Atmos. Meas. Tech. Discuss., 8, 8509–8562, 2015
www.atmos-meas-tech-discuss.net/8/8509/2015/
doi:10.5194/amtd-8-8509-2015

Atmospheric Measurement Techniques Discussions
Open Access

Known and unknown unknowns: the application of ensemble techniques to uncertainty estimation in satellite remote sensing data

A. C. Povey and R. G. Grainger

National Centre for Earth Observation, University of Oxford, Clarendon Laboratory, Parks Road, Oxford OX1 3PU, UK

Received: 23 June 2015 – Accepted: 20 July 2015 – Published: 10 August 2015

Correspondence to: A. C. Povey (adam.povey@physics.ox.ac.uk)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

This paper discusses a best-practice representation of uncertainty in satellite remote sensing data. An estimate of uncertainty is necessary to make appropriate use of the information conveyed by a measurement. Traditional error propagation quantifies the uncertainty in a measurement due to well-understood perturbations in a measurement and auxiliary data – known, quantified "unknowns". The underconstrained nature of most satellite remote sensing observations requires the use of various approximations and assumptions that produce non-linear systematic errors that are not readily assessed – known, unquantifiable "unknowns". Additional errors result from the inability to resolve all scales of variation in the measured quantity – unknown "unknowns". The latter two categories of error are dominant in underconstrained remote sensing retrievals and the difficulty of their quantification limits the utility of existing uncertainty estimates, degrading confidence in such data.

This paper proposes the use of ensemble techniques to present multiple self-consistent realisations of a data set as a means of depicting unquantified uncertainties. These are generated using various systems (different algorithms or forward models) believed to be appropriate to the conditions observed. Benefiting from the experience of the climate modelling community, an ensemble provides a user with a more complete representation of the uncertainty as understood by the data producer and greater freedom to consider different realisations of the data.

1 Introduction

All measurements are subject to error, the difference between the value obtained and the theoretical true value (or measurand). Errors are traditionally classified as "random" or "systematic" depending on if they would have zero or non-zero mean (respectively) when considering an infinite number of measurements of the same circumstances. The uncertainty on a measurement describes the expected magnitude of the error by

AMTD
8, 8509–8562, 2015

A. C. Povey and R. G. Grainger

characterising the distribution of error that would be found if the measurement was infinitely repeated. These concepts are sketched in Fig. 1.

Uncertainty is a vital component of data as it provides

- a means to efficiently and consistently communicate the strengths and limitations of data to users, and
- a metric with which to compare and consolidate different estimates of a measurand.

The importance of quoting the uncertainty on any measurement and the thorough validation of both are well accepted, but the terms "uncertainty" and "validation" are used inconsistently.

This paper aims to present a succinct outline of uncertainty and validation and their best-practice application to satellite remote sensing of the environment. Satellite remote sensing is a sequence of processes that estimate a geophysical quantity from a measurement of the current or voltage produced by a space-based detector in response to the radiation incident upon it. Each step in processing, formally described in Table 1, is subject to various sources of error. This formalisation was applied as early as 1970 for Nimbus 4 data processing (G. Peskett, personal communication, 2015), but did not enter the peer-reviewed literature until much later (Ducher, 1980).

Standardised methods for uncertainty estimation can be insufficient for satellite remote sensing data as they assume a well-constrained measurement where the sources of error are established – *known, quantifiable unknowns*. The dominance of systematic errors in satellite remote sensing data introduce *known, unquantified unknowns* (such as the impact of cloud filtering) and *unknown unknowns* (such as variability on scales smaller than that observed).

Ensemble techniques, a method widely used in the weather and climate communities, provide multiple self-consistent realisations of a data set as a means of representing non-linear error propagation and variations resulting from ambiguous representations of natural processes. This paper argues that such techniques provide an

AMTD
8, 8509–8562, 2015

effective means to represent and communicate the uncertainty resulting from the latter two categories of "unknowns" affecting satellite remote sensing data.

The discussions to follow aim to be accessible to both users and producers of satellite remote sensing data and the issues considered apply (theoretically) to all satellite-based instruments. The relative importance of each point will depend on the precise technique considered and the concepts will not be considered for all possible measurements. Illustrative examples will primarily draw from the characterisation of aerosol, cloud, and the surface with a hypothetical nadir-viewing radiometer in a low Earth orbit (~ 800 km) with a spatial resolution of ~ 1 km having bands in the visible and infrared. This specification is typical of a number of past and existing instruments such as the Along Track Scanning Radiometer (ATSR) series, the Advanced Visible High Resolution Radiometer (AVHRR) series, and the Moderate Resolution Imaging Spectroradiometer (MODIS) on the Aqua and Terra platforms.

Section 2 outlines the accepted definition of uncertainty and the use of ensemble techniques in characterising the distribution of systematic errors in satellite remote sensing data. These are discussed with respect to specific sources of error in Sect. 3. Retrieval validation is considered in Sect. 4. Section 5 discusses the importance of qualitative information in the communication of uncertainty to data users while Sect. 6 summarises some conclusions and recommendations.

2 Representing uncertainty

2.1 Within retrieval theory

A generalised description of a retrieval technique is that it uses observations $\boldsymbol{y}$ and auxiliary information $\boldsymbol{b}$ to find some quantities of interest $\boldsymbol{x}$ that satisfy

$$\boldsymbol{y} = F(\boldsymbol{x}, \boldsymbol{b}) + \boldsymbol{\epsilon}, \tag{1}$$

which is practically performed by evaluating

$$\boldsymbol{x} = \boldsymbol{G}(\boldsymbol{y}, \boldsymbol{b}), \tag{2}$$

where the forward model $\boldsymbol{F}$ approximates the process by which the instrument and environment translate the desired quantities $\boldsymbol{x}$ into the observation $\boldsymbol{y}$ and whose formulation will depend on the choice of basis $\boldsymbol{x}$. The error in the measurements and forward model is denoted $\boldsymbol{\epsilon}$ and the inverse function $\boldsymbol{G}$ is some statistical or approximate inversion of the forward model, for which many schemes exist (e.g. Rodgers, 2000; Twomey, 1997).

If a hat denotes the theoretical true value of a quantity or function, the error in the retrieval is given by $\boldsymbol{\varepsilon} = \boldsymbol{x} - \hat{\boldsymbol{x}}$. It is affected by sources that fall between three extremes:

– Random fluctuations in the measurement, such as thermal fluctuations and shot noise. These are unavoidable but generally linear and (at least approximately) normally distributed such that the uncertainty can be represented by the standard deviation of their distribution. When using Eq. (2), the uncertainty resulting from random errors in multiple measurements can be calculated using the standard "propagation of errors" (Clause 5.1.2 of Working Group 1, 2008)

$$\sigma_{\boldsymbol{x}_j} = \sqrt{\sum_{i=1}^{N} \left(\sigma_{\boldsymbol{y}_i} \frac{\partial \boldsymbol{G}_j}{\partial \boldsymbol{y}_i} \right)^2}, \tag{3}$$

where $\sigma_{\boldsymbol{x}_j}$ is the uncertainty in the jth element of $\boldsymbol{x}$ and N observations were considered, which are assumed to have uncorrelated errors.

– Simplifications and approximations made in the technique. These errors are systematic and are unlikely to be quantified (as they would have been included in the forward model if they were). Such errors are commonly characterised through validation.

AMTD
8, 8509–8562, 2015

A. C. Povey and R. G. Grainger

– The degree to which the observation is representative of the situation it is proposed to describe. These are especially important for satellite observations, where measurements are averaged over some volume of the atmosphere that does not necessarily correspond to the scale of physical perturbations, such as turbulent mixing or cloud contamination.

These considerations compound when considering the uncertainty resulting from the use of auxiliary parameters, $\boldsymbol{b}$. If the uncertainty on the auxiliary parameters is well known, it is straightforward to propagate it into the retrieval using Eq. (3) with the substitution $\boldsymbol{y} \rightarrow \boldsymbol{b}$. However, the data may not map directly onto the defined state (e.g. observations at a different spatial resolution taken at a different sub-solar time), introducing additional error. If an auxiliary parameter is very poorly known, it may be preferable to retrieve it as an additional element of $\boldsymbol{x}$, though in doing so the problem may become underconstrained (if it was not already). Even where it is possible to make additional measurements, it is often necessary to input an independently retrieved quantity rather than work from raw data.

2.2 Formal definition

The metrological community has prepared an extensive summary of best-practice in the assessment of uncertainty in measurements – the Guide to the expression of uncertainty in measurement (Working Group 1, 2008, known hereafter as the GUM). It defines uncertainty as a

> "parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand."

This definition has been adopted by the European Space Agency's (ESA) Climate Change Initiative (CCI project teams, 2010).

In clause 0.4, the GUM states that an ideal method for evaluating uncertainty should be *universal*, in that it is applicable to all types of data. The reported uncertainty should

AMTD
8, 8509–8562, 2015

A. C. Povey and R. G. Grainger

then be *internally consistent*, being directly derivable from the information that was used in its calculation, and *transferable*, such that it can be input to subsequent calculations. These are achieved by assuming that any probability distribution from which errors are sampled can be accurately described by a single variance. If a series of N observations x_i are made, the mean is $\langle x \rangle = \frac{1}{N}\sum_{i=1}^{N} x_i$ with variance

$$\sigma^2_{\langle x \rangle} = \frac{\sum_{i=1}^{N}\left(x_i - \langle x \rangle\right)^2}{N-1}. \tag{4}$$

Clause 4.3 provides guidelines for determining a pseudo-variance when observations are not repeated, such as where the measurand is known to fall between two limits. With that, Eq. (3) can be evaluated for the equations used to derive the measurement (outlined in clause 5).

2.3 Application to satellite remote sensing

These conventions apply equally to satellite remote sensing data but represent an impractical ideal that does not help an analyst fully represent their understanding of the uncertainty in their data. This is due to the simplistic treatment of systematic errors. Clause 3.2.4 of the GUM states that, "It is assumed that the result of a measurement has been corrected for all recognized significant systematic effects and that every effort has been made to identify such effects." While data producers put significant effort into identifying systematic errors, their quantification can be a difficult and occasionally impossible task. For such errors, it is unclear that their distribution is symmetric, such that the emphasis on traditional error propagation contributes to many analysts neglecting important systematic errors as they cannot be quantified with confidence (Li et al., 2009; Kokhanovsky et al., 2010). This applies primarily to highly underconstrained observations. A few measurements of the radiation at the top of atmosphere (TOA) cannot be used to deduce the intricate state of the atmosphere and surface in the observed column without substantial simplification of the physics and/or additional

AMTD
8, 8509–8562, 2015

A. C. Povey and R. G. Grainger

information on the variation of the state. Systematic errors are produced where these assumptions break down (e.g. using an inaccurate water vapour profile when evaluating measurements affected by water absorption).

The magnitude and nature of systematic errors experienced is a function of the state observed. A common example is the differing treatment of land and sea surfaces. Averaging adjacent retrievals will not necessarily combine errors sampled from the same distribution. As the uncertainty of a retrieval is a function of the environment observed, they must be ascertained on a pixel-by-pixel basis to be meaningful.

The basis chosen to describe a system also impacts the expression of uncertainty. Consider the retrieval of cloud top temperature or pressure from measurements by a nadir-viewing infrared radiometer (for a more detailed description, see King, 1992; Fischer and Grassl, 1991; Schiffer and Rossow, 1983). The observed signal is the radiance at TOA, which is converted (using the Planck function) into the radiating temperature of the droplets at the top of the cloud. As that transform is non-linear, a symmetric distribution of random error in the radiance will not be symmetric when considering temperature, as sketched in Fig. 2. Similarly, the cloud top pressure is calculated from the temperature by interpolating a meteorological profile. As temperature varies linearly with height while pressure varies logarithmically, the distribution will be further distorted in pressure space, in addition to the uncertainty introduced by the meteorological profile.

If errors are expected to be small (as in the radiance to temperature transform), the non-linearity will be minimal and a variance-based representation of error is sensible. Otherwise, the distribution of error may be skewed or asymmetric such that one value is insufficient to describe it. Ensemble techniques can provide the additional information required to properly characterise the distribution of error.

2.4 Ensemble techniques

The standard error propagation techniques do not properly represent the distribution of non-linear errors. In such situations, the uncertainty can be represented by the vari-

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

AMTD

8, 8509–8562, 2015

Uncertainty estimation in satellite remote sensing data

A. C. Povey and R. G. Grainger

ation in an ensemble of individually self-consistent predictions. An example is that of numerical weather prediction (NWP). Rather than predict the weather from the output of a single model run, multiple runs are performed (Buizza et al., 2005) with each initialised by a perturbed version of the initial state (the perturbations being consistent with the uncertainty in the observations used). The weather is chaotic, such that small changes in the input data produce significant and non-linear changes in the result (Lorenz, 1965). The ensemble of forecasts captures the variability as an approximation of the uncertainty in a forecast (Houtekamer and Lefaivre, 1997), such as the fraction of model runs in which a given feature is observed, in a way that standard error propagation cannot.

