METROLOGICAL APPROACH FOR EO

GENERAL GUIDANCE ON A METROLOGICAL APPROACH TO FUNDAMENTAL DATA RECORDS (FDR), THEMATIC DATA PRODUCTS (TDPS) AND FIDUCIAL REFERENCE MEASUREMENTS (FRMS) – METROLOGY THEORETICAL BASIS

Jacob Fahy
Nigel Fox
Tom Gardiner
Paul Green
Sam Hunt
Jonathan Mittaz
Bernardo Mota
Pieter De Vis
Emma Woolliams

MAY 2022

General guidance on a metrological approach to fundamental data records (FDR), thematic data products (TDP) and fiducial reference measurements (FRM) – Metrology Theoretical Basis

Jacob Fahy
Nigel Fox
Tom Gardiner
Paul Green
Sam Hunt
Jonathan Mittaz
Bernardo Mota
Pieter De Vis
Emma Woolliams

National Physical Laboratory

Hampton Road, Teddington, Middlesex, TW11 0LW

Version control

Issue	Date	Authors	Reviewed by	Notes
1.0	19.5.22	As above	Authors	First issue

Project Acknowledgement

This work was carried out in the frame of the Instrument Data quality Evaluation and Assessment Service - Quality Assurance for Earth Observation (IDEAS-QA4EO) contract funded by ESA-ESRIN (n. 4000128960/19/I-NS), and builds on the work of previous projects, see acknowledgments.

Table of contents

1 Introduction

1.1 This set of documents

This document is part of a set of documents that describe and give practical tools to support a metrological approach for a satellite fundamental data record (FDR), satellite-derived thematic data product (TDP), or fiducial reference measurement (FRM) network or campaign.

Document	Description
Executive Summary	Introduction and overview of what a metrological approach is and what is needed to implement a metrological approach to FDRs, FRMs and TDPs
Metrology Document	This document, describing the metrological principles behind the approach
Process Document	A document that describes the step-by-step processes needed to implement a metrological approach
Templates Document	Templates and examples of uncertainty tree diagrams and effects tables and how to structure an uncertainty report
Toolkit introduction	An introduction to the COMET Python tools

1.2 Scope of this document

This document describes the theory behind a metrological approach and links to the Guide to the Expression of Uncertainty in Measurement (the GUM) and the International Vocabulary of Metrology (VIM), both established by the Joint Committee for Guides in Metrology (JCGM) – a committee that involves most of the world’s main standardisation bodies (BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML). It helps readers understand these principles and provides practical guidance on notation and how to present uncertainty information.

1.3 General introduction common to all documents

Earth Observation (EO) satellite programmes are operated by a wide variety of space agencies, meteorological agencies and commercial operators and provide observations for a wide range of social, scientific, environmental, and commercial applications. Historical and current EO data provide information about environmental and climate change that is of great value to today’s scientists and to decision makers. These data are also a legacy of immense value to future generations. However, for this immediate and legacy value to be realised, EO data sets must be interoperable and temporally stable, so that data from different sensors can be combined. The quality and uncertainty associated with datasets is also needed to assess their fitness for purpose for the desired applications.

Metrology is the discipline responsible for maintaining the International System of Units (SI) and the associated system of measurement. It is core to the SI, that measurements are stable over very long time periods, that measurement standards are equivalent worldwide and that measurements are coherent – that is different types of measurement can be combined because, for example, an electrical watt is equivalent to an optical watt is equivalent to a mechanical watt.

These properties of metrology are desired for EO data records. It is for this reason that over the last two decades there has been considerable research in the collaborative field of EO Metrology. The 2010 endorsement of the Quality Assurance Framework for Earth Observation (QA4EO) by the Committee on Earth Observation Satellites (CEOS) in the frame of the Global Earth Observation System of Systems (GEOSS) set up the basic principle that EO data should be accompanied by a fully traceable indicator of its quality, allowing users to readily assess the fitness for purpose for their applications. Traceability requires that this quality indicator be based on “a documented and quantifiable assessment of evidence demonstrating the level of traceability to internationally-agreed (where possible SI) reference standards.” QA4EO stops short of requiring robust metrological traceability, but the accompanying guidelines are based on principles adapted from the metrology community.

Since 2010, collaborative EO-metrology projects have been developing robust methods to facilitate broader use of metrological principles in EO applications. In Europe, such projects have been led through European research funding (FP-7, Horizon 2020) and by projects from the European Space Agency, and more recently the broader Copernicus Programme via institutes such as EUMETSAT and ECMWF. In this document we build on this legacy of activity and expand the concepts, nascent in the FIDUCEO project [Mittaz et al 2019], generalising them beyond passive radiometric band sensors, to establish FDRs, TDPs and FRMs.

1.4 Common introduction to FDRs, TDPs and FRMs

The terms Fiducial Reference Measurement (FRM), Fundamental Data Record (FDR) and Thematic Data Product (TDP) were applied initially by the European Space Agency to describe metrologically-rigorous observations of specific relevance to space-based observations. While not yet formally endorsed by a committee, these terms are increasingly being used by the broader Earth observation community. Here we present possible definitions for consideration[1].

