Uncertainty information in climate data records from Earth observation

Merchant, C. J. ORCID: https://orcid.org/0000-0003-4687-9850, Paul, F., Popp, T., Ablain, M., Bontemps, S., Defourny, P., Hollmann, R., Lavergne, T., Laeng, A., de Leeuw, G., Mittaz, J., Poulsen, C., Povey, A. C., Reuter, M., Sathyendranath, S., Sandven, S., Sofeiva, V. F. and Wagner, W. (2017) Uncertainty information in climate data records from Earth observation. Earth System Science Data, 9 (2). pp. 511-527. ISSN 1866-3516 doi: 10.5194/essd-9-511-2017 Available at https://centaur.reading.ac.uk/69384/

Published version at: http://dx.doi.org/10.5194/essd-2017-16
To link to this article DOI: http://dx.doi.org/10.5194/essd-9-511-2017

Publisher: Copernicus Publications

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2017-16, 2017
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 28 February 2017

Open Access Earth System Science Data Discussions

Uncertainty information in climate data records from Earth observation

Christopher J. Merchant[1,2], Frank Paul[3], Thomas Popp[4], Michael Ablain[5], Sophie Bontemps[6], Pierre Defourny[6], Rainer Hollmann[7], Thomas Lavergne[8], Alexandra Laeng[9], Gerrit de Leeuw[10], Jonathan Mittaz[1,11], Caroline Poulsen[12], Adam C. Povey[13], Max Reuter[14], Shubha Sathyendranath[15], Stein Sandven[16], Viktoria F. Sofieva[11] and Wolfgang Wagner[17].

[1]Department of Meteorology, University of Reading, Reading, RG6 6AL, UK
[2]National Centre for Earth Observation, University of Reading, Reading, RG6 6AL, UK
[3]Department of Geography, University of Zurich, Winterthurerstr. 190, 8057 Zurich, Switzerland
[4]Deutsches Zentrum für Luft-und Raumfahrt e. V., Deutsches Fernerkundungsdatenzentrum, 82234 Oberpfaffenhofen, Germany
[5]Collecte Localisation Satellite, 11 rue Hermès, 31520 Ramonville Saint-Agne, France
[6]Earth and Life Institute, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium
[7]Deutscher Wetterdienst, Frankfurterstr. 135, 63500 Offenbach, Germany
[8]Norwegian Meteorological Institute , NO-0313 Oslo, Norway
[9]Karlsrhue Institute for Technology, Institut fur Meteorologie und Klimaforschung, 76021 Karlsruhe Germany
[10]Finnish Meteorological Institute, FIN-00101 Helsinki, Finland
[11]National Physical Laboratory, Teddington, TW11 0LW, UK
[12]Science and Technology Facilities Council, Rutherford Appleton Laboratory, Didcot, OX11 0QX, UK
[13]National Centre for Earth Observation, University of Oxford, Oxford, OX1 3PU, UK
[14]Institute of Environmental Physics, University of Bremen, Bremen, Germany
[15]Plymouth Marine Laboratory, Prospect Pl, Plymouth PL1 3DH, UK
[16]Nansen Environmental and Remote Sensing Center, Thormohlensgate 47, N_5006 Bergen, Norway
[17]Department of Geodesy and Geoinformation, Vienna University of Technology, 1040 Wien, Austria

Correspondence to: Christopher Merchant (c.j.merchant@reading.ac.uk)

Abstract. Climate data records (CDRs) derived from Earth observation (EO) should include rigorous uncertainty information, to support application of the data in policy, climate modelling and numerical weather prediction reanalysis. Uncertainty, error and quality are distinct concepts, and CDR products should follow international norms for presenting quantified uncertainty. Ideally, uncertainty should be quantified per datum in a CDR, and the uncertainty estimates should be able to discriminate more and less certain data with confidence. In this case, flags for data quality should not duplicate uncertainty information, but instead describe complementary information (such as the confidence held in the uncertainty estimate provided, or indicators of conditions violating retrieval assumptions). Errors have many sources and some are correlated across a wide range of time and space scales. Error effects that contribute negligibly to the total uncertainty in a single satellite measurement can be the dominant sources of uncertainty in a CDR on large space and long time scales that are highly relevant for some climate applications. For this reason, identifying and characterizing the relevant sources of uncertainty for CDRs is particularly challenging. Characterisation of uncertainty caused by a given error effect involves

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2017-16, 2017
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 28 February 2017

assessing the magnitude of the effect, the shape of the error distribution, and the propagation of the uncertainty to the geophysical variable in the CDR accounting for its error correlation properties. Uncertainty estimates can and should be validated as part of CDR validation, where possible. These principles are quite general, but the form of uncertainty information appropriate to different essential climate variables (ECVs) is highly variable, as confirmed by a quick review of the different approaches to uncertainty taken across different ECVs in the European Space Agency's Climate Change Initiative. User requirements for uncertainty information can conflict with each other, and again a variety of solutions and compromises are possible. The concept of an ensemble CDR as a simple means of communicating rigorous uncertainty information to users is discussed. Our review concludes by providing eight recommendations for good practice in providing and communicating uncertainty in EO-based climate data records.

1. Introduction

Few scientists would dispute the principle that an estimate of uncertainty should be given with every measured value. However, meaningful adherence to this simple principle can be challenging, and in practice researchers commonly encounter datasets where uncertainty information is generic, misleading or absent. Climate data records (CDRs) are not immune from this problem, despite the fact that climatic signals usually are subtle (e.g., Kennedy, 2014; Mahlstein et al., 2012; Flannaghan et al., 2014; Barnett et al., 2005), adding to the importance of rigorous uncertainty characterization of CDRs (e.g., Immler et al., 2010).

The question of how to derive and present uncertainty information in CDRs has received sustained attention within the European Space Agency (ESA) Climate Change Initiative (CCI; Hollman et al., 2013). Like the National Oceanic and Atmospheric Administration CDR program (Bates et al., 2016), the CCI program generates CDRs addressing a range of essential climate variables (ECVs: Global Climate Observing System, 2010; Bojinski et al., 2014). Here, we review the nature, mathematics, practicalities and communication of uncertainty information in CDRs from Earth observations, we highlight some of the challenges that developing good uncertainty information presents, and we give examples of recent progress drawn from the experience of several CCI projects.

2. The requirement for uncertainty information

The environment and climate of Earth are changing (e.g., Intergovernmental Panel on Climate Change, 2013), and these changes reflect both profound human influences on the Earth system and natural variability. Scientific progress in understanding contemporary changes has great importance in constraining future changes that may have far-reaching consequences for society. For public understanding, policy development and climate assessments, climatic changes and trends in recent decades need to be calculated. In this context, quantified observational uncertainties are required which reflect the degree to which the observing system is stable. The "system", here, includes all components that can affect the

values in the CDR, from platform and sensor to software parameters and (where relevant) human judgements. Stability is the time-rate at which systematic errors in the CDR may accumulate, and needs to be understood so that artifacts arising from the limitations of observing systems are not mis-interpreted as real climatic changes or trends.

There is major international investment of scientific effort in modeling the climate and its many component systems, and this is a major application that needs CDRs with quantified uncertainties. CDRs underpin climate model evaluation and improvement by providing references that can be used to identify model deficiencies. Model-data comparisons require appropriate skepticism about both model and data, since errors in both can mislead (e.g., Notz, 2015; Bellprat, 2016). Modelers need confidence in discriminating model-data discrepancies that unambiguously indicate model deficiencies from those where observational errors are significant. Feedback gathered by CDR producers (e.g., Rayner et al., 2015) shows that modelers find it too time consuming to develop a level of appreciation of observational datasets that allows them to make confident judgements about such matters. For this reason, CDRs need to include validated uncertainty information that modelers trust for contextualizing model-data discrepancies. Until this is achieved, modelers will continue to rely on heuristics such as interpreting differences between observational datasets as being representative of observational uncertainty, a strategy that may or may not be valid depending on the case in point.

A third example of why uncertainty in CDRs matters is the case of data assimilation. Re-analysis runs of atmospheric forecasting models (e.g. Dee et al., 2011; Kobayashi et al., 2015) provide useful, dynamically consistent information about the climate system over recent decades. The analyses include inferred fields of variables that are practically unobservable and/or were not historically observed, on a global scale. Re-analyses are among the most widely used datasets in geosciences, because of their information content and spatio-temporal completeness. Re-analyses are created by data assimilation, which brings observations and model together, using the observations to constrain the evolution of the model towards reality. The combination involves weighting the impact of different observations together and weighting the influence of observations relative to the internal evolution of the model. Ideally, uncertainty estimates should be available for each observation, so that more certain observations have more influence on the analysis. Densely sampled, numerous data (such as from satellites) can inappropriately overwhelm other observations, if these data are subject to errors that correlate across space and time and therefore do not “average out”. Ideally, spatio-temporal correlation should be understood and represented in the observational covariance matrices that weight satellite observations, to avoid undue influence on the analysis. The requirement for uncertainty information goes beyond generic estimates at dataset level: information is needed on which data are more or less certain, and how their errors are structured in space and time. Where information provided in CDRs about observational uncertainties is limited, generic assumptions are generally made, leading to sub-optimal outcomes; an example is shown in Figure 1.