The representation of non-linear error propagation via ensembles is applicable to satellite remote sensing observations. Components of the retrieval's uncertainty can be determined using an ensemble of retrievals where each member of the ensemble adds a random perturbation to the measurements $\boldsymbol{y}$ and ancillary parameters $\boldsymbol{b}$ (in accordance with their respective error distributions). The feasibility of doing this in large-scale processing is limited by computational cost so it is primarily useful as a method to validate the calculated uncertainties.

Ensembles are also widely used in the climate modelling community (for example, Flato et al., 2013; Crucifix et al., 2005; Meehl et al., 2000). Many processes cannot be accurately modelled at the coarse resolutions practical for climate modelling. These are parametrised, but there are many possible schemes and each has associated unquantifiable systematic errors. The uncertainty in climate models is approximated by considering the diversity in an ensemble of models using different assumptions and approximations.

Such ensembles are useful where a measurement does not fully constrain a problem. To illustrate the concept, consider estimating the volume of an aluminium bucket knowing only its mass. As the density of aluminium is known and assuming the thickness of metal used to make the bucket, the mass can be converted into a surface area. The volume is then determined from the surface area by assuming the shape and

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |
AMTD
8, 8509–8562, 2015
Uncertainty estimation in satellite remote sensing data
A. C. Povey and R. G. Grainger
Title Page
Abstract Introduction
Conclusions References
Tables Figures
Back Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

height of the bucket. That choice of shape (i.e. the forward model) will greatly affect how the retrieval interprets the mass measurement.

This is portrayed in Fig. 3. Each line represents a different forward model for converting mass into volume. A slice (lines of the same colour) shows the impact of shape on the form of the forward model. Looking through the slices (different colours of the same line style) shows the impact of the assumed height. Note:

- When the bucket is assumed to have a height of 12 cm (purple), the three different models produce consistent results between 0.15 and 0.3 kg. The error due to using an inappropriate model there will be small, but increases for masses > 0.3 kg. The error is a function of the true state.
- For a height of 24 cm (red) the models diverge greatly; a 0.32 kg bucket could have a volume between 0.10 and 11 L. Thus, the use of an incorrect model will introduce substantial error. The error is a function of the forward model's parameters.
- In this example the actual shape of the bucket is not known, so it is not possible to rigorously quantify the error resulting from the choice of forward model. Without additional information, the results for a hemispherical bucket are just as valid as a conical one despite their significantly different interpretations of the data (e.g. a hemispherical bucket has a minimum mass for a given height whilst a conical one does not).

The form of the ensemble will depend on its intended use and a priori knowledge. In this example, the ensemble would be three estimates of the volume (one for each shape). The uncertainty resulting from errors in the weight, density, and thickness would be given separately for each ensemble member. If genuinely nothing was known about the height, the ensemble could be extended to represent a range of heights. In reality, some auxiliary information will exist that should constrain the values.

The standard deviation across ensemble members may be a useful proxy where the models are consistent, as in the 12 cm slice, but not generally. Non-linear errors

AMTD
8, 8509–8562, 2015

Uncertainty estimation in satellite remote sensing data

A. C. Povey and R. G. Grainger

can be most meaningfully described through an ensemble, with which many users already have extensive experience (Rayner et al., 2014). Ensemble techniques are universal, being a generalisation of the GUM's techniques to a poorly constrained problem (i.e. a well-constrained problem has a one-member ensemble). Each realisation of the data is internally consistent and the ensemble presents a more complete understanding of the data, as ambiguities are explicitly highlighted. The information is transferable using the well-established techniques of the modelling community.

This example is artificial but illustrates the utility of ensemble techniques to satellite remote sensing data:

- Retrievals of aerosol optical depth are strongly affected by the choice of aerosol microphysical properties. Analogous to the choice of bucket shape, these properties alter the form of the forward model and introduce unquantifiable errors. An ensemble can be produced by evaluating the observations with various models, as currently performed by the MISR and ORAC algorithms (Liu et al., 2009; Thomas et al., 2009, respectively).

- A variety of techniques can be used to merge multiple satellite sensors into a single, long-term product, such as the Jason-1 and 2 mean sea-level missions (Ablain et al., 2015) or the SeaWiFS and MODIS Terra and Aqua ocean colour data (Maritorena and Siegel, 2005). These correspond to the choice of bucket height – a poorly constrained retrieval parameter.

- Retrieval parameters and auxiliary data have associated uncertainties. Where the propagation of these is highly non-linear, they can be estimated via ensemble techniques analogous to the NWP approach, as done by Liu et al. (2015). Rather than present an ensemble of retrievals, Mears et al. (2011) produced an ensemble of estimated errors (as perturbations about the measured value presume it is the mean of the true distribution).

- Errors that are correlated over large temporal and/or spatial scales are impractical to calculate and represent with traditional covariance matrices. Ensembles have

AMTD
8, 8509–8562, 2015

been used to represent these in SST products (Kennedy et al., 2011a, b), with less problematic errors represented by separate uncertainty estimates.

In essence, the ensemble approach is useful for characterising the error resulting from an incomplete description of the situation observed. At the expense of increased data volume, an ensemble provides the user with

1. a more appropriate representation of the uncertainty resulting from the realisation of the problem, and

2. the freedom to select the portrayal(s) of the data most appropriate to their purposes.

An ensemble also facilitates the intercomparison of different methodologies, through which techniques can be refined or rejected.

3 Evaluating errors in a satellite observation

Despite their extensive use in the community (and this paper), the classification of errors as random or systematic is limited. A random error can appear to introduce a systematic bias after propagation through a non-linear equation due to its asymmetric distribution and the distribution of a systematic error has finite width. The use of these terms is better understood as synonyms for the non-technical meanings of noise and bias, respectively.

The GUM chose to eschew classification of error altogether, instead classifying uncertainties as type A and B dependent on if they were calculated from an observed frequency distribution (i.e. traditional statistical techniques) or an assumed probability density function. This provides an important focus on the different techniques through which uncertainty is calculated, but does not address the interest of data users in understanding the cause of errors in a measurement. The source of an error affects how it is realised and its relative importance in the eyes of data producers and users. Five classifications of error by source are proposed, which will be discussed in turn.

AMTD
8, 8509–8562, 2015

A. C. Povey and R. G. Grainger

3.1 Measurement errors

Measurement errors result from statistical variation in the measurand or random fluctuations in the detector and electronics. To assess these accurately, it is important that a measurement is traceable to a well-documented standard. This requires the straightforward (if not simple) comparison of an instrument to a thoroughly characterised reference. Further, the response of any instrument will evolve over time, necessitating the periodic repeat of calibration procedures.

Satellite radiometers are characterised prior to launch (e.g. Hickey and Karoli, 1974; Barnes et al., 1998; Tanelli et al., 2008), to varying levels of accuracy, providing a traceable assessment of uncertainty. However, the stresses of launch can irrevocably and unpredictably alter the behaviour of an instrument, such that this assessment merely provides a first guess of the performance in practice (e.g. Kummerow et al., 2000). It is impossible to perform calibration in orbit analogous to the laboratory-based format. Some instruments carry calibration sources to provide continual, in-situ evaluation (e.g. Smith et al., 2012). Though designed to be more robust than the instrument itself, these have been shown to have stability issues (Xiong et al., 2010). Hence, it is unreasonable to expect a traceable assessment of uncertainty for a satellite-borne sensor analogous to any ground-based instrument.

Vicarious methods of calibration can be used, whereby the response of the instrument to a known Earth-bound stimulus is considered (e.g. Slater et al., 1996; Fougnie et al., 2007; Powell et al., 2009; Kuze et al., 2014). For example, radiometers have been calibrated by observing an area of the Libyan desert known to have a very stable surface reflectance over time (Smith et al., 2002). For some instruments, this is the only direct calibration possible (Heidinger et al., 2003). Calibrations are periodically re-evaluated and new data sets released (e.g. the recent ATSR V1.2 or MODIS L1B Collection 6). For such calibrations to be traceable, it is necessary to establish international standard reference sites that are independently and regularly monitored.

AMTD
8, 8509–8562, 2015

3.2 Parameter errors

Retrievals using satellite observations virtually always require auxiliary information as there is insufficient information available to retrieve all parameters of the atmosphere and surface simultaneously. For example, the accuracy of line-by-line radiative transfer calculations depends upon the spectroscopic data used (see, for example, Fischer et al., 2008). Parameters will be produced by an independent retrieval and have associated uncertainties. If uncertainty is reported via a standard deviation, it can be propagated using Eq. (3). More complex uncertainties can be represented through an ensemble.

3.3 Approximation errors

It is not always practical to evaluate the most precise formulation of a forward model. For example, the atmosphere may be approximated as plane parallel to simplify the equations or look-up tables (LUT) may be used rather than solving the equations of radiative transfer. Such approximations will introduce error. Often known as "forward model error" (Rodgers, 2000), it can be assessed by comparing the performance of the full and simplified forward models with simulated data. These errors can be highly state-dependent but should also be small (as otherwise the approximation was misguided), such that it should be appropriate to quantify the maximum error and convert that into an effective standard deviation (GUM Clause 4.3).

3.4 Resolution errors

3.4.1 Definition of the measurand

How a measurand is defined affects which errors are relevant. Summarising clause D.3 of the GUM, consider the use of a micrometer to measure the thickness of a sheet of paper. As the sheet will not be uniform, the true value depends on the precise location of the measurement. Hence, when measuring "the thickness of this sheet of paper", the variation of thickness across the sheet is an additional source of error to be considered when estimating the uncertainty. This error can be neglected by defining the measurand as "the thickness of this sheet of paper at this point", but that is of little practical use. Similarly, "the thickness of a sheet of paper from this supplier" is a more useful measurand, for which the error due to variations between different sheets would also need to be considered.

A datum in a satellite product is understood to represent an average of some physical quantity over the observed pixel at a specified time. Compared to the situations considered in the GUM, these suffer a number of important limitations:

1. It is not possible to redefine the scope of the measurand (i.e. changing from "this sheet of paper" to "a sheet from this supplier") as that is prescribed by the optics of the instrument. What will be called the *resolution error* derives from the inability of the measurement to resolve the desired measurand. This generally results from variations in the quantity on scales smaller than a pixel, analogous to the variations in thickness over a sheet of paper.

2. The perturbations are not necessarily independent. For example, in the open ocean it is reasonable to expect that mixing will homogenise SST over a pixel, but in coastal waters variations in depth and sediment concentration introduce spatially correlated perturbations that will not average to zero.

3. Unlike the thickness example, it is not possible to repeat the observation. Atmospheric states evolve over minutes to hours and influence (to some extent) any environmental observation such that two instruments can never strictly observe the same state. This is unusual in the sciences, where experiments generally accumulate statistical confidence through repeated measurement of equivalent circumstances.

The last point can be addressed by averaging adjacent pixels from the same sensor. When done with Level 1 data, this is known as superpixeling (Munechika et al., 1993).

AMTD
8, 8509–8562, 2015

It is commonly used in aerosol retrievals to reduce measurement error (e.g. Sayer et al., 2010a), as aerosols are assumed to vary over scales much larger than a pixel (order 50 km, Anderson et al., 2003). Such averaging is not valid in the presence of cloud, which is fundamentally a stochastic feature with an extended region of influence (Grandey and Stier, 2010).

When Level 2 data is averaged, the result is Level 3 data. Averages over hundreds of kilometres and days to weeks are similar to the scales evaluated by climate models and the volume of data is vastly more manageable. Such data are susceptible to additional limitations:

– The definition of the measurand is even more important. It may appear sufficient to describe a product as (for example) "average SST in March 2005 over 30–31° N and 10–11° W", but the satellite's spatial sampling will greatly affect the value. Comparison of satellite products to model outputs can only be successful if the model is sampled as if observed by that satellite (so called "instrument simulators", e.g. Sayer et al., 2010b).

– Satellite products are only representative of the time they observe (Privette et al., 1995). If the quantity has a diurnal cycle, the measurand should be described as an average at a specific time. That time may evolve through a record due to satellite drift, such that data from the beginning of such a record may not be directly comparable to those at the end.

– Resolution errors are a function of the pixel size and the variability of the measured quantity. The independent pixel approximation (Chambers et al., 1997) commonly used assumes a constant quantity at the pixel scale. While this approximation holds in many circumstances, it is not universally true and certainly breaks down as pixels are aggregated to represent a larger spatial scale.

