index in deciduous forests. Silva Fennica 35 (4) 373–384.

Pendrill, L.R., 2007. Optimised measurement uncertainty and decision-making in conformity assessment. Measure 2 (2) 76–86.

Pendrill, L.R., 2014. Using measurement uncertainty in decision-making and conformity assessment. Metrologia 51, S206–S218.

Pickett-Heaps, C.A., Canadell, J.G., Briggs, P.R., Gobron, N., Haverd, V., Paget, M.J., Pinty, B., Raupach, M.R., 2013. Evaluation of six satellite-derived fraction of absorbed photosynthetic active radiation (FAPAR) products across the Australian continent. Remote Sens. Environ. 140, 241–256.

Pinty, B., Gobron, N., Widlowski, J.-L., Gerstl, S.A.W., Verstraete, M.M., Antunes, M., Bacour, C., Gascon, F., Gastellu, J.-P., Goel, N., Jacquemoud, S., North, P., Qin, W., Thompson, R., 2001. The radiation transfer model intercomparison (RAMI) exercise. J. Geophys. Res. 106, 11937–11956.

Posselt, R., Mueller, R., Trentmann, J., Stockli, R., Liniger, M.A., 2014. A surface radiation climatology across two meteosat satellite generations. Remote Sens. Environ. 142, 103–110.

Prieto-Blanco, A., North, P.R.J., Barnsley, M.J., Fox, N., 2009. Satellite-driven modelling of net primary productivity (NPP): theoretical analysis. Remote Sens. Environ. 113 (1) 137–147.

Richardson, A.D., Dail, D.B., Hollinger, D.Y., 2011. Leaf area index uncertainty estimates for model-data fusion applications. Agric. Forest Meteorol., http://dx.doi.org/10.1016/j.agroformet.2011.05.009.

Rochdi, N., Fernandes, R., 2010. Systematic mapping of leaf area index across Canada using 250-meter MODIS data. Remote Sens. Environ. 114 (5) 1130–1135.

Sommer K.D., Kochsiek, M., 2002. Role of measurement uncertainty in deciding conformance in legal metrology, OIML. Bulletin XLIII (2) 19–24.

Taberner, M.B., Pinty, Y., Govaerts, S., Liang, M.M., Verstraete, N., Gobron, N., Widlowski, J.-L., 2010. Comparison of MISR and MODIS land surface albedos: methodology. J. Geophys. Res. 115, http://dx.doi.org/10.1029/2009JD012665 D05101.

Verger, A., Camacho, F., Garcia-Haro, F.J., Melia, J., 2009. Prototyping of Land-SAF leaf area index algorithm with VEGETATION and MODIS data over Europe. Remote Sens. Environ. 113, 2285–2297.

Widlowski, J-L., 2010. On the bias of instantaneous FAPAR estimates in open-canopy forests. Agric. Forest Meteorol. 150, 1501–1522.

Widlowski, J.-L., Pinty, B., Lopatka, M., Atzberger, C., Buzica, D., Chelle, M., Disney, M., Gastellu-Etchegorry, J.-P., Gerboles, M., Gobron, N., Grau, E., Huang, H., Kallel, A., Kobayashi, H., Lewis, P.E., Qin, W., Schlerf, M., Stuckens, J., Xie, D., 2013. The 4th radiative transfer model intercomparison (RAMI-IV): proficiency testing of canopy reflectance models with ISO-13528. J. Geophys. Res. – Atmos. 118, http://dx.doi.org/10.1002/jgrd.50497.

Widlowski, J.-L., Mio, C., Disney, M.I., Adams, J., Andredakis, I., Atzberger, C., Brennan, J., Busetto, L., Chelle, M., Ceccherini, G., Colombo, R., Côté, J.-F., Eenmäe, A., Essery, R., Gastellu-Etchegorry, J.-P., Gobron, N., Grau, E., Haverd, V., Homolová, L., Huang, H., Hunt, L., Kobayashi, H., Koetz, B., Kuusk, A., Kuusk, J., Lang, M., Lewis, P.E., Lovell, J.L., Malenovský, Z., Meroni, M., Morsdorf, F., Mõttus, M., Nilson, T., Ni-Meister, W., Pinty, B., Rautiainen, M., Schlerf, M., Somers, B., Stuckens, J., Verstraete, M.M., Yang, W., Zhao, F., Zenone, T., 2015. The fourth phase of the radiative transfer model intercomparison (RAMI) exercise: actual canopy scenarios and conformity testing. Remote Sens. Environ. (submitted).

WELMEC, 2006. Elements for Deciding the Appropriate Level of Confidence in Regulated Measurements, WELMEC 4.2, European co-operation in legal metrology. http://www.welmec.org/fileadmin/user_files/publications/4-2.pdf.

WMO-2008, 2008. Guide to Meteorological Instruments and Methods of Observation, WMO-No. 8, Geneva. .

WWW-1, Monitoring Atmospheric Composition and Climate, MACC-2, http://www.gmes-atmosphere.eu/

WWW-2, OSCAR Requirements Database of Physical Variables, http://www.wmo-sat.info/oscar/

WWW-3, Global Climate Observing System (GCOS), http://www.wmo.int/pages/prog/gcos/

WWW-4, Quality Assurance for Earth Observation, QA4ECV, http://qa4eo.org

WWW-5, RAdiative transfer Model Intercomparison, RAMI, http://rami-benchmark.jrc.ec.europa.eu/HTML/RAMI-IV/RAMI-IV.php

Zibordi, G., Mélin, F., Voss, K.J., Johnson, B.C., Franz, B.A., Kwiatkowska, E., Huot, J-P., Wang, M., Antoine, D., 2015. System vicarious calibration for ocean color climate change applications: requirements for in situ data. Remote Sens. Environ. 159, 361–369.