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2017-16, 2017
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 28 February 2017

3. Error, uncertainty and quality

The terms 'error' and 'uncertainty' are often unhelpfully conflated. Usage should follow international standards from metrology (the science of measurement), which bring clarity to thinking about and communicating uncertainty information. Formal definitions are found in the International Vocabulary of Metrology (VIM, 2008). Adopting the "Error Approach" therein to describe the process of measurement, we have:

- *measurand* – a quantity to be measured
- *measurement* – process of experimentally obtaining one or more measured values that can reasonably be attributed to a quantity
- *measured value* – result of a measurement, obtained to quantify the measurand
- *error* – measured value minus the true value of the measurand; in practice the error is unknowable, except where the measured value can be compared with a reference value of negligible uncertainty
- *uncertainty* – non-negative parameter characterizing the dispersion (spread) of the quantity values attributed to a measurand, given the measured value and understanding of the measurement

Thus, a measured value results from measurement of a target quantity, called the measurand. It is only an estimate of the measurand, because various effects introduce errors into the process of measurement. These errors are unknown. Uncertainty information characterizes the distribution of values that it is reasonable to attribute to the measurand, given both the measured value and our characterization of effects causing error. Error is thus the 'wrongness' of the measured value (and is unknown). Uncertainty describes the 'doubt' we have about the measurand's value, given the result of a measurement and our estimate of the error distribution. A classic question at a scientific meeting is "What is the error in your measurement?", perhaps after someone has shown a plot without "error" bars. The questioner is asking for information about uncertainty, but the technically correct answer to this question would be "I don't know the error, and if I did, I would correct for it."

Note that these technical definitions correspond well to the plain meaning of the words 'error' (mistake) and 'uncertainty' (doubt) as used by non-scientists. As well as improving communication between scientists, careful usage will help scientists communicate beyond their community.

It is common for satellite datasets to include quality flags, as a simple means to guide users about the usability and validity of data. This raises the question of the relationship between quality and uncertainty.

Where a quantitative uncertainty estimate is provided for each pixel or datum, as advocated here, quality and uncertainty can be cleanly decoupled, giving different information to the user. The quality indicator should indicate whether both the

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2017-16, 2017
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 28 February 2017

measured value and its uncertainty estimate have been obtained under conditions such that they are expected to be quantitatively valid. With this approach, a highly uncertain measured value is not of lower quality provided that the high uncertainty is validly estimated. Data are flagged as lower quality data in circumstances that violate the assumptions behind the measured value and its uncertainty estimate.

For example, consider a case where the uncertainty estimates are known to be unrealistically small under certain conditions of illumination by the Sun. There may be contamination of the signal by stray radiance, for example, and no means to quantify the contamination. For these situations, a quality indicator can be used to indicate that an assumption or condition underlying the retrieval or the uncertainty estimate provided is not valid – i.e., that stray radiance may have biased the measured values by a non-negligible amount not accounted for in the uncertainty estimate.

4. Lessons from metrology

In addition to precise language for describing measurement uncertainty, metrology has developed rigorous understanding of issues around measurement uncertainty in the context of developing and promulgating international measurement standards, not least the Système International d'Unités (SI; Bureau International des Poids et Mesures, 2006). A key metrological concept is traceability through the chain of processes from the primary standard to an end-point measurement.

More generally, any measurement can be thought of as a series of transformations from the event observed to some final value. These include physical processes (such as the emission of light by a gas), measurement techniques (such as the observation of light by a detector), classifications (e.g. cloudy or clear sky) and mathematical analyses (e.g., inversion algorithms). Each transformation may be influenced by multiple effects that accumulate and propagate error. To develop a full uncertainty budget, every effect that may introduce error at any point in the chain needs to be considered, quantified (by one of various defined approaches), documented, and (if not negligible) appropriately propagated through the remainder of the chain.

Developing a more rigorous metrology of Earth observation (EO; Mittaz et al., 2017) is particularly important for CDR generation, compared to EO applications in general. The applications of CDRs involve the analysis of data on a wide range of space and time scales—from process studies that are highly resolved in time and space, to decadal and/or continental scale assessments of subtle climate changes. To provide valid quantitative uncertainty information across this range of scales, all sources of error need to be assessed and uncertainty propagation across scales needs to be rigorous. At larger scales of analysis, systematic effects that are small contributors to uncertainty in individual measured values may become the dominant sources of uncertainty (see Figure 2).

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2017-16, 2017
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 28 February 2017

Classic metrological concerns are, firstly, to assess and quantify all known sources of error and, secondly, to propagate uncertainty rigorously through all steps to the end result. The analogy between problems of EO-based climatology and metrology has prompted a developing dialogue and joint projects between these communities in recent years (e.g., World Meteorological Organisation and Bureau Internationale de Poids et Mesures, 2010; Woolliams et al., 2016).

5. Origin and characterization of errors

A datum in a CDR is the end result of a sequence of transformations. Consider a simplified scenario for the transformations involved in passive remote sensing using an infra-red radiometer to create a multi-mission CDR. (1) Infra-red radiation emitted from a particular field of view (originating from the Earth's surface and the atmosphere path above it) is collected by the aperture of a sensor and filtered during its passage through sensor optics. (2) The filtered radiance falls on a solid-state detector, causing a voltage signal. (3) The voltage is amplified electronically. (4) The amplified signal is quantized to "counts" and recorded. (5) The scene counts are compared with counts obtained when viewing two reference targets whose temperatures are measured; via this on-board calibration process, channel-integrated brightness temperature is determined using various parameters and assumptions. (6) This brightness temperature is input to processing software that retrieves a geophysical variable to generate a CDR. This sixth step can itself be decomposed into many transformations and dependencies. (6.a) Auxiliary information is also accessed by the processor, which may include a wide range of information. Some information is intrinsic to the observation and is highly certain (e.g., satellite view zenith angle, time). External geophysical datasets may be used, such as numerical weather prediction fields or surface classification, and these may or may not be provided with quantified uncertainties. All auxiliary information influences the CDR, and gives rise to uncertainty (6.b) The processor typically involves a step to determine that the pixel properties are valid for the intended retrieval (screening cloudy pixels, for example). This influences the CDR through the sampling distribution of the observations. (6.c) The set of observations is inverted to obtain an estimate of a geophysical quantity, such as an ECV. This inversion may be sensitive to the auxiliary information, and may vary in its complexity and degree of non-linearity. (6.d) Many ECV estimates may be aggregated to a coarser space-time grid for the purpose of (say) evaluating the results of a climate model run, resulting in a particular datum in a gridded dataset for each particular sensor. (6.e) A multi-mission CDR is created from datasets for several similar sensors by harmonising discrepancies between sensors (using sensor overlap periods, or other means), which modifies the datum to its final value.

Every step in the above sequence is a transformation that is subject to effects that introduce errors. Characterising these effects is the significant core of work required to develop good uncertainty information in a CDR. The errors from each

effect have certain properties which can be estimated to the degree that the effect is understood. There are several aspects to characterizing the errors from a given effect, which are discussed in turn with reference to the above scenario.

5.1 Magnitude of uncertainty

The magnitude of uncertainty characterizes the dispersion (width) of the estimated distribution of errors. Standard uncertainty is the standard deviation of the distribution, although other coverage factors can also be used. The value of the standard uncertainty can be estimated from basic principles in some cases. An example is the uncertainty introduced by quantization of the signal, which in older sensors using relatively few bits could be a significant source of noise. In other cases, the estimate of uncertainty may rely on empirical information. For example, the noise of an amplifier circuit may have been measured during pre-launch testing. Using pre-launch noise levels in an uncertainty estimate involves the assumption of stable behaviour of the amplifier during and after launch; that assumption itself can be tested for consistency with other instrument data or the noisiness apparent when observing relatively uniform targets.

In generating CDRs, we often have to deal with the multi-variate case because several channels are combined to estimate a geophysical quantity. Errors in these channels are not necessarily independent, and in this case the generalization of the standard uncertainty is the error covariance matrix, which has as many rows and columns as there are channels (or other variates). The square root of an element on the diagonal of this matrix corresponds to the standard uncertainty for a particular variate.

5.2 Shape of the error distribution

If the error distribution is zero-mean Gaussian, then the standard uncertainty fully describes the error distribution arising from the effect. Not all effects cause Gaussian-distributed errors. One example is the logarithmic distribution of radar backscatter errors associated with speckle. Another example is quantization, as illustrated by Figure 3, which shows a simulation of the distributions of brightness temperature for an Advanced Very High Resolution Radiometer (AVHRR) viewing a pixel with a true scene temperature of 230 K and of 300 K. This distribution was obtained by simulating detector noise, amplifier noise, quantization and ideal (unbiased) on-board calibration. The separated peaks are the effect of the AVHRR's 10-bit digitization of the detector and amplifier noise. Each separated spike has a nearly Gaussian distribution whose spread arises from errors in the calibration process: the calibration applied for a given observation arises from a finite sample of views of the calibration targets (an internal black body and a space view), which therefore implies some statistical uncertainty. Where quantization is negligible, which is often the case for contemporary sensors, the Gaussian distribution may realistically describe the signal noise.