– For retrievals that use an a priori constraint, each retrieved value contains a contribution from the a priori. When averaging, if the a priori is not "removed" from the

AMTD
8, 8509–8562, 2015

Uncertainty estimation in satellite remote sensing data

A. C. Povey and R. G. Grainger

value, it will contribute repeatedly to the average, biasing it. Neglecting covariance between state vector elements, this can be done via

$$x'_i = \left(\frac{x_i}{\sigma_i^2} - \frac{x_a}{\sigma_a^2}\right) \Big/ \left(\frac{1}{\sigma_i^2} - \frac{1}{\sigma_a^2}\right). \quad (5)$$

To account for covariance, see Eq. (10.47) of Rodgers (2000). The values x'_i can then be averaged as desired, explicitly including the a priori value once.

3.4.2 Impact of sampling

The interaction of cloud with the radiation field is sufficiently complex and variable that it is not generally possible to retrieve its properties simultaneously with the surface and/or other atmospheric constituents. Hence, most atmospheric measurements are pre-filtered for the presence of cloud via one of a plethora of empirical techniques (e.g. Ackerman et al., 1998; Stowe et al., 1999; Pavolonis and Heidinger, 2004; Curier et al., 2009). This constrains the retrieval to observations believed to be appropriate to the forward model used.

The filtering process impacts the sampling of the product, as regions with persistent cloud cover will be neglected. Level 3 products are particularly susceptible to these sampling effects. The concept is also known as “fair-weather bias” as the exclusively clear-sky conditions considered are not necessarily representative of the long-term average conditions that the measurand purports to describe. Ensemble techniques can be used to characterise this error either by demonstrating the changes in coverage as a function of the cloud filter used or by explicitly considering cloudy conditions as an alternative realisation of the system (for which the state vector will likely be different).

Filtering can remove exceptional events. Aerosol retrievals often assume all data with optical thickness above some threshold are cloud contaminated, but it is possible for dust or volcanic ash to achieve an optical thickness above any useful threshold. This systematically removes high optical depths from long-term averages, producing a low

AMTD
8, 8509–8562, 2015

A. C. Povey and R. G. Grainger

bias in average products and failing to characterise the largest (and potentially most important) events. Such limits should be stated within the product definition to make this distinction clear.

3.5 System errors

The stochastic change in TOA radiance due to the presence of cloud (or other optically thick layer such as smoke or volcanic ash) is a long-standing problem in satellite remote sensing. The issue is that the forward model, $\boldsymbol{F}$ in Eq. (1), has a significantly different form for each stochastic realisation of the environment. One realisation will be referred to as a *system*.

If there is no a priori knowledge of which system is appropriate, the forward model could be formed from the linear sum of all possible systems, e.g.

$$\boldsymbol{y} = a\boldsymbol{F}_{\text{clear sky}}(\boldsymbol{x}_a, \boldsymbol{b}_a) + b\boldsymbol{F}_{\text{cloud}}(\boldsymbol{x}_b, \boldsymbol{b}_b) + c\boldsymbol{F}_{\text{smoke}}(\boldsymbol{x}_c, \boldsymbol{b}_c) + \cdots + \boldsymbol{\epsilon}, \tag{6}$$

where $a, b, c, \ldots$ are the weighting of each system, which sum to unity. Each system is represented by a unique state $\boldsymbol{x}_a, \boldsymbol{x}_b, \boldsymbol{x}_c, \ldots$ and there may be degeneracies between them (e.g. each state may quantify the surface reflectance). While this approach may be successful for some multispectral observation systems, in most cases it makes an underconstrained problem worse.

Another technique is to assume the measurements are of a specific system (i.e. one of the weights is unity and the others are zero). The choice of system is based on prior knowledge, usually relative values of radiances or their spatial variability (e.g. the cloud flagging discussed in Sect. 3.4.2). However, the choice of thresholds is often application dependent, leading to gross error (e.g. Sect. 3.2 of Holzer-Popp et al., 2015) as there is a substantial difference between asking “Is this an observation of X?” and “Is this observation suitable for analysis with my model of X?” The former desires an appraisal of the state based on data; the latter seeks to minimise forward model errors.

An alternative approach is to perform a retrieval with each relevant system in turn and choose a posteriori the best system (e.g. Levy et al., 2013). Ideally, the fit to the

AMTD
8, 8509–8562, 2015

A. C. Povey and R. G. Grainger

Title Page
Abstract Introduction
Conclusions References
Tables Figures
◀◀ ▶▶
◀ ▶
Back Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

measurements would indicate a best choice of system, shown schematically in Fig. 4. Difficulty emerges when multiple systems produce values with indistinguishable fits to the measurements (e.g. the measurements can be fit equally well by a water cloud or thick aerosol haze). In either case, analogous to the 24 cm slice of Fig. 3, an unquantified error may be present due to deviations between the forward model and reality. This manner of reporting an ensemble of all the systems evaluated allows the error to be at least sampled.

3.6 Existing terminology

The combined impact of approximation, resolution, and system errors was defined as “structural uncertainty” by Thorne et al. (2005). Their emphasis was that the choices made by different investigators in the analysis of the same data can produce discrepancies. The terminology proposed above clarifies the type of choices which introduce such errors to an analysis and delineates by the manner in which they would be assessed. Regardless, this paper would prefer “structural error” as it is the error that is structural, not its uncertainty. The term “structural uncertainty” is used by Draper (1995) to describe system errors, though with respect to statistical rather than physical models.

4 Retrieval validation

Validation is a vital step in the production of any data set, confirming that the data and methodology are fit for their purpose. Often thought of as the conclusion of data generation, it provides guidance for future development of the algorithm and so is better considered a step in the cycle of retrieval development (see Fig. 5). Validation should be traceable and repeatable and can take two forms that will be discussed in this section:

- *Internal validation* – the comparison of measurements from a single instrument;

AMTD
8, 8509–8562, 2015

– *External validation* – the comparison of measurements with correlative measurements made by a different instrument.

These can be thought of as assessing the precision and accuracy of the retrieval, respectively, and can establish that the methodology produces physically consistent results. The process should demonstrate that new data are consistent with independent results, estimate the relative error between the techniques considered, and show that the predicted uncertainties accurately describe the distribution of that error.

This paper construes a validation as a comparison against real data only. There is use in evaluating the performance of an algorithm against simulated data, but that is considered a step in retrieval refinement (confirming it behaves as expected in controlled conditions) rather than a validation.

4.1 External validation

Users will be most familiar with external validation – the comparison of observations from two or more instruments. This focuses on quantifying the correlation and difference between data sets. While such validation activities are fundamental to the characterisation and minimisation of systematic errors, they should not be confused with a quantification of uncertainty. Validation techniques are neither universal (being dependant on the collocation criteria), internally consistent (as external data are used), nor transferable (being representative of only the conditions considered).

4.1.1 Weighting functions

When comparing two data sets, neither quantifies "the truth" (even when one is substantially more precise than the other). Both have associated errors, random and systematic, such that all that can be said is the products are consistent with each other. Also, simply because two measurements purport to quantify the same measurand does not mean they actually do. Weighting functions illustrate the difference in sensitivity between instruments.

AMTD
8, 8509–8562, 2015

Consider cloud top height (CTH). The entire cloud emits thermal radiation, much of which will be scattered or absorbed within the cloud. Radiation from the cloud observed by a satellite corresponds to photons that found an unimpeded path to TOA. Hence, a radiometer quantifies an average of the cloud's temperature profile weighted by the probability that a photon from that level can arrive at TOA. The distribution of the weight is known as the weighting function, and is sketched in red in Fig. 6a. Due to the lack of information about the vertical extent of the cloud, it is common to assume the cloud is infinitely thin (e.g. Poulsen et al., 2012) and the measurand would be more accurately described as the "effective cloud radiating height".

A very simple model of this situation assumes that radiation increases linearly with optical path τ measured in the direction away from the observer. That radiance is attenuated with the exponential of τ so the observed radiance R can be approximated as

$$R = a\tau e^{-\tau}, \tag{7}$$

where a is some constant. This function has a maximum at $\tau = 1$. This result approximately holds in more detailed calculations, such that a useful rule-of-thumb is that a radiance can be thought of as emanating from the level of the atmosphere at unit optical path.

The Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) is commonly used to validate CTH (e.g. Holz et al., 2008; Stengel et al., 2013). CALIOP measures the backscatter from a pulsed laser beam as a function of height, which is predominatly a function of the number of particles in the beam. CTH is identified by the rapid increase in signal at the edge of the cloud as particle density increases. This results in a weighting function that is substantially sharper and peaked at the physical top of the cloud (black in Fig. 6).

A direct comparison of these two products will find that radiometer-retrieved CTH are consistently lower than those from the lidar. To properly validate the satellite against the lidar, it is necessary to use the satellite's weighting function to calculate an "effective cloud radiating height" from the lidar profile. When measurements are compared, it must be done on a common basis.

More formally, a weighting function describes the dependence of a measurement on the underlying state. When the state chosen to describe a measurement is not an orthogonal basis of the observed state, a variable in the state vector will not uniquely determine an element of the true state. The relationship between the retrieved state and true state is expressed by the averaging kernel $\mathbf{A} = \partial\hat{\boldsymbol{x}}/\partial\boldsymbol{x}$, which satisfies

$$\boldsymbol{x} - \boldsymbol{x}_a = \mathbf{A}(\hat{\boldsymbol{x}} - \boldsymbol{x}_a) + \boldsymbol{\epsilon}', \tag{8}$$

where $\boldsymbol{\epsilon}'$ represents the action of $\boldsymbol{G}$ on $\boldsymbol{\epsilon}$.

Consider where $\boldsymbol{x}$ has two elements: the CTH and total optical thickness. In the lidar retrieval, these two variables are independent; $\mathbf{A}_{\mathrm{lidar}}$ is a unit matrix. In the radiometer retrieval, the CTH retrieved is a function of the optical depth profile and $\mathbf{A}_{\mathrm{rad}}$ contains off-diagonal elements. To illustrate, consider when an optically thin cloud ($\tau \ll 1$) lies above a thicker cloud (Fig. 6b). The lidar will identify CTH as the physical top of the thin cloud but the radiometer will retrieve a CTH between the clouds. As the upper cloud's thickness increases, the weighting function is increasingly dominated by the upper cloud. The retrieved CTH is dependent on the upper cloud's optical thickness. The averaging kernel would be

$$\mathbf{A}_{\mathrm{rad}} = \begin{pmatrix} 1 - \frac{\partial \mathrm{CTH}}{\partial \tau} & \frac{\partial \mathrm{CTH}}{\partial \tau} \\ 0 & 1 \end{pmatrix}. \tag{9}$$

The off-diagonal elements of the averaging kernel represent aspects of the state that cannot be resolved by the chosen basis and forward model. Here, a two-layer cloud cannot be properly represented when the basis only describes the properties of a single-layer cloud. The characterisation of an averaging kernel may require the use of an extended state vector and simulations with a more detailed model. (If the retrieval had been posed over that extended state vector, the averaging kernel would have been diagonal.)

AMTD
8, 8509–8562, 2015

A. C. Povey and R. G. Grainger

4.1.2 Comparing retrieved quantities

Retrievals will be compared over some collection of observations representing only a subset of the realisable state vectors (e.g. a SST product compared to ship-based measurements will only encapsulate the variation in SST over major shipping lanes rather than globally). As systematic errors are circumstantial, this collection represents only a sample of the complete distribution – just as the definition of a measurand frames how its value can be understood and used, the scope of a validation frames the understanding of systematic errors.

Towards the aim of repeatability, validation should be performed in a manner such that, if an additional source of data were introduced (e.g. a new instrument site or satellite orbit), the conclusions would not be expected to change. In the highly common case that there are insufficient data to achieve this, the scope of the validation should be clearly outlined.

One would naïvely judge if two retrievals are consistent by considering,

$$\chi^2 = (\boldsymbol{x}_1 - \boldsymbol{x}_2)^{\mathrm{T}}(\mathbf{S}_1 + \mathbf{S}_2)^{-1}(\boldsymbol{x}_1 - \boldsymbol{x}_2), \tag{10}$$

where $\mathbf{S}_i$ is the covariance of a retrieved solution. Rodgers and Connor (2003) noted that this does not apply for retrievals with differing averaging kernels, and developed an alternative formalism that will be briefly summarised. The collection of states compared is assumed to have a mean state $\boldsymbol{x}_c$ with covariance $\mathbf{S}_c$. This could be the mean of one of the data sets considered, or represent prior information, such as a climatology from a previous measurement campaign.

Equation (8) linearises the retrieved state about the a priori state. The two retrievals are unlikely to share an a priori. Hence, to consider compatible averaging kernels it is necessary to translate both data sets to a common linearisation point, for which $\boldsymbol{x}_c$ and $\mathbf{S}_c$ are suitable. The necessary translation is

$$\bar{\boldsymbol{x}}_i = \boldsymbol{x}_i - \boldsymbol{x}_c + (\mathbf{A}_i - \mathbf{I})(\boldsymbol{x}_{ai} - \boldsymbol{x}_c) \tag{11}$$

$$\equiv \mathbf{A}_i(\hat{\boldsymbol{x}} - \boldsymbol{x}_c) + \boldsymbol{\epsilon}'_i. \tag{12}$$

The difference between retrievals is then,

$$\boldsymbol{\delta} = \bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2 \equiv (\mathbf{A}_1 - \mathbf{A}_2)(\hat{\boldsymbol{x}} - \boldsymbol{x}_c) + \boldsymbol{\epsilon}'_1 - \boldsymbol{\epsilon}'_2, \tag{13}$$

which has covariance,

$$\mathbf{S}_\delta = (\mathbf{A}_1 - \mathbf{A}_2)^\mathrm{T} \mathbf{S}_c (\mathbf{A}_1 - \mathbf{A}_2) + \mathbf{S}_1 + \mathbf{S}_2. \tag{14}$$

Thus, rather then Eq. (10), an appropriate comparison metric is

$$\chi^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\mathrm{T} \mathbf{S}_\delta^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2). \tag{15}$$

This is useful for the comparison of optimal retrievals and is widely used in the trace gas community (e.g. Froidevaux et al., 2008; Wunch et al., 2010). It is less straightforward but equally important for any comparison of data products. Different algorithms will have distinct sensitivities to the same input information. Products from different sensors will consider different inputs, which will react differently to the unconstrained atmospheric states. Even where channels with similar wavelengths are used, they will have different band-passes which subtly affect their sensitivity (weighting functions). For example, the scattering properties of smaller droplets change more rapidly with wavelength than those of larger droplets. Thus, in Fig. 6b, a second radiometer with a wider band could have significantly different weighting functions depending on the droplet size in the upper cloud (because it transmits differently at the edges of the band). If the averaging kernel is not calculated, it is not possible to rigorously compare the data from different sensors, even from the same algorithm.

When one product is of much higher resolution, such as the comparison against CALIOP described in Sect. 4.1.1, it may be possible to transform it onto the basis of the other via

$$\bar{\boldsymbol{x}}_2^* = \boldsymbol{x}_c + \mathbf{A}_1 (\bar{\boldsymbol{x}}_2 - \boldsymbol{x}_c), \tag{16}$$

for which

$$\boldsymbol{\delta}^* = \bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2^* \equiv (\mathbf{A}_1 - \mathbf{A}_1\mathbf{A}_2)(\hat{\boldsymbol{x}} - \boldsymbol{x}_c) + \boldsymbol{\epsilon}'_1 - \mathbf{A}_1\boldsymbol{\epsilon}'_2, \tag{17}$$

AMTD
8, 8509–8562, 2015

A. C. Povey and R. G. Grainger

which has covariance,

$$\mathbf{S}_{\delta*} = (\mathbf{A}_1 - \mathbf{A}_1\mathbf{A}_2)\mathbf{S}_c(\mathbf{A}_1 - \mathbf{A}_1\mathbf{A}_2)^\mathrm{T} + \mathbf{S}_1 + \mathbf{A}_1\mathbf{S}_2\mathbf{A}_1^\mathrm{T}. \tag{18}$$

As Eq. (11) casts each observation on the same linearisation point, these techniques can be directly applied to the comparison of more than two instruments. When independent observations are not available to externally validate data, one can compare a product to model output provided the model is sampled as if viewed by a satellite. The retrieval's averaging kernel and weighting functions are necessary to translate the physical variables quantified by the model (e.g. particle number density) into the observed measurand.

4.1.3 Expected error envelopes

Expected error envelopes are a common means of presenting the result of a validation of, for example, aerosol optical depth τ (e.g. Kahn et al., 2005; Levy et al., 2010). The difference between the retrieved value and that reported by the Aerosol Robotic Network (AERONET) approximates the "error" in the retrieval. The "expected error envelope" is the width of the observed distribution of "error" and is described like an uncertainty. The value is an "envelope" because the distribution widens with increasing retrieved optical depth, such that the final value is reported as $\pm(a + b\tau)$, where a represents the minimum width of the "error" distribution and b represents the rate at which it widens with increasing optical depth.

This is an efficient means of communicating the results of the validation against AERONET and does convey a quantitative measure of the degree of certainty the data producer has in their product. It is not, though, an estimation of uncertainty. Validation techniques are neither universal (being dependant on the collocation criteria), internally consistent (as external data are used), nor transferable (being representative of only the conditions considered). Treating such values as an uncertainty has led to significant difficulty integrating data from different sensors as global and local sources of error

AMTD
8, 8509–8562, 2015

are disconnected, even when the envelopes are stratified by observing conditions and retrieval assumptions (Holzer-Popp et al., 2014).

This application of a single uncertainty value for all retrievals conveys an incorrect appreciation of the uncertainty to users as it implies well-constrained random and systematic components. Though stratification by relevant circumstances (e.g. over desert, high aerosol loading) indicates that the error depends on the state observed, a single number cannot usefully communicate the distribution of error in any particular measurement. Only pixel-level estimates provide an uncertainty consistent with its widely accepted definition and the presentation of ensembles, already used in the calculation of these envelopes, can better represent the distribution of errors not quantified in that estimate.

4.2 Internal validation

Internal validation is a less frequently discussed means to assess the precision and consistency of measurements.

4.2.1 Self-consistency

Repeated observations of an unchanged target should sample the distribution of error, such that a histogram of the observations should be Gaussian with a standard deviation equivalent to the uncertainty. An opportunity for this type of repeated observation is rare with satellite instruments. More common is the sampling of the same point in successive orbits (often near the poles), assembling pairs of measurements of similar (if not identical) atmospheric states (e.g. Lambert et al., 1996). If the first observation is x_1 with uncertainty σ_1 and the second x_2 with σ_2 then a histogram of

$$\Delta = \frac{x_1 - x_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \quad (19)$$

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

AMTD

8, 8509–8562, 2015

Uncertainty estimation in satellite remote sensing data

A. C. Povey and R. G. Grainger

should have a mean of zero and a standard deviation of unity. The covariance of simultaneously retrieved quantities can be considered by evaluating Eq. (10) instead.

Atmospheric variation may increase the observed variability so a larger standard deviation is not questionable. A variance less than one usually indicates an underestimation of the uncertainty. Significant departure from a Gaussian distribution is indicative of unidentified systematic errors. If the variable is expected to be homogeneous across a region, all observations there can be used to validate the uncertainty directly, as the variance of the observations should be greater than the average of the uncertainties.

4.2.2 Against other algorithms

Using different forward model assumptions, statistical techniques, and/or filtering methods can produce results that may be consistent with themselves and external validation but not with each other. Differences between retrievals, in the absence of external validation data or a programming error, indicate variations in the state within the unconstrained state space. They form an ensemble that illuminates where the formulation of the problem is most relevant, highlighting where future research could be concentrated to better represent the observations (Holzer-Popp et al., 2013). Belief that one representation is "better" than others independent of external validation is an expression of a priori knowledge. Such knowledge can be very useful in identifying "unknown unknowns" in a retrieval, but it is important to appreciate that any constraint not made by the data is an expression of a priori data, be it as formal as knowing that surface temperatures are generally within 40 of 10 °C or as simple as believing surface pressure shouldn't vary across a land–sea boundary.

5 Communication with users

Confidence in data is communicated to users through uncertainty estimates and quality assurance statements. The quantification of uncertainty illustrates how new data relate to the existing body of knowledge, but there is also the user's qualitative sense of the "worth" of data. To what extent does it constrain the variables they are investigating? When and where are the data most robust and when and where do they effectively convey no information? What do they quantify that was not already known? The aims of the user frame these questions. A detailed case study requires reliable uncertainty estimates to incorporate varied measurements and understand the limitations of the information provided but it is impractical for a twenty-year model climatology to consider a single measurement, its uncertainty even more so.

Further, the "unknown unknowns" affecting satellite remote sensing data are not completely indescribable. Information such as "results are often unreliable over deserts" is still important to users, even if the uncertainty can't be quantified. A dialogue with users is important in improving the understanding of data and receiving feedback on that data for future improvement.

5.1 Error budget

The aim of an error budget is to classify the contributions to the uncertainty by their source. At its simplest this may be in the form of a table, as suggested in Table 2. The total uncertainty estimated in this way can be compared with that found through validation activities. Discrepancy between the two can potentially indicate an error source has been over-looked.

5.2 Quality assurance

Quality assurance (or flagging) is a qualitative judgement of the performance of a retrieval and the suitability of that technique for processing the data. This complements the uncertainty, whose calculation assumes that the forward model is appropriate to the observed circumstances. Statistical distributions are unsuited to show when an algorithm fails to converge, converges to an unphysical state, encounters incomprehensible data, or observes circumstances beyond the ability of its model to describe. Provided it

AMTD
8, 8509–8562, 2015

is described in the language of a statement of confidence, quality assurance provides useful information.

The difficulty is that a simple flag is a coarse means of communication. For example, MODIS aerosol products provide a data quality flag that takes values 0, 1, 2, or 3 to describe increasing confidence in the retrieval method (Sect. 2.5, Remer et al., 2006). This is widely used as a simple filter, rejecting data below some level. The level selected varies widely and it neglects, for example, that all low magnitude retrievals have confidence 1 due to the small signal. This will bias analyses to circumstances ideal for the chosen formulation, which aren't necessarily representative of the environment (Sect. 3.4.2).

However, such filtering is a logical response to this presentation of information. A more useful scheme would provide multiple separate flags (e.g. presence of cloud, challenging surface conditions, failure to converge, etc) in a bit mask. When these are properly documented they allow an attentive user to evaluate the impact of using data degraded by a specific feature and the disinterested user may be inspired to briefly consider the most appropriate flags for their purposes.

5.3 Algorithm maturity

Satellite remote sensing data have existed for several decades, but the retrieved geophysical quantities evolve as additional auxiliary data become available and new scientific problems appear. For example, AVHRR measurements from 1978 are still reprocessed for climate studies (Stengel et al., 2013; Heidinger et al., 2014). Figure 5 outlines the interlinking cycles of algorithm and operational development. Figure 7 illustrates how the repeated refinement and validation of data is a fundamental expression of the scientific method in data analysis. The cycle describes the ongoing conversation through which measurements and algorithms are improved in response to their use until a consensus is built that either:

1. the data set sufficiently addresses the needs of its users; or

AMTD
8, 8509–8562, 2015

2. the maximal amount of information has been extracted from the measurement and additional information is required to meet the needs of users.

The progress of a data set from initial conception to the achievement of one of these goals is known as its *maturity*.

Bates and Barkstrom (2006) and Bates and Privette (2012) have outlined the system maturity matrix as a standardised metric to quantify the maturity of a product, briefly summarised in Table 3. It provides a means to track the development of an algorithm and data set from initial concept to an operational setting, highlighting areas of a project that could benefit from additional resources to achieve increased impact. The CORE-CLIMAX project (Coordinating Earth observation data validation for re-analysis for climate services) has adapted and implemented such a scheme to rate the suitability of current data products for use as a Climate Data Record (CDR), introduced in Table 4. These matrices concentrate on goal 1 above, specifically the ability for "end-users to realize the strengths and weaknesses of the dataset" (Work Package 2, 2013).

The appropriate presentation of data with thorough documentation and metadata produced using a publicly available, consistently realised computer code is a desirable aim. Such features should be included in any algorithm from inception to minimise simple mistakes and the misunderstanding of data by users. However, the presence of such features does not address the scientific quality or importance of the data.

The proposed metric simply counts the citations the data has received, disregarding the variety of applications and their impact upon scientific understanding. Participation in international data assessments works towards this aim, but only when there are multiple means of observing or evaluating a measurand. These are not available for many environmental variables and they should not be considered immature if they make the best use of the information available (goal 2).

It is important that an inexperienced user should not misinterpret data with a high maturity index as being more accurate or suited to a particular study. A mature data set is one which is near the end of its development cycle in that it is agreed to be fit-

AMTD
8, 8509–8562, 2015

for-purpose by the scientific community. This must not be confused with a data set that fully constrains the measurand.

With specific regard to the evaluation of uncertainty:

- As discussed in Sect. 3.1, SI traceability is not possible for a satellite instrument in the traditional meaning of that phrase. The environmental science community as a whole must develop ground-based, traceable standards for satellite instruments, such as well-characterised and monitored surfaces. The current metric penalises products that have no such standard to reference.