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2017-16, 2017
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 28 February 2017

Open Access Earth System Science Data Discussions

5.3 Propagation of uncertainty

Uncertainty from effects associated with a particular transformation ultimately propagate to the contents of the CDR. Gaussian errors can be propagated through linear and nearly-linear transformations by standard analytic means (Joint Committee for Guides in Metrology, 2008). Let $\boldsymbol{Y} = f(\boldsymbol{X})$ represent any of the transformations between the admitting Earth-leaving radiance into the aperture of a sensor and writing a datum in a climate data record. The function f describes how the one-or-more inputs in vector $\boldsymbol{X}$ give rise to the output(s) of the transformation in vector $\boldsymbol{Y}$. The uncertainty in the output(s) is characterized by an error covariance matrix

$$\boldsymbol{U_y} = \boldsymbol{C_y U_x C_y^{\mathrm{T}}}, \quad (1)$$

where $\boldsymbol{U_x}$ is the error covariance matrix of the inputs, and $\boldsymbol{C_y}$ is the matrix of sensitivity coefficients, in which $\frac{\partial f_i}{\partial x_j}$ quantifies the influence that the i^{th} input in $\boldsymbol{X}$ has on the j^{th} output in $\boldsymbol{Y}$. If there are several effects, indexed by e, then

$$\boldsymbol{U_x} = \sum_e \boldsymbol{U_{x,e}}. \quad (2)$$

These analytic propagation equations are a first order approximation, and are strictly valid for Gaussian distributed errors that are sufficiently small that f is linear over the range of likely errors.

For non-Gaussian distributions and/or transformations that are not linear, Monte Carlo approaches are necessary to propagate uncertainty. A common non-linear transformation in generating some CDRs is threshold-based categorization of a set of observations, either because the CDR comprises a classification (such as land cover), or because the retrieval of the geophysical variable is valid only for certain classes (such as cloud-free scenes). When observations are near a threshold, errors can cause a change in classification. Simulating the retrieval process many times can characterise the propagation of uncertainty in observations into the classification results.

5.4 Correlation structure

The importance of understanding the correlation of errors is that failing to account for correlation generally leads to underestimation of uncertainty, and unfounded confidence in the interpretation of the CDR.

A common example of error correlation arises when a geophysical variable is retrieved from satellite imagery. Estimation of geophysical quantities from radiance measurements is usually an inverse problem in which there is some ambiguity and dependence on auxiliary parameters (whether explicit or hidden). Both ambiguity and parameter dependence tend to cause retrieval errors that are shared to some degree between nearby image pixels. The correlation length scale for such retrieval errors depends on the effect.

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2017-16, 2017
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 28 February 2017

For example, aerosol optical depth may be estimated across a particular scene in reflectance imagery assuming a size distribution and refractive index that systematically differs from the reality; errors are therefore expected to be correlated between pixels on the scales of variation in true aerosol properties.

More generally, retrieval errors are correlated on the space and time scales of atmospheric variability whenever retrieval ambiguity is related to atmosphere conditions (e.g., Merchant and Embury, 2014; Buchwitz et al., 2013). The errors may be decorrelated between different overpasses (because atmospheric conditions change, e.g., Reuter et al, 2014), but are strongly related for adjacent pixels from a single orbit overpass. Figure 4 illustrates this for the case of sea surface temperature retrieval (SST) simulated retrieval errors that are correlated geographically, and decorrelated in time.

Systematic effects are those causing errors with structure across a whole data set, or at least across large space and long time scales within a data set. The term "systematic error" is sometimes loosely equated to "bias", but the concept of a systematic effect is in truth more subtle since a systematic effect can produce zero-mean errors, which means there is no bias overall. Systematic effects can be defined as those that cause errors which one could in principle correct, if one had the understanding required. For example, a CDR may be derived from a series of sensors whose calibrations differ. Even if the series is adjusted to compensate for inconsistency between the calibration of different sensors, there is uncertainty in doing this; errors in the adjustment parameters affect, potentially, the entire data record from a particular sensor. These systematic errors may correspond to an overall bias, but more commonly they have some geographical and/or temporal structure. However, in principle, given better information, corrections for these errors could be devised.

6. Which types of uncertainty information are used?

The previous section introduced four considerations useful in thinking about uncertainty from a given effect: 'what is the typical magnitude of error?', 'what is the shape of the distribution of error?', 'how does this error propagate?' and 'what is the correlation structure of the error across many observations?'. These considerations apply quite generally. However, the nature of the answers depends on the particularities of the CDR being considered. There is a range of forms which uncertainty information can take. This range is illustrated in the CCI programme by the varied contents of 'Uncertainty Characterisation Reports' prepared for each CDR. (For these reports and other documentation, refer to www.esa-cci.org.)

Quantitative measures of uncertainty describe the doubt we have about the measurand, given the measured value, in numerical terms. Conceptually, the provided numbers quantify the dispersion (i.e., spread) of the estimated error probability distribution function (PDF). Options for characterization are varied, including percentiles, confidence intervals, maximum range of error, multiples of the standard deviation, covariance matrices, distribution histograms, misclassification rates, etc.

Standard uncertainty is a highly informative measure when the error distribution is close to Gaussian. For example, in the case of sea surface temperature (SST), errors are reasonably well described by a Gaussian distribution whose standard deviation can be modeled by uncertainty propagation (Merchant and Le Borgne, 2004; Embury et al. 2012). Even in this relatively simple case, there are subtleties. Sea-water freezes around –1.8°C. Even though the measurement error distribution remains Gaussian when the retrieved temperature approaches the freezing point, the distribution of credible SST errors becomes asymmetric given the additional knowledge that SST below –1.8°C is precluded.

The dispersion of errors is sometimes better described using fractional uncertainty. This approach is typically more appropriate for data such as ocean chlorophyll concentration or atmospheric aerosol optical depth (AOD). In both these cases there is a strict lower limit to valid data of zero, and both the measured values and standard uncertainty can vary in value over orders of magnitude, with larger uncertainty in absolute terms when the measured values are large. Quoting a fractional uncertainty is thus more appropriate, and is equivalent to stating a standard uncertainty on logarithm-transformed data. However, for values near zero, standard uncertainty may be more representative. For example, effects associated with surface brightness introduce an uncertainty in AOD that is the dominant uncertainty for low-aerosol scenes. Thus, GCOS (2011) recommends the combination of absolute and fractional uncertainty models for CDRs of aerosol optical depth.

Some CDRs refer to categorical ECVs, such as the status of the land cover at a given place, whether the land at a given location has recently burned, or whether the land is covered by a glacier. Here an appropriate statement of uncertainty can be probabilistic: how probable is the status to be other than indicated? When the classification uses a Bayesian approach like the maximum likelihood estimation, the probability to belong to the output class is naturally available. For non-probabilistic classifiers (‘random forest’ for instance), a proxy to class membership probability can be defined as the number of trees in the ensemble voting for the final class (Loosvelt et al., 2012). Similarly, the distance to the optimal separating hyperplane in the feature space can be used in support vector machine classifications (Giacco et al., 2010). Table 1 shows the variety of ECVs and corresponding uncertainty information in the CCI program. The maturity of uncertainty information presently provided varies, and for some cases, uncertainty estimation is not yet achieved. Given the limited uncertainty information available in the “level 1” radiance products from which the CDRs are derived, it is clear that in every case the uncertainty information could in principle be improved further. Despite this, the comments describing the basis of the uncertainty information in different products illustrate application of the principles of uncertainty estimation discussed above.

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2017-16, 2017
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 28 February 2017

7. Validation of uncertainty

Quantified uncertainty information provided in CDRs needs to be validated – i.e., evaluated by independent means to establish quantitative realism and credibility of the uncertainty estimates. Many validation studies in the literature consider the validation of measured values, but validation of attached uncertainty information is less common. Indeed, where specific uncertainty estimates are not provided with measured values, measured-value validation is often seen as a method for deriving generic uncertainty information (based on the validation discrepancies).

Validating uncertainty information in a CDR is challenging because it requires quantification of three contributions to the observed differences between the values measured from space and on the ground (e.g., Wimmer et al., 2012; Dils et al., 2014):

- the uncertainty for each CDR data value;
- the uncertainty for each reference measured value being used as a validation point; and
- the magnitude of real geophysical variability caused by the different nature of the satellite and validation measurements.

Real geophysical variability between measurements of nominally the same measurand arises for many reasons, depending on the ECV considered. The spatial location of the measurements can differ (including the tolerance for spatial mismatch and the effect of point measurement vs. area-average over a satellite pixel). The measurements are likely not perfectly synchronized, and the geophysical state may have evolved in the intervening time. Definitional differences are common between measurands, even though nominally equivalent, such as the remotely sensed measurement being sensitive to a weighted average of some vertical profile of a variable, whereas the reference measurement is made at discrete heights/depths. In some cases, validation must be performed using reference data for a measurand that is closely related, but not exactly the same (a definitional discrepancy).

In the case of satellite CDR data, x_{sat}, containing standard uncertainty estimates, u_{sat}, validation of the CDR uncertainty information can be based on the distribution of the ratio:

$$\frac{x_{sat} - x_{ref}}{\sqrt{u_{sat}^2 + u_{ref}^2 + u_{mis}^2}}, \quad (3)$$

where x_{ref} is the value of the reference (validation) data, u_{ref} is the uncertainty in the reference data, and u_{mis} is the geophysical variability arising from temporal, spatial and definitional mis-match between the satellite and reference data. If the uncertainties and variability are correctly quantified, this ratio will be normally distributed with standard deviation equal

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2017-16, 2017
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 28 February 2017

to unity. The better the quality of the reference data (the smaller u_{ref}) and the better the match of satellite to validation data (the smaller u_{mis}), the more sensitive is the validation of u_{sat}.