- The spatial covariance of error in a product can only be quantified through validation against spatially distributed, independent data. Satellite remote sensing is used for many environmental products because they are impractical to measure from the ground. In such cases it is not possible to independently assess covariance errors. Ensemble techniques may be useful there.

- A distinction must be made between internal and external validation activities. An international assessment of multiple, independent products from different measurement techniques that quantify equivalent measurands represents the external validation of a mature research area. An internal validation of differing algorithms from the same sensor evaluates the relative properties of the algorithms, not their suitability for quantifying the measurand.

Monitoring the progress of algorithm development must be done in a manner which encourages researchers to follow the fundamental scientific method (Fig. 7) whereby the interpretation of geophysical properties or processes is underpinned by a description of instrument calibration, the retrieval algorithm, and product validation. Maturity is an expression of confidence, not uncertainty, and should use appropriate language.

AMTD
8, 8509–8562, 2015

6 Conclusions

An appreciation of the range of values consistent with a measurement is necessary to properly apply and contextualise data. Three qualities were identified by the Guide to Uncertainty in Measurement (Working Group 1, 2008) as necessary for an expression of uncertainty to be useful:

- *universality*: all manners of observation can apply the techniques to calculate their uncertainty;
- *internal consistency*: the calculation of uncertainty requires no information beyond that used in the analysis;
- *transferability*: the uncertainty must be of use to a data user.

This paper classifies errors affecting satellite remote sensing data with five groups:

- *measurement*: intrinsic variability in the observation;
- *parameter*: errors propagated from auxiliary data;
- *approximation*: explicit simplifications in the formulation of the forward model;
- *system*: differences between the chosen description of the environment and reality;
- *resolution*: variability at scales smaller than that observed.

In the terminology of Thorne et al. (2005), the first two result in parametric errors and the remainder structural errors.

Measurement and parameter errors are generally well represented by the traditional propagation of random perturbations through an analysis. These are useful but only describe one aspect of the uncertainty – the “unknowns” that are known and quantifiable. Approximation and system errors represent the inability of the analysis to describe the

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

AMTD

8, 8509–8562, 2015

Uncertainty estimation in satellite remote sensing data

A. C. Povey and R. G. Grainger

environment observed and are the dominant source of error in most passive satellite remote sensing data (as it is not possible to constrain the complex behaviour of the environment with a few TOA radiances). Data producers are aware of these additional "unknowns", such as the representation of the surface's bi-directional reflectance, but cannot quantify them in the manner required for traditional error propagation (i.e. they are known, unquantifiable unknowns). Even well-constrained analyses will be affected by system errors resulting from quality control, cloud filtering being the most common. Resolution errors describe the disconnect between what occurs in nature and the means by which it is observed, primarily resulting from the instrument's sampling.

The difficulty with the last three categories of error is that they can be highly non-linear – their magnitude and nature depend upon the state observed and the ability of the forward model to describe it. Propagation of errors assumes that the equations used are accurate and that errors affect them linearly. Uncertainties currently reported with satellite remote sensing data neither represent the actual (non-linear) distribution of errors nor the full range of information known about the errors.

This can be addressed in various ways. Firstly, uncertainty estimates in satellite remote sensing data must be presented at pixel level. Pervasive quantifications misrepresent the dependence of error upon state and rely on external information. While pixel-level estimates will not represent the impact of unquantified unknowns, it is important that uncertainty be presented in a context that represents the data producer's confidence in and understanding of their data.

Ensemble techniques can be used to represent unquantifiable unknowns. The underconstrained nature of many satellite observations means that multiple realisations of a data set that are consistent with measurements can be derived by using conflicting descriptions of the environment, such as assumptions of particle microphysical properties or differing calibration coefficients. In the absence of a priori constraints, each of these realisations is feasible and should be presented together. This is common practice in the climate modelling community and the satellite remote sensing commu-

AMTD
8, 8509–8562, 2015

nity should capitalise on user's experience to better communicate the uncertainty in products.

The manner in which a measurand is defined affects both the sources of error that must be considered (e.g. resolution errors) and the manner in which the data must be compared with other measurements. In an underconstrained problem, it is often not possible to report a value that is uniquely constrained by those conditions (i.e. the state vector is not an orthogonal basis of the observed conditions). This can result in the retrieved value being sensitive to multiple features of the environment, as quantified by the averaging kernel. When comparing data sets, it is important to ensure that equivalent quantities are being compared or biases will be observed that are a function of the system definition rather than an error in the retrieval. The necessary transforms were outlined in Rodgers and Connor (2003).

As not all errors can be quantified, there is also qualitative information necessary to appreciate the applicability of data and, as a data set evolves, it is important to assess both the degree to which it represents a scientific advancement and to which it satisfies the needs of its users. This information can be conveyed through product user guides, validation studies, quality assurance flags, and/or measures of a retrieval system's maturity. It is both important that this information is readily available to users and that it is communicated in the language of a statement of confidence. Continuous interaction with users will be necessary to improve these reports to ensure they communicate the desired information. Of particular importance is

– An error budget outlining the quantified sources of error;

– A description of the available quality control information and its physical meaning to enable users to apply it in an educated fashion;

– Known weaknesses of the data that are not represented by the uncertainty.

Evaluating the quality of an algorithm using existing metrics limits the ability of the satellite remote sensing community to communicate their understanding of the uncertainties in their products to users in an efficient or effective manner. Without that

AMTD
8, 8509–8562, 2015

dialogue, users cannot appropriately use data and cannot feedback to data producers to improve it. The hope is that by representing uncertainties in satellite remote sensing data through ensembles, understanding of the limitations of the data will increase, highlighting areas for future research. Through continual communication among the entire scientific community, unknown unknowns can become known and, eventually, make the usage of ensembles unnecessary as understanding of the environment converges upon the truth.

Acknowledgements. This research was supported in part by the aerosol and cloud components of the ESA CCI and in part through the Natural Environment Research Council's support of the National Centre for Earth Observation. For their inspirational and insightful conversations, thanks must be given to the participants of the Aerosol CCI uncertainty workshop on 4 September 2014; the AeroSat meeting on 27–28 September 2014; and SST CCI Uncertainty Workshop of 18–20 November 2014. The authors are indebted to Claire Bulgin, Gerrit de Leeuw, Thomas Holzer-Popp, John Kennedy, Greg McGarragh, and Chris Merchant for their useful comments.

References

Ablain, M., Cazenave, A., Larnicol, G., Balmaseda, M., Cipollini, P., Faugère, Y., Fernandes, M. J., Henry, O., Johannessen, J. A., Knudsen, P., Andersen, O., Legeais, J., Meyssignac, B., Picot, N., Roca, M., Rudenko, S., Scharffenberg, M. G., Stammer, D., Timms, G., and Benveniste, J.: Improved sea level record over the satellite altimetry era (1993–2010) from the Climate Change Initiative project, Ocean Sci., 11, 67–82, doi:10.5194/os-11-67-2015, 2015. 8519

Ackerman, S. A., Strabala, K. I., Menzel, W. P., Frey, R. A., Moeller, C. C., and Gumley, L. E.: Discriminating clear sky from clouds with MODIS, J. Geophys. Res., 103, 32141–32157, doi:10.1029/1998JD200032, 1998. 8525

Anderson, T. L., Charlson, R. J., Winker, D. M., Ogren, J. A., and Holmén, K.: Mesoscale variations of tropospheric aerosols, J. Atmos. Sci., 60, 119–136, doi:10.1175/1520-0469(2003)060<0119:MVOTA>2.0.CO;2, 2003. 8524

AMTD
8, 8509–8562, 2015

Uncertainty estimation in satellite remote sensing data

A. C. Povey and R. G. Grainger

Barnes, W., Pagano, T., and Salomonson, V.: Prelaunch characteristics of the Moderate Resolution Imaging Spectroradiometer (MODIS) on EOS-AM1, IEEE T. Geosci. Remote, 36, 1088–1100, doi:10.1109/36.700993, 1998. 8521

Bates, J. J. and Barkstrom, B. R.: A maturity model for satellite-derived climate data records, in: 14th Conference on Satellite Meteorology and Oceanography, Atlanta, GA, USA, p. 2.11, available at: ams.confex.com/ams/Annual2006/techprogram/paper_100658.htm (last access: 7 August 2015), 2006. 8538, 8554

Bates, J. J. and Privette, J. L.: A maturity model for assessing the completeness of climate data records, EOS T. Am. Geophys. Un., 93, 441, doi:10.1029/2012EO440006, 2012. 8538

Buizza, R., Houtekamer, P. L., Pellerin, G., Toth, Z., Zhu, Y., and Wei, M.: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems, Mon. Weather Rev., 133, 1076–1097, doi:10.1175/MWR2905.1, 2005. 8517

CCI project teams: CCI project guidelines, Tech. Rep. EOP-DTEX-EOPS-SW-10-0002, European Space Agency, Frascati, Italy, available at: ionia1.esrin.esa.int/files/ESA_CCI_Project_Guidlines_V1.pdf(last access: 7 August 2015), 2010. 8514

Chambers, L. H., Wielicki, B. A., and Evans, K. F.: Accuracy of the independent pixel approximation for satellite estimates of oceanic boundary layer cloud optical depth, J. Geophys. Res.-Atmos., 102, 1779–1794, doi:10.1029/96JD02995, 1997. 8524

Chase, R. R.: Report of the EOS data panel, Earth observing system, data and information system, Data Panel Report, Vol. IIa, NASA Technical Memorandum 87777, NASA, Hampton, VA, USA, available at: http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19860021622.pdf(last access: 7 August 2015), 1986. 8552

Crucifix, M., Braconnot, P., Harrison, S., and Otto-Bliesner, B.: Second phase of paleoclimate modelling intercomparison project, EOS T. Am. Geophys. Un., 86, 264, doi:10.1029/2005EO280003, 2005. 8517

Curier, L., de Leeuw, G., Kolmonen, P., Sundström, A.-M., Sogacheva, L., and Bennouna, Y.: Aerosol retreival over land using the (A)ATSR dual-view algorithm, in: Satellite Aerosol Remote Sensing Over Land, edited by: Kokhanovsky, A. and de Leeuw, G., Springer, Berlin, Germany, 135–160, 2009. 8525

Draper, D.: Assessment and propagation of model uncertainty, J. Roy. Stat. Soc. B, 57, 45–97, available at: www.jstor.org/stable/2346087(last access: 7 August 2015), 1995. 8527

Ducher, G.: Cartographic possibilities of the SPOT and Spacelab projects, Photogramm. Rec., 10, 167–180, doi:10.1111/j.1477-9730.1980.tb00019.x, 1980. 8511

Fischer, H., Birk, M., Blom, C., Carli, B., Carlotti, M., von Clarmann, T., Delbouille, L., Dudhia, A., Ehhalt, D., Endemann, M., Flaud, J. M., Gessner, R., Kleinert, A., Koopman, R., Langen, J., López-Puertas, M., Mosner, P., Nett, H., Oelhaf, H., Perron, G., Remedios, J., Ridolfi, M., Stiller, G., and Zander, R.: MIPAS: an instrument for atmospheric and climate research, Atmos. Chem. Phys., 8, 2151–2188, doi:10.5194/acp-8-2151-2008, 2008. 8522

Fischer, J. and Grassl, H.: Detection of cloud-top height from backscattered radiances within the oxygen A band. Part 1: Theoretical study, J. Appl. Meteor., 30, 1245–1259, doi:10.1175/1520-0450(1991)030<1245:DOCTHF>2.0.CO;2, 1991. 8516

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C., and Rummukainen, M.: Evaluation of climate models, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, UK and New York, NY, USA, 741–866, available at: http://www.ipcc.ch/pdf/assessment-report/ar5/wg1/WG1AR5_Chapter09_FINAL.pdf(last access: 7 August 2015), 2013. 8517

Fougnie, B., Bracco, G., Lafrance, B., Ruffel, C., Hagolle, O., and Tinel, C.: PARASOL in-flight calibration and performance., Appl. Optics, 46, 5435–5451, doi:10.1364/AO.46.005435, 2007. 8521

Froidevaux, L., Jiang, Y. B., Lambert, A., Livesey, N. J., Read, W. G., Waters, J. W., Browell, E. V., Hair, J. W., Avery, M. A., Mcgee, T. J., Twigg, L. W., Sumnicht, G. K., Jucks, K. W., Margitan, J. J., Sen, B., Stachnik, R. A., Toon, G. C., Bernath, P. F., Boone, C. D., Walker, K. A., Filipiak, M. J., Harwood, R. S., Fuller, R. A., Manney, G. L., Schwartz, M. J., Daffer, W. H., Drouin, B. J., Cofield, R. E., Cuddy, D. T., Jarnot, R. F., Knosp, B. W., Perun, V. S., Snyder, W. V., Stek, P. C., Thurstans, R. P., and Wagner, P. A.: Validation of Aura Microwave Limb Sounder stratospheric ozone measurements, J. Geophys. Res.-Atmos., 113, 1–24, doi:10.1029/2007JD008771, 2008. 8532