An example validation of uncertainty based on this principle is shown in Figure 5. In this case, the data are cloud-top height (CTH) from Cloud CCI retrievals, driven by interpretation of the cloud-top temperature in thermal imagery, matched to independent CTH measurements made by CALIPSO, using laser ranging. The CALIPSO validation data have, in this case, negligible uncertainty, and mismatch uncertainty is also neglected. The plots therefore show the histogram of discrepancy in CTH between the two observations, divided by the uncertainty estimated in the Cloud CCI retrieval process. The Gaussian that best fits the main peak is also shown, with its calculated width. In the case of ice clouds, the product uncertainty seems to be well estimated for the majority of data. For liquid clouds, the analysis reveals a systematic effect that is not accounted for in the product uncertainty estimates, since a significant fraction of data are found to disagree with the validation data by around six times the standard uncertainty. Such disagreement would be very rare if the standard uncertainty were appropriate to these matches.

Triple collocation techniques (McColl et al., 2014) have been used for assessing uncertainty estimates in near-surface wind speed (Stoffelen et al., 1998), soil moisture (Gruber et al., 2016) and other remotely sensed variables. For valid quantitative estimation or validation of uncertainty, the technique requires three sources of collocatable data that have errors that are independent and random (both between the data sources, and within each data source), and assumes that sampling mis-matches and differences of definition of the measurands between the three types of data are negligible. Other methods of uncertainty validation methods are briefly reviewed in Sofieva et al. (2014).

The uncertainty arising from instrument noise can also be validated using an Earth target that is assumed not to vary – e.g. white sands in New Mexico for reflectance validation. In this case, validation is not against independent measurements, but using repeated observations by the same instrument. Such analyses would be more robust if the geophysical standard could be traced to a more controlled reference, which would require more support for repeated, accurate measurement of the Earth target from the ground (Schaepman-Strub et al., 2016). For categorical ECVs such as land cover type, a degree of validation of uncertainty information can be obtained by verifying that estimated mis-classification rates in the product are stable with respect to reasonable ranges of classification parameters. For instance, if classification is based on training a classifier using a dataset split into calibration and validation ("train" and "test") subsets, the process can be repeated many times with a different random division into train and test subsets, which allows the dispersion in the mis-classification rates to be characterized.

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2017-16, 2017
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 28 February 2017

8. Presenting uncertainty information in climate datasets

When determining how uncertainty information is to be included in the CDR, various requirements can conflict (Table 2). The core conflict is between providing for applications requiring only summary information that discriminates more and less uncertain data, and providing for applications that demand detail about uncertainties sufficient to calculate uncertainty in quantities derived from the CDR (averages in space and time, temporal differences, integrals, trends, fluxes, etc). Data producers themselves are a user of their low-level (e.g., full resolution, orbital) products when they create higher-level products (e.g., gridded datasets and gap-filled analyses). In order to provide realistic uncertainty information at the higher level, they may require fine-grained uncertainty information for the low-level CDR, such as separate quantification of uncertainty at pixel level from effects with distinct spatio-temporal correlation properties. Such detailed information is complex for non-expert users, and an unnecessary data volume for those whose application requires, for example, only the total uncertainty.

The increase in volume of data involved in providing uncertainty information is far from being a minor point. The volume of data required for a comprehensive description of uncertainty, including the degree of error correlation, can be many times the volume of the measured values. For example, a full error covariance matrix for N measured values is $N \times N$. Data volume and processing limits are thus significant obstacles to comprehensive brute-force calculations of uncertainty. Insight and imagination are required to develop treatments of uncertainty that meet the requirements for rigor in CDR applications and are computationally tractable. Data producers can develop different versions of products that are light and heavy with respect to uncertainty information. Data delivery systems can be developed that allow users to select on download consistent uncertainty information to the degree of detail they require. There is likely no single strategy that is optimal for every ECV.

A user consultation meeting on uncertainty information in SST CDRs (Rayner et al., 2015) explored these issues with a range of users, including "power users" in applications such as data assimilation for re-analyses and centennial-scale climate modeling. An interesting conclusion from the workshop is that many users are interested in ensemble versions of EO-based CDRs, despite the multiplied data volume this implies. The purpose of the ensemble CDR is to represent the effect of all sources of error on all spatio-temporal scales. The motivation of the ensemble approach is two-fold (e.g., Morice et al., 2012). First, the user doesn't need to engage deeply with the origins and correlation structure of errors in the CDR and their implications for their application, since these are captured in the differences between ensemble members. Second, for some applications it is simpler to re-run a process several times with different ensemble members than to propagate uncertainties through the process, particularly when error structures exist across a wide range of scales. These motivations don't apply for every application, and the ensemble approach is less attractive to users facing constraints of data volume or processing power. The ensemble approach raises issues and opportunities for the data provider. Uncertain auxiliary parameters to the processing can be sampled across their plausible range rather than relying on a single best estimate. However, the strategy

for creating an ensemble requires careful design, and there are subtleties to be addressed, such as whether a "best" member is supplied, how large an ensemble is appropriate, and what the ensemble spread represents. Within the CCI program, the ensemble approach has been adopted only experimentally thus far (e.g., Reuter, 2013).

9. Good practice for uncertainty quantification

One perspective on what constitutes good practice in uncertainty quantification has been embedded in metrics of CDR maturity recently proposed. Building on the work of Bates and Privette (2012) for the NOAA Climate Data Records Program, Schulz et al. (2015) have proposed a system maturity matrix (SMM) for assessing CDR generating capacity. The SMM includes criteria for assessing the maturity of uncertainty characterisation, including linkage to standards, degree of validation, the approach to uncertainty quantification and the degree of automation of quality monitoring. The originators are clear that the purpose of assessing a CDR system against the SMM is to identify priorities for investment in developing a CDR in support of routine climate information and assessments. The overall maturity score is not an indicator of the scientific value of a dataset, which could be very high for a new variable obtained by a system whose maturity is low.

For multiple factors in CDR generation, the SMM maps the status of a CDR system onto a scale from 1 (low maturity) to 6 (high maturity). The content of the SSM relevant to uncertainty, validation and quality is reproduced as Table 3. A score of 2 on the uncertainty quantification criterion corresponds to provision of limited information, such as estimates of uncertainty that are generic (i.e., describe the typical uncertainty for the dataset as a whole). At the next maturity score, the provided information is still at the level of the dataset, but is comprehensively described and quantified, which suggests that the nature of the effects causing error is determined. To move to a score of 4, this understanding is applied to develop uncertainty information in the product that is specific to each datum, and capable of discriminating between more and less certain data. A score of 5 corresponds to providing quantification of the correlation structures in errors, via covariance information or other means. For practical purposes, since covariance matrices can be large, this provision is not necessarily required to be within the product per datum. However, feasible approaches may be found that satisfy this maturity criterion at a per-datum level, such as decomposition of total uncertainty into dominant components arising from effects with distinct, quantified correlation structures (e.g., Bulgin et al., 2016). The highest maturity score of 6 is obtained when the estimated uncertainty magnitudes and error correlation structures are thoroughly validated.

It is not the purpose of this paper to discuss the general merits of the maturity matrix approach to evaluating CDR systems. However, it is clear that if a CDR producer address uncertainty using the perspectives in this paper, they will achieve a high maturity score in this aspect of the SMM.

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2017-16, 2017
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 28 February 2017

Open Access Earth System Science Data Discussions

This paper has demonstrated the complexity of developing good uncertainty information for users of climate datasets. The aspiration to provide per-datum uncertainty estimates at all product levels and for all versions of products at all spatio-temporal scales is very challenging and not fully solved. It is clear that developing and validating uncertainty estimates involves effort comparable to developing the retrieval itself. There is a lot of diversity in the nature of CDRs and of the errors present in them. The details of good practice for describing uncertainty in CDRs vary accordingly. Nonetheless, it is useful to state some general principles that emerge from the previous sections:

1. Make quantitative uncertainty information available within the dataset. (Don't expect users to find uncertainty information from reading related papers.)
2. Use well-defined metrological concepts, such as "standard uncertainty", to quantify uncertainty.
3. Provide uncertainty information that discriminates which data are more and less certain. Per-datum uncertainties should be given, if possible, in CDRs where uncertainty varies significantly.
4. Assuming per-datum uncertainty information is provided, avoid redundancy of this information with quality flags: do not flag high-uncertainty data as "bad" if a valid estimate of that high uncertainty is provided; instead, use quality flags to indicate the level of confidence in the validity of the provided uncertainty and retrieval assumptions.
5. Define what uncertainty information is given in the CDR in the product documentation.
6. Describe in the product documentation the main effects causing errors, how uncertainty varies within the dataset, how errors may be correlated in time and space, and under what circumstances estimated uncertainty may be invalid (and flagged as such).
7. Use validation to evaluate both retrieved quantities and uncertainty estimates.
8. Propagate uncertainty appropriately (accounting for error correlation) and consistently when creating aggregated products.

10. Conclusion

Quantifying and validating uncertainty information is challenging. The challenge is particularly great when using complex observational systems (satellite sensors and their processing chains) to meet the requirements of data for climate applications (understanding of uncertainty across a wide range of space and time scales, provided with a high level of rigour and transparency). The form of uncertainty information must differ according to the nature of the target essential climate variable. In general, however, the aim is to provide justified (validated) quantification of uncertainty that allows users to know which data are more or less certain within the product.

There are many sources of error (effects) that influence the values that end up populating a climate data record. Uncertainty in fundamental climate data records (level 1 products) are not presently provided in a form sufficient to support per-datum

propagation to estimate uncertainty in derived climate data records, so there are constraints on what is practical. The efforts of CDR producers must focus on identifying dominant sources of error, bearing in mind that effects of relatively small magnitude in a single datum may be the dominant effect on a large space-time scale, and therefore may be relevant for climate applications. There is therefore unavoidably a need to develop a good understanding of many error sources, and not just "instrument noise". At the same time, one cannot wait for the perfect uncertainty budget: producers must provide CDR using the best available knowledge meanwhile. Where some error sources are as yet unquantifiable, users benefit from simple, accessible descriptions of the potential uncertainty not estimated in the product.

The means of quantifying uncertainty vary across ECVs, depending on factors such as the nature of the geophysical retrieval (ranging across physics-based inversion methods, empirical relationships and manual interpretation) and the availability of validation data. Uncertainty contributions may be modelled using a detailed uncertainty budget, or estimated from the spread of outcomes across Monte Carlo simulations. Again, pragmatism is often required to obtain a timely estimate.

The idea of validation should encompass validation of data and of uncertainty information associated with data. Validation of uncertainties (as well as the measured values themselves) can be limited by the availability of reference data.

Uncertainty concepts can be confusing, and users' needs vary. CDR producers can help by providing versions of products with "simple" (but inevitably partial and approximate) uncertainty information. Documentation must make clear what the provided information is (and is not) telling users. We have noted that ensemble methods may be able to provide users with conceptual simplicity and quantitative rigour, although at the expense of practical issues around data volumes.

Use in CDRs of well-defined, internationally agreed standards for naming and calculating uncertainty information is highly desirable wherever possible, and will clarify interaction with and feedback from user communities. These standards come from the field of metrology, and cover most situations encountered in developing CDRs. Engagement between Earth observers and metrologists is increasing. These interactions will make progress on the aspects of EO that go beyond definitions developed for laboratory-based metrology. In particular, quantifying the uncertainty over large scales of space and time (the low time and spatial frequencies in CDRs) remains a major research challenge, and involves understanding of complex error correlation structures (effects that cause neither independent random or fixed systematic errors). This area of research cannot be neglected, because users apply climate data to the full range of space-time scales spanned by Earth observation. Significant progress needs to be made in order to be able to provide users of climate data records with the certainty they need regarding uncertainty.

Acknowledgements

The primary funding for the work represented in this paper was the Climate Change Initiative of the European Space Agency, Phases 1 and 2. CJM was additionally supported by the National Centre for Earth Observation of the UK's Natural Environment Research Council. Further contributions were developed within the project "Fidelity and Uncertainty in Climate data records from Earth Observation" which received funding from the European Union's H2020 Research and Innovation programme, under Grant Agreement 638822. We thank Angela Benedetti of ECWMF for support in summarizing the MACC data assimilation experiments. MR was in part funded by the State and the University of Bremen.

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.

Barnett, T., F. Zwiers, G. Hengerl, M. Allen, T. Crowly, N. Gillett, K. Hasselmann, P. Jones, B. Santer, R. Schnur, P. Scott, K. Taylor, and S. Tett, Detecting and Attributing External Influences on the Climate System: A Review of Recent Advances. *J. Climate*, 18, 1291–1314, doi: 10.1175/JCLI3329.1., 2005.

Bates, J. J. and J. L.Privette, A maturity model for assessing the completeness of climate data records, Eos Trans. AGU, 93(44), 441, 2012.

Bates, J., J. Privette, E. Kearns, W. Glance, and X. Zhao, 2016: Sustained Production of Multidecadal Climate Records: Lessons from the NOAA Climate Data Record Program. *Bull. Amer. Meteor. Soc.*, 97, 1573–1581, doi: 10.1175/BAMS-D-15-00015.1., 2016.

Benedetti, A., Morcrette J.-J., Boucher O., Dethof A., Engelen R.J., Fisher M., Flentjes H., Huneeus N., Jones L., Kaiser J.W., Kinne S., Mangold A., Razinger M., Simmons A.J., Suttie M., and the GEMS-AER team, Aerosol analysis and forecast in the ECMWF Integrated Forecast System: Data assimilation. Technical Memoranda ECMWF 571., 2008.

Bojinski, S., M. Verstraete, T. Peterson, C. Richter, A. Simmons, and M. Zemp, The Concept of Essential Climate Variables in Support of Climate Research, Applications, and Policy. *Bull. Amer. Meteor. Soc.*, 95, 1431–1443, doi: 10.1175/BAMS-D-13-00047.1., 2014.

Buchwitz, M., Reuter M., Bovensmann H., Pillai D., Heymann J., Schneising O., Rozanov V., Krings T., Burrows J. P., Boesch H., Gerbig C., Meijer Y., and Löscher A., Carbon Monitoring Satellite (CarbonSat): assessment of atmospheric CO_2 and CH_4 retrieval errors by error parameterization. Atmospheric Measurement Techniques, 6, 3477-3500, doi:10.5194/amt-6-3477-2013, 2013.

Bulgin, C. E., Embury, O., Corlett, G. and Merchant, C. J., Independent uncertainty estimates for coefficient based sea surface temperature retrieval from the Along-Trck Scanning Radiometer instruments. Remote Sensing of Environment, 178. pp. 213-222. ISSN 0034-4257 doi: 10.1016/j.rse.2016.02.022, 2016.

Bureau International des Poids et Mesures, The International System of Units (SI), 8th edition, available from http://www.bipm.org/en/publications/si-brochure/, 2006.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N. and Vitart, F., The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q.J.R. Meteorol. Soc., 137: 553–597. doi:10.1002/qj.828, 2011.

Dils, B., Buchwitz, M., Reuter, M., Schneising, O., Boesch, H., Parker, R., Guerlet, S., Aben, I., Blumenstock, T., Burrows, J. P., Butz, A., Deutscher, N. M., Frankenberg, C., Hase, F., Hasekamp, O. P., Heymann, J., De Mazière, M., Notholt, J., Sussmann, R., Warneke, T., Griffith, D., Sherlock, V., and Wunch, D.: The Greenhouse Gas Climate Change Initiative (GHG-CCI): comparative validation of GHG-CCI SCIAMACHY/ENVISAT and TANSO-FTS/GOSAT CO2 and CH4 retrieval algorithm products with measurements from the TCCON, Atmos. Meas. Tech., 7, 1723-1744, doi:10.5194/amt-7-1723-2014, 2014.

Embury, O. and Merchant, C. J., A reprocessing for climate of sea surface temperature from the Along-Track Scanning Radiometers: a new retrieval scheme. Remote Sensing of Environment, 116. pp. 47-61. ISSN 0034-4257 doi: 10.1016/j.rse.2010.11.020, 2012.

Flannaghan, T. J., S. Fueglistaler, I. M. Held, S. Po-Chedley, B. Wyman, and M. Zhao, Tropical temperature trends in Atmospheric General Circulation Model simulations and the impact of uncertainties in observed SSTs, J. Geophys. Res. Atmos., 119, 13,327–13,337, doi:10.1002/2014JD022365, 2014.

Global Climate Observing System, Implementation Plan for the Global Observing System for Climate in Support of the UNFCCC (2010 Update), GCOS-138 WMO-TD/No. 1523. , 2010.

Giacco F., Thiel C., Pugliese L., Scarpetta S., and Marinaro M., Uncertainty analysis for the classification of multispectral satellite images using SVMs and SOMs. IEEE Trans Geosci Remote Sens 48(10):3769–3779, 2010.

Gruber, A., Su, C.H., Zwieback, S., Crowd, W., Dorigo, W., and Wagner, W., Recent advances in (soil moisture) triple collocation analysis, International Journal of Applied Earth Observation and Geoinformation, 45: 200-211, 2016.

Hollmann, R., C. J. Merchant, R. Saunders, C. Downy, M. Buchwitz, A. Cazenave, E. Chuvieco, P. Defourny, G. de Leeuw, R. Forsberg, T. Holzer-Popp, F. Paul, S. Sandven, S. Sathyendranath, M. van Roozendael, and W. Wagner, The ESA Climate Change Initiative: Satellite Data Records for Essential Climate Variables. *Bull. Amer. Meteor. Soc.*, 94, 1541–1552, doi: 10.1175/BAMS-D-11-00254.1., 2013.

Joint Committee for Guides in Metrology, Evaluation of measurement data – Guide to the expression of uncertainty in measurement, JCGM 100:2008, available from http://www.bipm.org/en/publications/guides/gum.html, 2008.

Immler, F. J., Dykema, J., Gardiner, T., Whiteman, D. N., Thorne, P. W., and Vömel, H.: Reference Quality Upper-Air Measurements: guidance for developing GRUAN data products, Atmos. Meas. Tech., 3, 1217-1231, doi:10.5194/amt-3-1217-2010, 2010.

Intergovernmental Panel on Climate Change, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovern- mental Panel on Climate Change* [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp., 2013.

Kennedy, J. J., A review of uncertainty in in situ measurements and data sets of sea surface temperature, Rev. Geophys., 52, 1–32, doi:10.1002/2013RG000434, 2014.

Loosvelt, L., Peters, J., Skriver, H., De Baets, B., Verhoest, N.E., Impact of reducing polarimetric SAR input on the uncertainty of crop classifications based on the random forests algorithm. IEEE Trans Geosci Remote Sens 50(10):4185–4200, 2012.