Grandey, B. S. and Stier, P.: A critical look at spatial scale choices in satellite-based aerosol indirect effect studies, Atmos. Chem. Phys., 10, 11459–11470, doi:10.5194/acp-10-11459-2010, 2010. 8524

Heidinger, A. K., Sullivan, J. T., and Nagaraja Rao, C. R.: Calibration of visible and near-infrared channels of the NOAA-12 AVHRR using time series of observations over deserts, Int. J. Remote Sens., 24, 3635–3649, doi:10.1080/0143116021000023907, 2003. 8521

AMTD

8, 8509–8562, 2015

A. C. Povey and R. G. Grainger

Heidinger, A. K., Foster, M. J., Walther, A., and Zhao, X. T.: The pathfinder atmospheres–extended AVHRR climate dataset, B. Am. Meteorol. Soc., 95, 909–922, doi:10.1175/BAMS-D-12-00246.1, 2014. 8537

Hickey, J. R. and Karoli, A. R.: Radiometric calibrations for the earth radiation budget experiment, Appl. Optics, 13, 523–533, doi:10.1364/AO.13.000523, 1974. 8521

Holz, R. E., Ackerman, S. A., Nagle, F. W., Frey, R., Dutcher, S., Kuehn, R. E., Vaughan, M. A., and Baum, B.: Global Moderate Resolution Imaging Spectroradiometer (MODIS) cloud detection and height evaluation using CALIOP, J. Geophys. Res.-Atmos., 113, D00A19, doi:10.1029/2008JD009837, 2008. 8529

Holzer-Popp, T., de Leeuw, G., Griesfeller, J., Martynenko, D., Klüser, L., Bevan, S., Davies, W., Ducos, F., Deuzé, J. L., Graigner, R. G., Heckel, A., von Hoyningen-Hüne, W., Kolmonen, P., Litvinov, P., North, P., Poulsen, C. A., Ramon, D., Siddans, R., Sogacheva, L., Tanre, D., Thomas, G. E., Vountas, M., Descloitres, J., Griesfeller, J., Kinne, S., Schulz, M., and Pinnock, S.: Aerosol retrieval experiments in the ESA Aerosol_cci project, Atmos. Meas. Tech., 6, 1919–1957, doi:10.5194/amt-6-1919-2013, 2013. 8535

Holzer-Popp, T., Kahn, R., de Leeuw, G., Munchak, L. A., Pinnock, S., Povey, A. C., Sayer, A. M., and Thomas, G. E.: Minutes of pixel-level uncertainty discussion, in: AEROSAT 2, Steamboat Springs, CO, USA, 1–3, available at: http://www.aero-sat.org/aero-sat-meeting-2.html(last access: 7 August 2015), 2014. 8534

Holzer-Popp, T., de Leeuw, G., and Martynenko, D.: Phase 1 Final report, Tech. rep., ESA Climate Change Initiative: Aerosol, Frascati, Italy, 2015. 8526

Houtekamer, P. L. and Lefaivre, L.: Using ensemble forecasts for model validation, Mon. Weather Rev., 125, 2416–2426, doi:10.1175/1520-0493(1997)125<2416:UEFFMV>2.0.CO;2, 1997. 8517

Kahn, R. A., Gaitley, B. J., Martonchik, J. V., Diner, D. J., Crean, K. A., and Holben, B.: Multiangle Imaging Spectroradiometer (MISR) global aerosol optical depth validation based on 2 years of coincident Aerosol Robotic Network (AERONET) observations, J. Geophys. Res.-Atmos., 110, 1–16, doi:10.1029/2004JD004706, 2005. 8533

Kennedy, J. J., Rayner, N. A., Smith, R. O., Parker, D. E., and Saunby, M.: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties, J. Geophys. Res.-Atmos., 116, D14103, doi:10.1029/2010JD015218, 2011a. 8520

Kennedy, J. J., Rayner, N. A., Smith, R. O., Parker, D. E., and Saunby, M.: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization, J. Geophys. Res.-Atmos., 116, D14104, doi:10.1029/2010JD015220, 2011b. 8520

King, M. D.: Remote sensing of cloud, aerosol, and water vapor properties from MODIS, IEEE T. Geosci. Remote, 30, 2–27, doi:10.1109/36.124212, 1992. 8516

Kokhanovsky, A. A., Deuzé, J. L., Diner, D. J., Dubovik, O., Ducos, F., Emde, C., Garay, M. J., Grainger, R. G., Heckel, A., Herman, M., Katsev, I. L., Keller, J., Levy, R., North, P. R. J., Prikhach, A. S., Rozanov, V. V., Sayer, A. M., Ota, Y., Tanré, D., Thomas, G. E., and Zege, E. P.: The inter-comparison of major satellite aerosol retrieval algorithms using simulated intensity and polarization characteristics of reflected light, Atmos. Meas. Tech., 3, 909–932, doi:10.5194/amt-3-909-2010, 2010. 8515

Kummerow, C., Simpson, J., Thiele, O., Barnes, W., Chang, A. T. C., Stocker, E., Adler, R. F., Hou, A., Kakar, R., Wentz, F., Ashcroft, P., Kozu, T., Hong, Y., Okamoto, K., Iguchi, T., Kuroiwa, H., Im, E., Haddad, Z., Huffman, G., Ferrier, B., Olson, W. S., Zipser, E., Smith, E. A., Wilheit, T. T., North, G., Krishnamurti, T., and Nakamura, K.: The status of the Tropical Rainfall Measuring Mission (TRMM) after two years in orbit, J. Appl. Meteorol., 39, 1965–1982, doi:10.1175/1520-0450(2001)040<1965:TSOTTR>2.0.CO;2, 2000. 8521

Kuze, A., Taylor, T. E., Kataoka, F., Bruegge, C. J., Crisp, D., Harada, M., Helmlinger, M., Inoue, M., Kawakami, S., Kikuchi, N., Mitomi, Y., Murooka, J., Naitoh, M., O'Brien, D. M., O'Dell, C. W., Ohyama, H., Pollock, H., Schwandner, F. M., Shiomi, K., Suto, H., Takeda, T., Tanaka, T., Urabe, T., Yokota, T., and Yoshida, Y.: Long-term vicarious calibration of GOSAT short-wave sensors: techniques for error reduction and new estimates of radiometric degradation factors, IEEE T. Geosci. Remote, 52, 3991–4004, doi:10.1109/TGRS.2013.2278696, 2014. 8521

Lambert, A. L., Grainger, R., Remedios, J., Reburn, W., Rodgers, C., Taylor, F., Roche, A., Kumer, J., Massie, S., and Deshler, T.: Validation of aerosol measurements from the Improved Stratospheric and Mesospheric Sounder, J. Geophys. Res., 101, 9811–9830, doi:10.1029/95JD01702, 1996. 8534

Levy, R. C., Remer, L. A., Kleidman, R. G., Mattoo, S., Ichoku, C., Kahn, R., and Eck, T. F.: Global evaluation of the Collection 5 MODIS dark-target aerosol products over land, Atmos. Chem. Phys., 10, 10399–10420, doi:10.5194/acp-10-10399-2010, 2010. 8533

AMTD

8, 8509–8562, 2015

Levy, R. C., Mattoo, S., Munchak, L. A., Remer, L. A., Sayer, A. M., Patadia, F., and Hsu, N. C.: The Collection 6 MODIS aerosol products over land and ocean, Atmos. Meas. Tech., 6, 2989–3034, doi:10.5194/amt-6-2989-2013, 2013. 8526

Li, Z., Zhao, X., Kahn, R., Mishchenko, M., Remer, L., Lee, K.-H., Wang, M., Laszlo, I., Nakajima, T., and Maring, H.: Uncertainties in satellite remote sensing of aerosols and impact on monitoring its long-term trend: a review and perspective, Ann. Geophys., 27, 2755–2770, doi:10.5194/angeo-27-2755-2009, 2009. 8515

Liu, W., Huang, B., Thorne, P. W., Banzon, V. F., Zhang, H.-M., Freeman, E., Lawrimore, J., Peterson, T. C., Smith, T. M., and Woodruff, S. D.: Extended reconstructed sea surface temperature version 4 (ERSST.v4): Part II. Parametric and structural uncertainty estimations, 4, J. Climate, 931–951, doi:10.1175/JCLI-D-14-00007.1, 2015. 8519

Liu, Y., Chen, D., Kahn, R. A., and He, K.: Review of the applications of Multiangle Imaging SpectroRadiometer to air quality research, Sci. China Ser. D, 52, 132–144, doi:10.1007/s11430-008-0149-6, 2009. 8519

Lorenz, E. N.: A study of the predictability of a 28-variable atmospheric model, Tellus A, 17, 321–333, doi:10.3402/tellusa.v17i3.9076, 1965. 8517

Maritorena, S. and Siegel, D. A.: Consistent merging of satellite ocean color data sets using a bio-optical model, Remote Sens. Environ., 94, 429–440, doi:10.1016/j.rse.2004.08.014, 2005. 8519

Mears, C. A., Wentz, F. J., Thorne, P., and Bernie, D.: Assessing uncertainty in estimates of atmospheric temperature changes from MSU and AMSU using a Monte-Carlo estimation technique, J. Geophys. Res.-Atmos., 116, 1–16, doi:10.1029/2010JD014954, 2011. 8519

Meehl, G., Boer, G., Covey, C., Latif, M., and Stouffer, R.: The Coupled Model Intercomparison Project (CMIP), B. Am. Meteorol. Soc., 81, 313–318, doi:10.1175/1520-0477(2000)081<0313:TCMIPC>2.3.CO;2, 2000. 8517

Munechika, C. K., Warnick, J. S., Salvaggio, C., and Schott, J. R.: Resolution enhancement of multispectral image data to improve classification accuracy, Photogramm. Eng. Rem. S., 59, 67–72, 1993. 8523

Pavolonis, M. J. and Heidinger, A. K.: Daytime cloud overlap detection from AVHRR and VIIRS, J. Appl. Meteorol., 43, 762–778, doi:10.1175/2099.1, 2004. 8525

Poulsen, C. A., Siddans, R., Thomas, G. E., Sayer, A. M., Grainger, R. G., Campmany, E., Dean, S. M., Arnold, C., and Watts, P. D.: Cloud retrievals from satellite data using opti-

AMTD

8, 8509–8562, 2015

mal estimation: evaluation and application to ATSR, Atmos. Meas. Tech., 5, 1889–1910, doi:10.5194/amt-5-1889-2012, 2012. 8529

Powell, K. A., Hostetler, C. A., Liu, Z., Vaughan, M. A., Kuehn, R. E., Hunt, W. H., Lee, K. P., Trepte, C. R., Rogers, R. R., Young, S. A., and Winker, D. M.: CALIPSO lidar calibration algorithms. Part I: Nighttime 532-nm parallel channel and 532-nm perpendicular channel, J. Atmos. Ocean. Tech., 26, 2015–2033, doi:10.1175/2009JTECHA1242.1, 2009. 8521

Privette, J. L., Fowler, C., Wick, G. A., Baldwin, D., and Emery, W. J.: Effects of orbital drift on advanced very high resolution radiometer products: normalized difference vegetation index and sea surface temperature, Remote Sens. Environ., 53, 164–171, doi:10.1016/0034-4257(95)00083-D, 1995. 8524

Rayner, N. A., Merchant, C. J., Corlett, G. K., Mittaz, J., Bulgin, C., Atkinson, C. P., Good, S. A., and Kennedy, J. J.: Sea surface temperature user workshop on uncertainty, Tech. rep., ESA SST CCI, Frascati, Italy, available at: http://www.esa-sst-cci.org/PUG/pdf/CombinedSSTUserWorkshopReport.pdf(last access: 7 August 2015), 2014. 8519

Remer, L. A., Tanré, D., and Kaufman, Y. J.: Algorithm for remote sensing of tropospheric aerosol from MODIS: collection 5, Tech. Rep. MOD04/MYD04, NASA Goddard Space Flight Center, Greenbelt, MD, USA, available at: http://modis.gsfc.nasa.gov/data/atbd/atbd_mod02.pdf(last access: 7 August 2015), 2006. 8537

Rodgers, C. D.: Inverse Methods for Atmospheric Sounding: Theory and Practice, vol. 2, 2nd edn., World Scientific, Singapore, 2000. 8513, 8522, 8525

Rodgers, C. D. and Connor, B. J.: Intercomparison of remote sounding instruments, J. Geophys. Res., 108, 4116, doi:10.1029/2002JD002299, 2003. 8531, 8542

Sayer, A. M., Thomas, G. E., and Grainger, R. G.: A sea surface reflectance model for (A)ATSR, and application to aerosol retrievals, Atmos. Meas. Tech., 3, 813–838, doi:10.5194/amt-3-813-2010, 2010a. 8524

Sayer, A. M., Thomas, G. E., Palmer, P. I., and Grainger, R. G.: Some implications of sampling choices on comparisons between satellite and model aerosol optical depth fields, Atmos. Chem. Phys., 10, 10705–10716, doi:10.5194/acp-10-10705-2010, 2010b. 8524

Schiffer, R. and Rossow, W.: The International Satellite Cloud Climatology Project (ISCCP) – the first project of the World Climate Research Programme, B. Am. Meteorol. Soc., 64, 779–784, available at: http://rda.ucar.edu/datasets/ds742.0/docs/1983.SchifferRossow.pdf (last access: 7 August 2015), 1983. 8516

Slater, P. N., Biggar, S. F., Thome, K. J., Gellman, D. I., and Spyak, P. R.: Vicarious radiometric calibrations of EOS sensors, J. Atmos. Ocean. Tech., 13, 349–359, doi:10.1175/1520-0426(1996)013<0349:VRCOES>2.0.CO;2, 1996. 8521

Smith, D. L., Mutlow, C. T., and Nagaraja, R. C. R.: Calibration monitoring of the visible and near-infrared channels of the Along-Track Scanning Radiometer-2 by use of stable terrestrial sites, Appl. Optics, 41, 515–523, doi:10.1364/AO.41.000515, 2002. 8521

Smith, D. L., Mutlow, C. T., Delderfield, J., Watkins, B., and Mason, G.: ATSR infrared radiometric calibration and in-orbit performance, Remote Sens. Environ., 116, 4–16, doi:10.1016/j.rse.2011.01.027, 2012. 8521

Stengel, M., Mieruch, S., Jerg, M., Karlsson, K.-G., Scheirer, R., Maddux, B., Meirink, J., Poulsen, C., Siddans, R., Walther, A., and Hollmann, R.: The clouds climate change initiative: assessment of state-of-the-art cloud property retrieval schemes applied to AVHRR heritage measurements, Remote Sens. Environ., 162, 363–379, doi:10.1016/j.rse.2013.10.035, 2013. 8529, 8537

Stowe, L. L., Davis, P. A., and McClain, E. P.: Scientific basis and initial evaluation of the CLAVR-1 global clear/cloud classifcation algorithm for the advanced very high resolution radiometer, J. Atmos. Ocean. Tech., 16, 656–681, doi:10.1175/1520-0426(1999)016<0656:SBAIEO>2.0.CO;2, 1999. 8525

Tanelli, S., Durden, S. L., Im, E., Pak, K. S., Reinke, D. G., Partain, P., Haynes, J. M., and Marchand, R. T.: CloudSat's cloud profiling radar after two years in orbit: performance, calibration, and processing, IEEE T. Geosci. Remote, 46, 3560–3573, doi:10.1109/TGRS.2008.2002030, 2008. 8521

Thomas, G. E., Poulsen, C. A., Sayer, A. M., Marsh, S. H., Dean, S. M., Carboni, E., Siddans, R., Grainger, R. G., and Lawrence, B. N.: The GRAPE aerosol retrieval algorithm, Atmos. Meas. Tech., 2, 679–701, doi:10.5194/amt-2-679-2009, 2009. 8519

Thorne, P. W., Parker, D. E., Christy, J. R., and Mears, C. A.: Uncertainties in climate trends: lessons from upper-air temperature records, B. Am. Meteorol. Soc., 86, 1437–1442, doi:10.1175/BAMS-86-10-1437, 2005. 8527, 8540

Twomey, S.: Introduction to the Mathematics of Inversion in Remote Sensing and Indirect Measurements, Dover Publications, Inc., Amsterdam, the Netherlands, 1997. 8513

Work Package 2: Protocol for verifying, monitoring, calibrating and validating FCDRs and TCDRs of the CRDs/ECVs, Tech. Rep. D331, CORE-CLIMAX, Enschede, the Netherlands, available at: http://www.coreclimax.eu/sites/coreclimax.itc.nl/files/documents/Deliverables/

AMTD
8, 8509–8562, 2015
Uncertainty estimation in satellite remote sensing data
A. C. Povey and R. G. Grainger

WP_Reports/Deliverable-D331-CORECLIMAX.pdf (last access: 7 August 2015), 2013. 8538, 8555

Working Group 1: Evaluation of measurement data – guide to the expression of uncertainty in measurement, Tech. Rep. JCGM 100:2008, Joint Committee for Guides in Metrology, Paris, France, available at: http://www.iso.org/sites/JCGM/GUM-introduction.htm (last access: 7 August 2015), 2008. 8513, 8514, 8540

Wunch, D., Toon, G. C., Wennberg, P. O., Wofsy, S. C., Stephens, B. B., Fischer, M. L., Uchino, O., Abshire, J. B., Bernath, P., Biraud, S. C., Blavier, J.-F. L., Boone, C., Bowman, K. P., Browell, E. V., Campos, T., Connor, B. J., Daube, B. C., Deutscher, N. M., Diao, M., Elkins, J. W., Gerbig, C., Gottlieb, E., Griffith, D. W. T., Hurst, D. F., Jiménez, R., Keppel-Aleks, G., Kort, E. A., Macatangay, R., Machida, T., Matsueda, H., Moore, F., Morino, I., Park, S., Robinson, J., Roehl, C. M., Sawa, Y., Sherlock, V., Sweeney, C., Tanaka, T., and Zondlo, M. A.: Calibration of the Total Carbon Column Observing Network using aircraft profile data, Atmos. Meas. Tech., 3, 1351–1362, doi:10.5194/amt-3-1351-2010, 2010. 8532

Xiong, X., Sun, J., Xie, X., Barnes, W. L., and Salomonson, V. V.: On-orbit calibration and performance of Aqua MODIS reflective solar bands, IEEE T. Geosci. Remote, 48, 535–546, doi:10.1109/TGRS.2009.2024307, 2010. 8521

AMTD

8, 8509–8562, 2015

Uncertainty estimation in satellite remote sensing data

A. C. Povey and R. G. Grainger

Table 1. Satellite data processing levels, adapted from Chase (1986).

Level 0	Reconstructed, unprocessed instrument data at full resolution.
Level 1A	Reconstructed, unprocessed instrument data, time-referenced and annotated with ancillary information such as radiometric and geometric calibration coefficients and geolocation parameters. Data may be at full resolution or an average over some retrieval area.
Level 1B	Level 1A data that has been converted to physical units (e.g. brightness temperature rather than voltage). Not all instruments will have a Level 1B equivalent.
Level 2	Derived environmental variables (e.g. ocean wave height, soil moisture) at the same resolution and location as the Level 1 source data.
Level 3	Variables mapped onto uniform space-time grid scales, usually with some corrections for completeness and consistency (e.g. interpolation of missing points, interlacing multiple orbits).

AMTD
8, 8509–8562, 2015

A. C. Povey and R. G. Grainger

Table 2. Example of an error budget.

	Uncertainty Term	Uncertainty	Bias	Sensitivity	Random Uncertainty	Systematic Uncertainty
Measurement elements	y_1 $\vdots$	σ_{y_1}	δ_{y_1}	$\frac{\partial x_1}{\partial y_1}$	$\frac{\partial x_1}{\partial y_1}\sigma_{y_1}$	$\frac{\partial x_1}{\partial y_1}\delta_{y_1}$
Parameter elements	b_1 $\vdots$	σ_{b_1}	δ_{b_1}	$\frac{\partial x_1}{\partial b_1}$	$\frac{\partial x_1}{\partial b_1}\sigma_{b_1}$	$\frac{\partial x_1}{\partial b_1}\delta_{b_1}$
Total Uncertainty					Add above values in quadrature	Add above values

AMTD
8, 8509–8562, 2015

Table 3. Levels of system maturity, as defined in Bates and Barkstrom (2006).

Level 1	Initial Research	Results are based on environmental data records or a research satellite mission. Time series is short (usually less than 10 years). Validation is not yet complete.
Level 2	Managed Development	Initial validation complete with peer-reviewed journal paper(s) published, etc.
Level 3	Validated	Continuous validation for greater than 10 years. Data from multiple investigators with understood differences in results. Provisionally used in assessments and societal benefit areas with positive impact demonstrated.
Level 4	Certified Validated (a preponderance of the evidence)	Full provenance demonstrated; fully compliant with national and international standards; regularly used for identified societal benefit areas.
Level 5	Benchmark (beyond a reasonable doubt)	Variable critical to defining long-term climate change that is observed on the global scale. A measurement that is tied to irrefutable standards, usually with a broad laboratory base. Observation strategy designed to reveal systematic errors through independent cross-checks, open inspection, and continuous interrogation. Limited number of carefully selected observables, with highly confined objectives defining (a) climate forcings, (b) climate response.

AMTD
8, 8509–8562, 2015

A. C. Povey and R. G. Grainger

Table 4. Excerpts of the system maturity matrix defined by Work Package 2 (2013), available at http://www.coreclimax.eu/sites/coreclimax.itc.nl/files/documents/Deliverables/WP_Reports/Deliverable-D222-CORECLIMAX-Maturity_Matrix.xlsx.

Category	Maturity 1–2	Maturity 3–4	Maturity 5–6
Software readiness	Conceptual development	Portable and numerically reproducible code with draft user manual	Turnkey system fully compliant with coding standards
Metadata	None	Standardised formatting sufficient to use and understand data and trace data heritage	Regularly updated metadata, fully compliant with international standards
User documentation	Limited scientific description of the methodology available from PI	Published methodology with product descriptions and validation exercises available from PI	Publications outlining product updates and comprehensive validation (including uncertainty information)
Uncertainty characterisation	None	Quantitative estimates of uncertainty provided using standard nomenclature and procedures to establish SI traceability	Data provider has participated in multiple international assessments, incorporated feedback into the product development cycle, and quantified temporal and spatial error covariances
Public access and feedback	Restricted availability through PI	Version-controlled, documented computer codes available through PI	Source code available to public with capability for continuous data provisions
Usage	None	Product use cited in literature; societal and economic benefits discussed	Product and its applications have become the reference in multiple research fields with demonstrated influence on policy making

AMTD
8, 8509–8562, 2015

A. C. Povey and R. G. Grainger

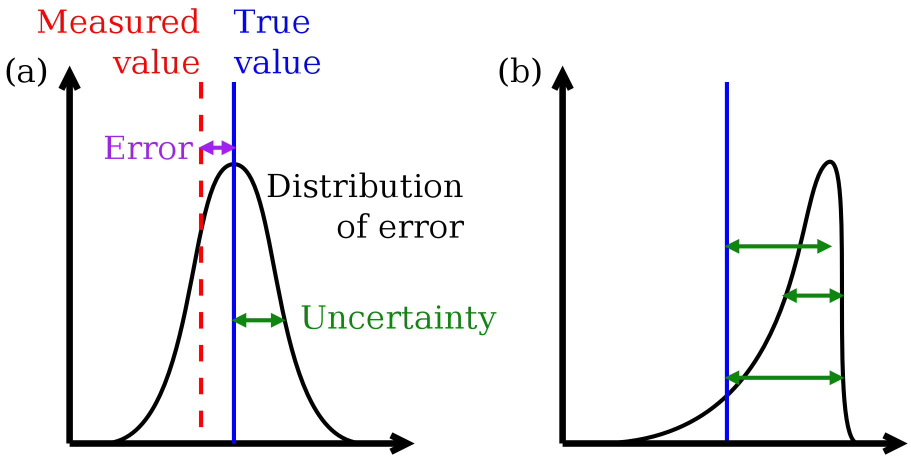

Figure 1. An illustration of error and uncertainty. The error in a measurement (purple arrow) is the difference between the true value of the measurand (solid blue) and the value measured (dashed red). The black line shows the frequency distribution of values that would be obtained if the measurement were infinitely repeated, referred to as the distribution of error. **(a)** A conventional random error. The uncertainty (green arrow) characterises the distribution of error by its width. **(b)** An error with a systematic component. This cannot be characterised with a single value.