Mahlstein, I., G. Hegerl, and S. Solomon, Emerging local warming signals in observational data, Geophys. Res. Lett., 39, L21711, doi:10.1029/2012GL053952, 2012.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K., Takahashi, K., The JRA-55 Reanalysis: General Specifications and Basic Characteristics. *Journal of the Meteorological Society of Japan*, *93* (1), 5-48, 2015

Massonnet, F., Bellprat, O., Guemas, V., and Doblas-Reyes, F. J., Using climate models to estimate the quality of global observational data sets, Science, Vol 354, Issue 6211, pp452-455, doi: 10.1126/science.aaf6369, 2016.

McColl K.A., Vogelzang J., Konings A. G., Entekhabi D., Piles M., Stoffelen A., Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target. Geophysical Research Letters 2014; 41: 6229-6236, 2014.

Merchant, C. J. and Embury, O., Simulation and inversion of satellite thermal measurements. In: Zibordi, G., Donlon, C. J. and Parr, A. C. (eds.) Optical radiometry for ocean climate measurements. Experimental methods in the physical sciences, 47 (47). Academic Press, pp. 489-526. ISBN 9780124170117 doi: 10.1016/B978-0-12-417011-7.00015-5, 2014.

Merchant, C. J. and Le Borgne, P., Retrieval of sea surface temperature from space based on modeling of infrared radiative transfer: capabilities and limitations. Journal of Atmospheric and Oceanic Technology, 21 (11). pp. 1734-1746. ISSN 1520-0426 doi: 10.1175/JTECH1667.1, 2004.

Mittaz J., Woolliams, E. and Merchant, C. J., Applying Principles of Metrology to Historical Earth Observations from Satellites, Metrologia, in preparation, 2017.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set, J. Geophys. Res., 117, D08101, doi:10.1029/2011JD017187, 2012.

Notz, D., How well must climate models agree with observations?, Phil. Trans. R. Soc. A 2015 373 20140164, DOI: 10.1098/rsta.2014.0164, 2015.

Rayner, N. A., C. J. Merchant, and G. K. Corlett, Communicating uncertainties in sea surface temperature, Eos, 96, doi:10.1029/2015EO030289, 2015.

Reuter, M., Bösch, H., Bovensmann, H., Bril, A., Buchwitz, M., Butz, A., Burrows, J. P., O'Dell, C. W., Guerlet, S., Hasekamp, O., Heymann, J., Kikuchi, N., Oshchepkov, S., Parker, R., Pfeifer, S., Schneising, O., Yokota, T., and Yoshida, Y.: A joint effort to deliver satellite retrieved atmospheric CO2 concentrations for surface flux inversions: the ensemble median algorithm EMMA, Atmos. Chem. Phys., 13, 1771-1780, doi:10.5194/acp-13-1771-2013, 2013.

Reuter M., Buchwitz M., Hilker M., Heymann J., Schneising O., Pillai D., Bovensmann H., Burrows J. P., Bösch H., Parker R., Butz A., Hasekamp O., O'Dell C. W., Yoshida Y., Gerbig C., Nehrkorn T., Deutscher N. M., Warneke T., Notholt J., Hase F., Kivi R., Sussmann R., Machida T., Matsueda H., and Sawa Y., Satellite-inferred European carbon sink larger than expected, Atmos. Chem. Phys., doi:10.5194/acp-14-13739-2014, 14, 13739-13753, 2014.

Schulz, J., John, V., Kaiser-Weiss, A., Roebeling, R., Tan, D., and Swinnen, E., Core-Climax Climate Data Record Capacity Assessment Report. CORE-CLIMAX Technical Report, CC/EUM/REP/15/001, 253 pp., available at http://www.eumetsat.int/website/home/Data/ClimateService/index.html, 2015.

Sofieva, V. F., Tamminen, J., Kyrölä, E., Laeng, A., von Clarmann, T., Dalaudier, F., Hauchecorne, A., Bertaux, J.-L., Barrot, G., Blanot, L., Fussen, D., and Vanhellemont, F.: Validation of GOMOS ozone precision estimates in the stratosphere, Atmos. Meas. Tech., 7, 2147-2158, doi:10.5194/amt-7-2147-2014, 2014.

Stoffelen A., Toward the true near-surface wind speed: Error modeling and calibration using triple collocation. Journal of Geophysical Research-Oceans, 103: 7755-7766, 1998.

Wimmer, W., Robinson I. S., and Donlon, C. J., Long-term validation of AATSR SST data products using shipborne radiometry in the Bay of Biscay and English Channel, Remote Sensing of Environment, Volume 116, Pages 17-31, ISSN 0034-4257, http://dx.doi.org/10.1016/j.rse.2011.03.022., 2012.

Woolliams E., Mittaz J., Merchant C. J. and Dilo A., Harmonization and Recalibration: A FIDUCEO perspective, Global Space-based Inter-calibration System Quarterly, Vol. 10 (2), doi: 10.7289/V5GT5K7S, 2016.

World Meteorological Organisation and Bureau Internationale de Poids et Mesures, Measurement Challenges for Global Observation Systems for Climate Change Monitoring: Traceability, Stability and Uncertainty, WMO/TD-No. 1557, Rapport BIPM-2010/08, ISBN 13 978-92-822-2239-3, available from http://www.bipm.org/en/conference-centre/bipm-workshops/wmo-bipm_workshop/, 2010.

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2017-16, 2017
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 28 February 2017

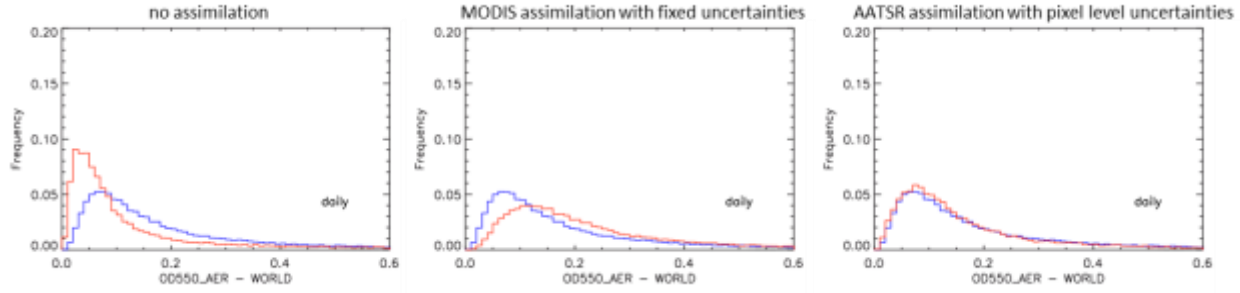

Figure 1. Benefit of pixel-level uncertainties in assimilating aerosol optical depth (AOD) estimated at 550 nm into the Monitoring Atmospheric Composition and Climate (MACC) atmospheric model. Each panel shows a distribution of AOD in the MACC model (in red) matched to 29528 AERONET ground-based AOD values (in blue): (left) no data assimilation; (centre) assimilation of MODIS retrievals; (right) assimilation of AATSR retrievals. The AERONET measured values have negligible uncertainty compared to satellite data. The MODIS data were the Dark Target AOD dataset (collection 5.1), which was operational in MACC, using fixed (generic) uncertainty estimates of 0.1 over land and of 0.05 over ocean. These values were chosen after bias correction and thorough testing of alternative uncertainty assumptions (Benedetti et al., 2008). The AATSR dataset was from Aerosol CCI, and its pixel-level uncertainty estimates were used (and no bias correction). The improved agreement in aerosol distribution suggests use of pixel-level uncertainties is beneficial.

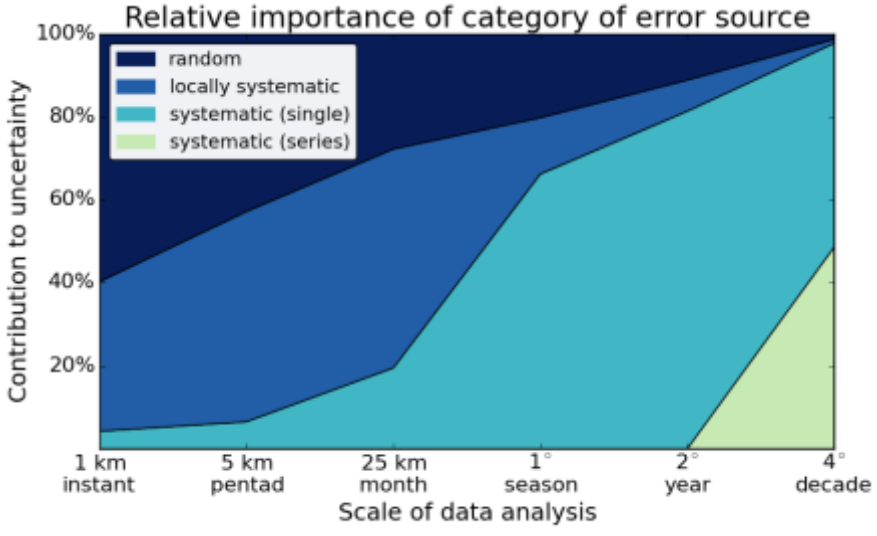

Figure 2. Contribution to the overall uncertainty from different error sources, for different spatio-temporal scales of analysis of a Climate Data Record (CDR). Conceptually, this figure is generally applicable to many climate CDRs. The particular case here is of a sea surface temperature (SST) CDR derived from a series of typical meteorological sensors. The effects causing errors are characterized by their correlation properties: noise causes random errors in SST that average out rapidly when analyzing change on larger/longer scales; retrieval errors for SST have a locally systematic aspect, and average out more gradually with scale; systematic errors, particularly in calibration, for a single sensor become more significant over time as the sensor ages and the calibration tends to drift; and a long CDR comprises data from a series of sensors which are, inevitably, imperfectly harmonized, so that systematic series effects become important for the longest time scales of analysis. Reproduced with permission from http://dx.doi.org/10.6084/m9.figshare.1483408, where full details of the scenario underlying the figure are available.