AMTD
8, 8509–8562, 2015

Uncertainty estimation in satellite remote sensing data

A. C. Povey and R. G. Grainger

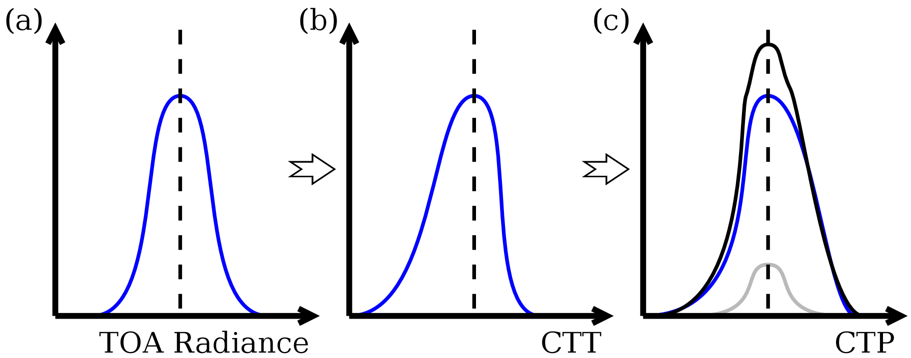

Figure 2. Distortion of the distribution of error for different selections of measurand when observing a cloud. (Non-linearities exaggerated for illustration.) **(a)** Measured TOA radiance suffers random errors, which have a symmetric distribution. **(b)** Transformation with the Planck function warps the distribution when reporting cloud top temperature. **(c)** These are further distorted when cloud top pressure is calculated. An additional error (grey; not to scale) is introduced by the auxiliary data used in that calculation, giving an irregular total distribution (black).

AMTD
8, 8509–8562, 2015

Uncertainty estimation in satellite remote sensing data

A. C. Povey and R. G. Grainger

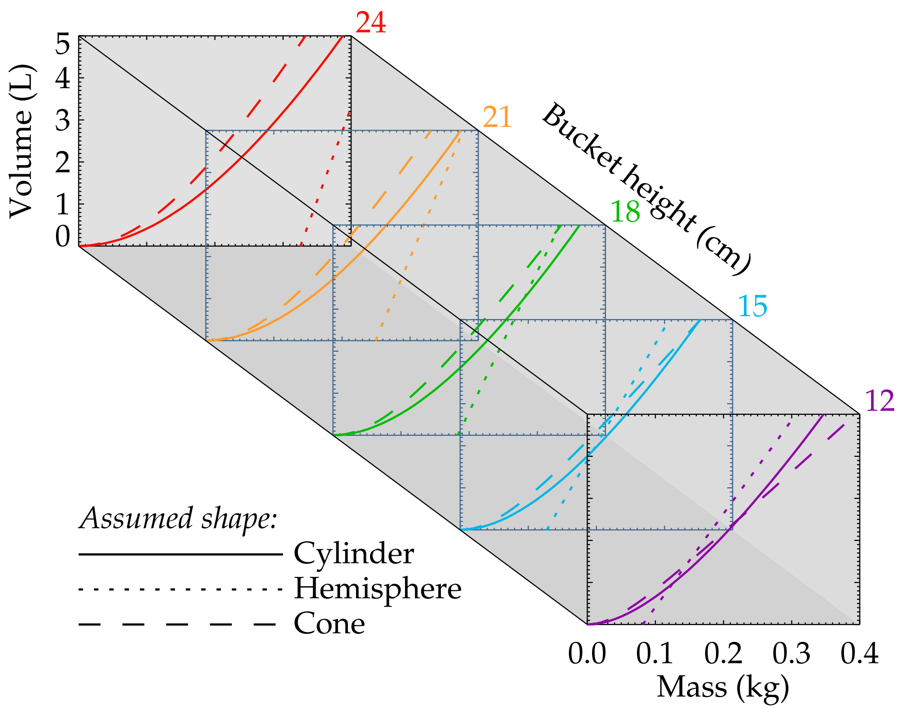

Figure 3. An ensemble of forward models for the volume of a bucket (*x* axis) as a function of its mass (*y* axis). A third parameter, the bucket's height, is not measured and so must be assumed. Its impact is shown over five slices of the *z* axis. Solid, dotted, and dashed lines denote cylindrical, hemispherical, and conical buckets respectively. The material is assumed to have thickness 1 mm and density 2.7 $\mathrm{g\,cm^{-3}}$.

AMTD
8, 8509–8562, 2015

Uncertainty estimation in satellite remote sensing data

A. C. Povey and R. G. Grainger

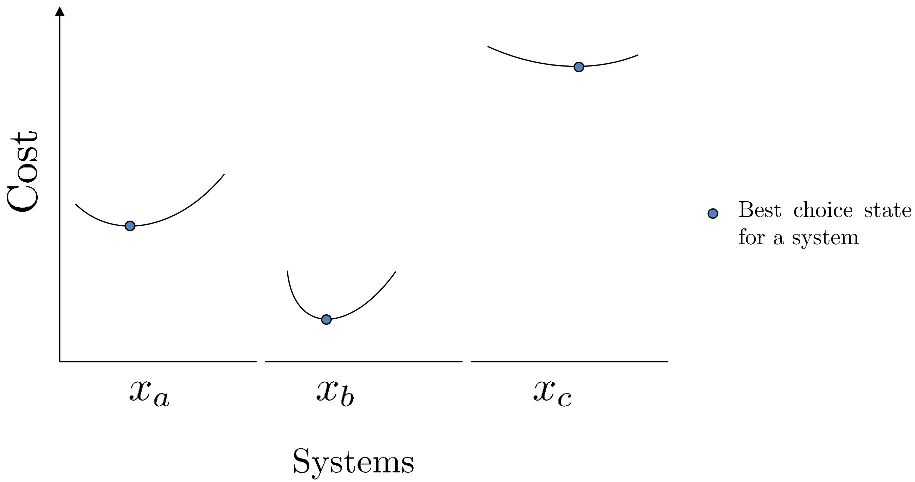

Figure 4. One-dimensional representation of a retrieval considering multiple systems (realisations of the forward model that do not necessarily retrieve the same variable). For a system, the retrieved state is the minimum of its cost function (indicated by a circle). The state with globally minimal cost (across all systems) is a posteriori taken as the best representation of the observed environment.

AMTD
8, 8509–8562, 2015

Uncertainty estimation in satellite remote sensing data

A. C. Povey and R. G. Grainger

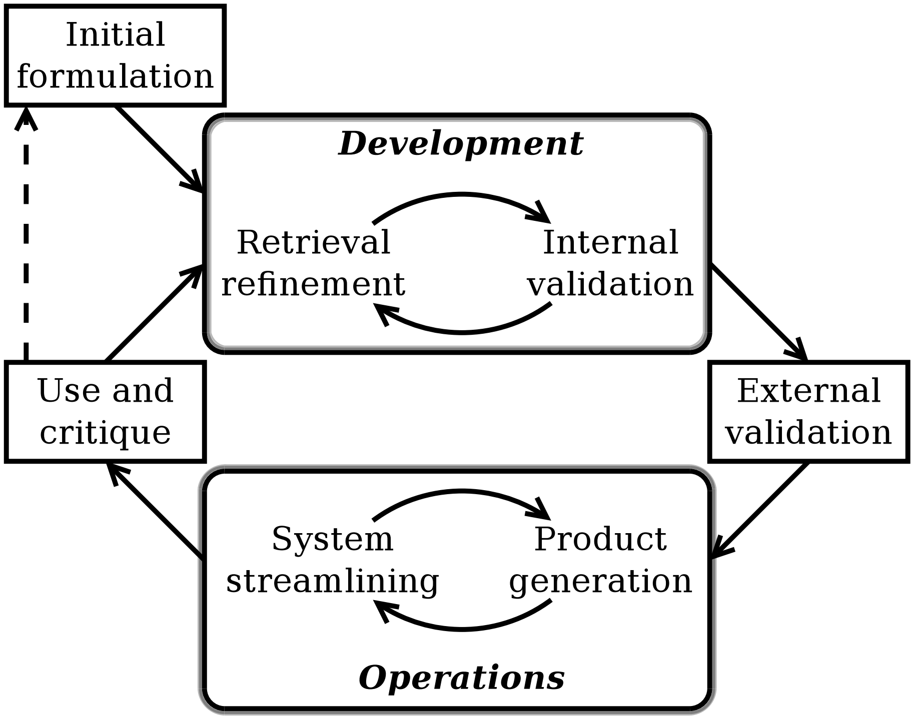

Figure 5. The cycle of retrieval development. The initial formulation and algorithm are repeatedly revised in light of internal validation activities. When consistent results are achieved, an external validation is performed (and published) to begin the operational cycle, where data are generated and disseminated. The application and critique of the data by the scientific community then feeds into further refinement of the algorithm (or entirely new algorithms). The development and operational cycles continue independent of the larger cycle but over time operations will increasingly dominate resources as the product becomes increasingly fit for purpose.

AMTD
8, 8509–8562, 2015

Uncertainty estimation in satellite remote sensing data

A. C. Povey and R. G. Grainger

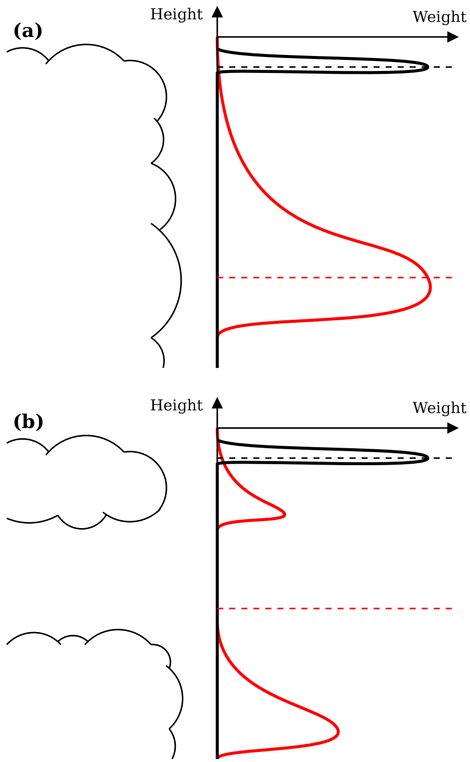

Figure 6. Schematic of the weighting functions for CTH for an infrared radiometer (red) and lidar (black), with dashed lines denoting the value retrieved. **(a)** For a thick cloud, the radiometer is most sensitive to the region one optical depth into the cloud while the lidar detects the physical cloud top. **(b)** The lidar's sensitivity is unchanged when a thin cloud lies over a thicker one, but the radiometer observes both clouds, resulting in an unphysical CTH somewhere between the two.

AMTD
8, 8509–8562, 2015

Uncertainty estimation in satellite remote sensing data

A. C. Povey and R. G. Grainger

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◀| |▶
◀ ▶
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

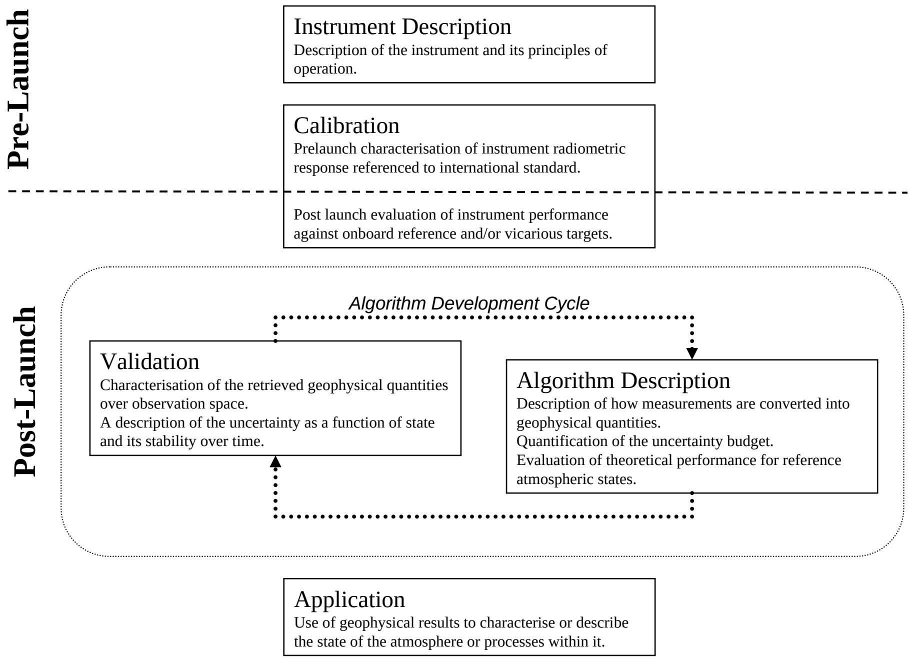

Figure 7. The sequence of scientific output needed to underpin satellite observations. The instrument, calibration, and algorithm descriptions may be contained in one or more publications. Significant iterations of the retrieval algorithm are usually described in a new publication.

AMTD
8, 8509–8562, 2015