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2017-16, 2017
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 28 February 2017

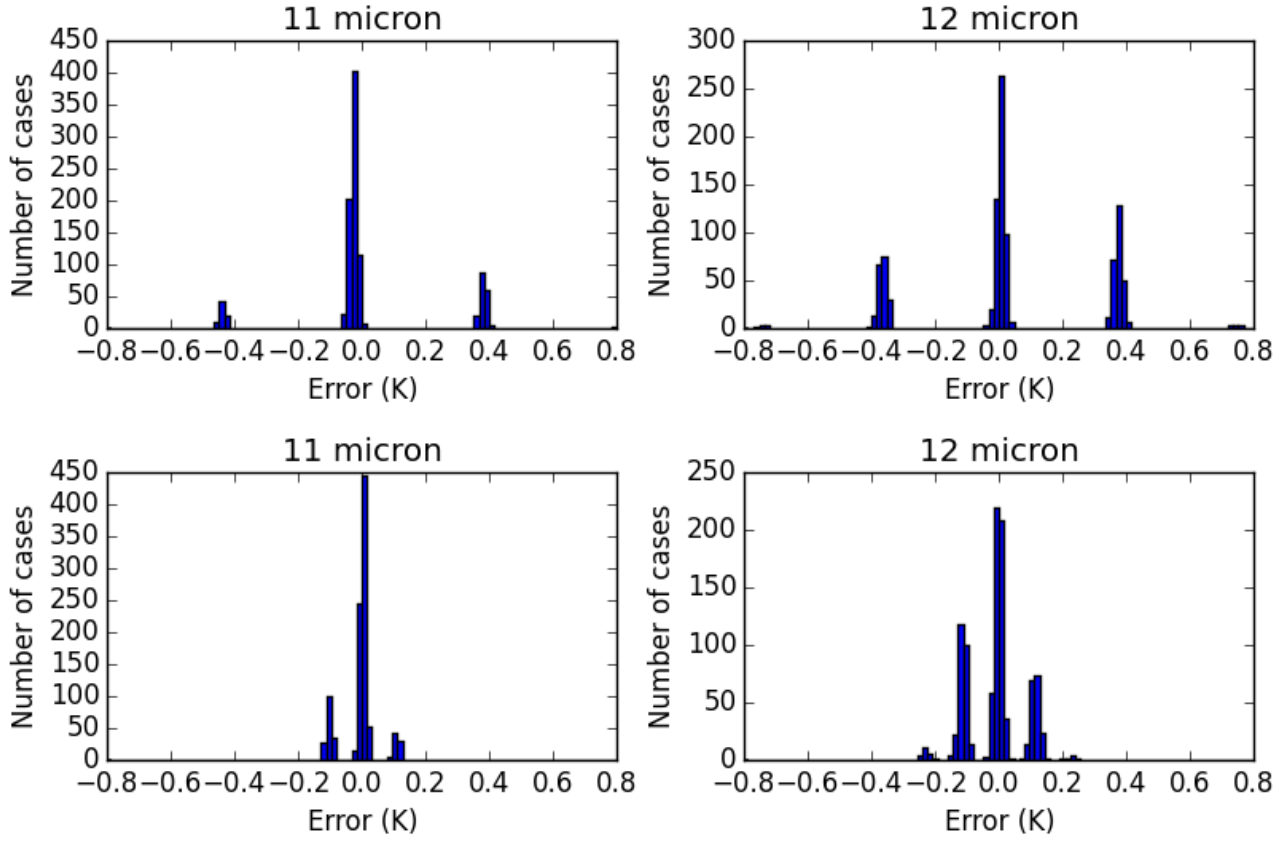

Figure 3. Distributions of single-pixel brightness temperature (BT) errors from a simulation of the detection and calibration system of an Advanced Very High Resolution Radiometer (AVHRR), for channels of different wavelength (columns) and two scene temperatures (rows: upper, 200 K scene; lower, 300 K scene). The unit of frequency of occurrence is per thousand.

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2017-16, 2017
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 28 February 2017

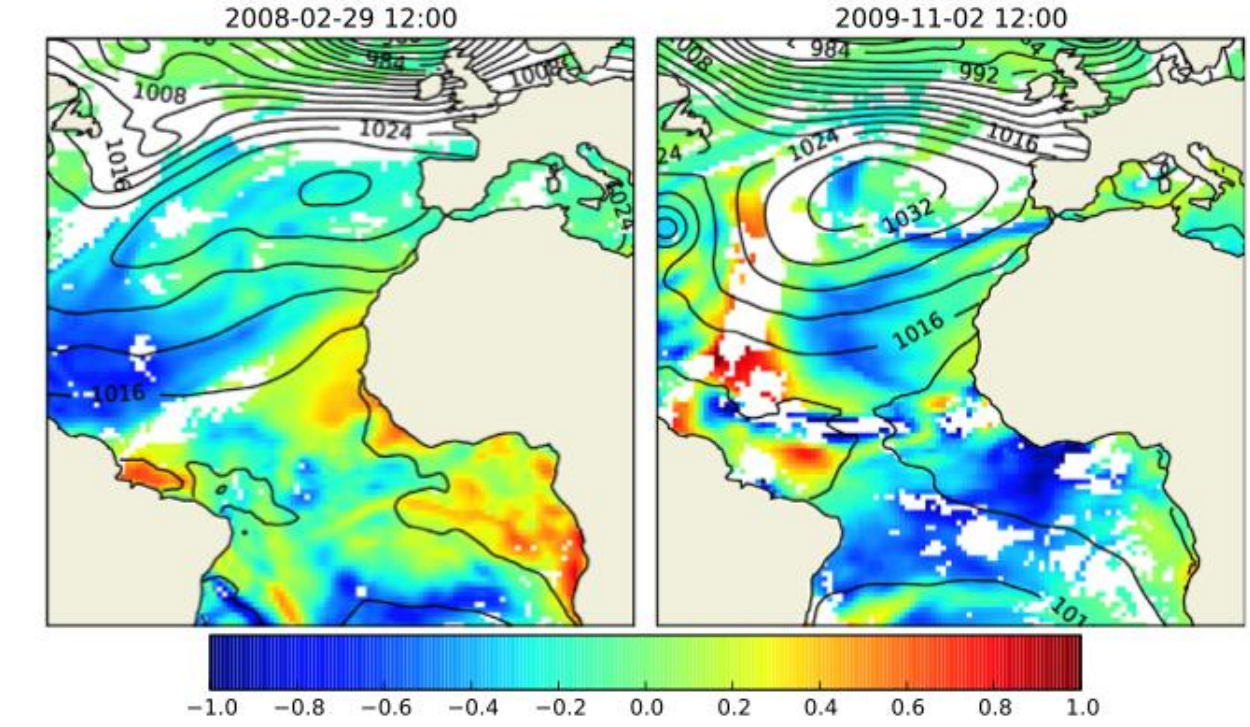

Figure 4. Simulation of locally correlated errors in retrieval of sea surface temperature (SST), overlaid with surface pressure contours to indicate length scales of atmospheric variability. The simulated retrieval errors are for a situation of a noise-free sensor whose calibration is perfectly known. The errors therefore arise solely from intrinsic ambiguity in inverting the observed radiances to SST. Note that there is no simple relationship between the SST errors and the atmospheric features associated with synoptic weather systems. White areas indicate 100% cloud cover. Reprinted from Experimental Methods in the Physical Sciences, 47, C J Merchant and O Embury, "Uncertainty information in climate data records from Earth observation", Pages 489-526, Copyright (2014), with permission from Elsevier.

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2017-16, 2017
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 28 February 2017

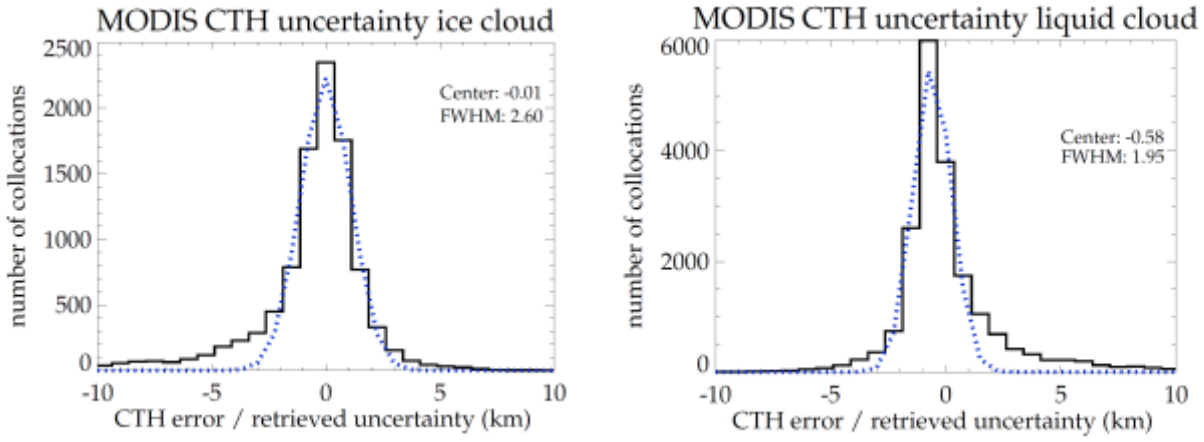

Figure 5. Example of validation of uncertainty using the distribution of differences between matched cloud top heights measured by Cloud CCI (data) and CALIPSO (CALIPSO values minus those from MODIS AQUA Cloud CCI) for a single day 2008/06/20 (solid black). Left: for ice clouds. Right: for liquid clouds. The plots show the histograms of the CTH error (the difference of retrieval compared to validation data, that is assumed to have negligible uncertainty) divided by the stated retrieval uncertainty. For ideal uncertainty estimates the full width at half maximum (FWHM) of the fitted Gaussian distribution (dashed blue) would be 2.

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2017-16, 2017
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 28 February 2017

Open Access Earth System Science Data Discussions

Table 1. Essential climate variables addressed in the ESA Climate Change Initiative program

Essential Climate Variable	Comments on nature of variable	Product characteristics	Uncertainty information provided	Basis on which uncertainty is estimated
Aerosol optical Depth (AOD)	AOD is a continuous, non-negative, log-normally distributed variable	Satellite swath (10x10 km^2 super-pixels) and gridded (1° grid daily and monthly)	Standard uncertainty given for each pixel level in swath product. Averaged uncertainty given for each cell in gridded products.	Propagation of sensor noise through retrieval process; context-specific (surface, aerosol type) estimate of retrieval uncertainty.
Cloud properties	Cloud properties are composed of several sub-variables (temperature, height, fraction) which are continuous non-negative variables	Satellite swath (≥ 5 km), gridded (0.05° grid) and averaged (0.5°, daily, monthly) estimates	Standard uncertainty given at pixel level in gridded swath product, and averaged for each cell in gridded products.	Propagation of sensor noise through retrieval process; (Optimal Estimation) context-specific (surface) estimate of retrieval uncertainty; Propagation of uncertainty to grid boxes accounting for correlation.
Glaciers	Glacier outlines derived from optical satellite data with manual intervention.	Outlines are provided in a vector format, scene-by-scene. A geo-spatial database addresses 200000 glaciers globally.	Not regularly determined; some tests have been published.	Various methods, including multiple digitizing by analysts. Appropriate validation data are generally missing.
Greenhouse Gases (XCO_2, XCH_4)	XCO_2 and XCH_4 are defined as atmospheric dry-air column-average mole fraction of CO_2 and CH_4.	One file per day including XCO_2 and XCH_4 plus additional information for surface flux inversions. Soundings have surface footprints of ~10 to ~60 km depending on sensor.	Standard uncertainty of XCO_2 and XCH_4 (per sounding) plus averaging kernels (AK) and a priori concentration profiles.	Propagation of sensor noise (and a priori uncertainty) through retrieval process and error scaling to match validation statistics.
Land Cover	A categorical variable describes the terrestrial surface annually in 22 discrete classes (from UN Land Cover Classification System)	Annual land cover maps at 300 m depicting land cover change from 1992 to 2015	Class uncertainty is available at the map and the class level. Composite surface reflectance uncertainty is	Class uncertainty is computed from confusion matrix built on independent statistical validation process.

			provided at pixel level.	
Ocean Colour	Variables of bio-optical relevance, with high dynamic range (4 decades).	Chlorophyll-a concentration, spectrally-resolved inherent optical properties, diffuse attenuation coefficient at 490 nm, membership of optical classes.	Estimated standard uncertainty and bias, for all products except back-scattering coefficient.	Uncertainty assignment based on product comparison with match-up in situ data for each optical class, applied per pixel according to class membership.
Ozone	Fields of averaged mole fraction of ozone, including vertically resolved profiles and total column" into "ozone total column and vertical profiles.	Ozone profiles from limb sounders with ~3 km vertical and ~300 km horizontal resolution. Ozone profiles from nadir sounders with ~4 km vertical resolution. Analysed/gridded versions of profiles and total column.	Uncertainty estimates are given for each ozone value in each record	Measurement noise propagated through the retrieval process and to the higher levels of data products, randomly varying parameter errors; sampling uncertainties.
Sea ice	Sea ice concentration (SIC), thickness (SIT).	SIC: daily, gridded data at between ~12 and ~50 km grid spacing. SIT: presently Arctic winter only, monthly 100 km-gridded freeboard and thickness.	SIC: Estimated standard uncertainty from retrieval and gridding. SIT: Presently no uncertainty provided.	SIC: The retrieval uncertainty is based on statistical spread of retrieval at tie points of known SIC. Parameterization for gridding uncertainty.
Sea level	Sea level is continuously variable in space and time; global variations should be consistent with the conservation of water mass in the climate system	Active remote sensing along ground tracks. Analysed to monthly 0.25° grid.	Uncertainty for each sea level determination along ground tracks. Uncertainty in inter-annual global mean sea level (Ablain et al. 2015).	Uncertainty is inferred by generalized least squares, where the error covariance matrix is built from altimeter correction uncertainties.
Sea surface temperature	Temperature is continuously variable in space and time, with a	Satellite swath (≥1 km), gridded (0.05° grid) and	Standard uncertainty given at pixel level in	Propagation of sensor noise through retrieval process;

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2017-16, 2017
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 28 February 2017

Open Access Earth System Science Data Discussions

(SST)	lower bound at the freezing temperature of sea water	gap-filled (0.05°, daily) SST estimates	swath product, and for each cell in gridded/gap-filled products. Component uncertainty contributions also available.	context-specific estimate of retrieval uncertainty; sampling uncertainty estimate in cell means.
Soil Moisture	Microwave retrievals represent moisture content in a thin surface layer (1-5 cm). No data when soil is frozen, snow-covered or overlain by very dense vegetation.	Daily (0:00 UTC) gridded (0.25°) global data. Three data records: (i) merged active, (ii) merged passive, (iii) merged active passive microwave data	Uncertainty given for each soil moisture value in each of the three data records. Additionally, quality flags are provided.	Propagation of sensor noise through retrieval process including context-specific estimate of retrieval uncertainty. Uncertainties introduced by sampling not yet characterized.

Table 2. Generic requirements for uncertainty information in climate data records, illustrating potential contradictions between the requirements for different applications of the data.

Requirement	Implications	Conflicts & solutions
1. Minimize data volume for users to download.	Provide only key summary information on uncertainty, such as the total uncertainty for each measured value.	Conflicts with need for detailed uncertainty information for some purposes (cf. 3, 4 and 5). More complete uncertainty information can be made available separately to core data products.
2. Data should be easy to read and understand.	Use standard metrological vocabulary to express uncertainty. Uncertainty data should be easy to associate to measured values.	Some established community standards/conventions include uncertainty vocabulary that is inconsistent with best practice. Work with community standards to converge practices.
3. Provide sufficient uncertainty information to allow correct propagation of uncertainty to spatial and temporal averages of data.	Uncertainty components from errors with different spatial correlations need to be separately quantified, with correlation information (e.g., length scales, covariance matrix).	Increases data volume (cf. 1). Increases complexity of dataset (cf. 2). Could provide two versions of data, one with summary and the other with comprehensive uncertainty information, with guidance as to which is needed for different purposes.
4. Provide information about temporal stability of observations and/or evolution of trend uncertainty over time (up to decades).	Information is provided on temporal correlation of errors, particularly arising from long-term systematic effects.	Full spatio-temporal covariance matrix for CDR is challenging to calculate or parameterize, and is likely infeasible to distribute. More general estimates of overall stability can be made. Ensemble approaches have been proposed.

5

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2017-16, 2017
Manuscript under review for journal Earth Syst. Sci. Data
Discussion started: 28 February 2017

Open Access Earth System Science Data Discussions

Table 3. Criteria for scoring the maturity of aspects of a CDR generation system, including criteria for uncertainty characterization, taken from the system maturity matrix of Schulz et al. (in review).

	CLIMATE DATA RECORD (CDR) MATURITY EVALUATION GUIDELINES		
Maturity	**Standards** [1]	**Validation**	**Uncertainty quantification**
1	None	None	None
2	Standard uncertainty nomenclature is identified or defined	Validation using external reference data done for limited locations and times	Limited information on uncertainty arising from systematic and random effects in the measurement
3	Score 2 + Standard uncertainty nomenclature is applied	Validation using external reference data done for global and temporal representative locations and times	Comprehensive information on uncertainty arising from systematic and random effects in the measurement
4	Score 3 + Procedures to establish SI traceability are defined	Score 3 + (Inter)comparison against corresponding CDRs (other methods, models, etc)	Score 3 + quantitative estimates of uncertainty provided within the product characterising more or less uncertain data points
5	Score 4 + SI traceability partly established	Score 4 + data provider participated in one inter-national data assessment	Score 4 + temporal and spatial error covariance quantified
6	Score 5 + SI traceability established	Score 4 + data provider participated in multiple inter-national data assessment and incorporating feedbacks into the product development cycle	Score 5 + comprehensive validation of the quantitative uncertainty estimates and error covariance