> A **fundamental data record** (FDR) is a record, of sufficient duration for its application, of uncertainty-quantified sensor observations calibrated to physical units and located in time and space, together with all ancillary and lower-level instrument data used to calibrate and locate the observations and to estimate uncertainty.

Generally, FDRs will be geolocated level 1 products. The FDR is a record of the physical quantity measured by the sensor. Although some applications in reanalyses ingest level 1 products, for many applications FDRs will be used to generate TDPs.

> A **thematic data product** (TDP) is a record, of sufficient duration for its application, of uncertainty-quantified retrieved values of a geophysical variable, along with all ancillary data used in retrieval and uncertainty estimation.

TDPs are higher level products that have been processed from FDRs, through algorithms which also often combine information from another FDR (e.g., from other satellite sensors) or from external information (such as reanalysis models).

> **Fiducial reference measurements** (FRMs) are a suite of independent, fully characterised, and traceable sub-orbital measurements that follow the guidelines outlined by the GEO/CEOS Quality Assurance framework for Earth Observation (QA4EO) and have value for space-based observations.

Thus, FRMs are the quality-assured in situ observations that can be used to calibrate and validate satellite-based sensor measurements. As ESA states ‘these FRM provide the maximum return on investment for a satellite mission by delivering, to users, the required confidence in data products, in

[1] The FDR and TDP definitions are taken from the FIDUCEO project FCDR and CDR definitions, which were discussed and refined at a workshop. The FRM definition comes from the ESA website

the form of independent validation results and satellite measurement uncertainty estimation, over the entire end-to-end duration of a satellite mission.'

Note that the terms 'Fundamental Climate Data Record' (FCDR) and 'Climate Data Record' (CDR) are used for FDRs and TDPs respectively, that are also typically of multi-decadal duration and come from a series of sensors that have been harmonised to a common reference.

2 Introduction to Metrology

2.1 Core principles of metrology

FDRs, TDPs and FRMs require a metrological approach. Metrology has three basic principles:

> **Traceability**: Metrological traceability is a property of a measurement that relates the measured value to a stated metrological reference through an unbroken chain of calibrations or comparisons. It requires, for each step in the traceability chain, that uncertainties are evaluated and that methodologies are documented.
>
> > For FDRs, traceability includes both the formal calibration of the instrument to appropriate references through pre-flight, in-orbit and vicarious calibrations, including the metrological traceability of any references (e.g., in situ observations). It also includes an analysis of any auxiliary information brought into the processing of level 1 data products, including geo-correction and ortho-rectification and the correction for instrument sensitivities to environmental conditions. For TDPs, understanding traceability also includes documenting the processing steps, and the origin of any auxiliary information brought into the processing, along with any atmospheric, land or topographic correction. For FRMs traceability includes the pre-deployment calibration of the onsite instruments, as well as the characterisation and modelling of sensitivities to the environmental conditions.
>
> **Uncertainty Analysis**: Uncertainty analysis is the review of all sources of uncertainty and the propagation of that uncertainty through the traceability chain. It is based on the Guide to the Expression of Uncertainty in Measurement (the GUM).
>
> > For FDRs, uncertainty analysis needs to consider not only pre-flight calibration, but also uncertainties associated with any instrument changes in orbit and sensitivities to onboard conditions. It has to consider the extent to which the measurement model (e.g., equation) used for FDR processing (generating the FDR from level 0 data) is an approximation of reality, as well as the uncertainties associated with the different quantities within that model. For TDPs, it also needs to consider uncertainties associated with auxiliary information brought into processing, the extent to which the retrieval algorithm retrieves the measurand and uncertainties associated with any approximations in the processing steps. For FRMs, uncertainty analysis needs to consider the pre-deployment calibration, any ageing or other changes of the instrument and sensitivities to environmental conditions.
>
> **Comparison:** Metrologists validate uncertainty analysis and confirm traceability through comparisons. The Mutual Recognition Arrangement (a formal arrangement to confirm the degree-of-equivalence between SI realisations and disseminations in different countries) requires regular, formal international comparisons that are conducted under strict rules.

Earth Observation comparisons are carried out between sensors and between sensors and sub-orbital observations using simultaneous observations, transfer standards (e.g., ground observations or pseudo-invariant sites/scenes), or large-scale averages. Traditionally, Earth Observation comparisons have been performed to estimate inter-sensor differences and to parametrise some retrieval algorithms; they have only recently been used to validate uncertainty estimates independently determined. FRMs are, of course, established to support the comparison to satellite sensors. Additionally, they benefit from comparisons of the instruments and methods used in the FRM itself.

2.2 Uncertainty analysis and the GUM

The *Guide to the Expression of Uncertainty in Measurement*, known as 'the GUM', provides guidance on how to determine, combine and express uncertainty. It was developed by the JCGM (Joint Committee for Guides in Metrology), a joint committee of all the relevant standards organisations (e.g., ISO) and the BIPM (*Bureau International des Poids et Mesures*). The JCGM continues to develop the GUM and has recently produced several supplements which cover topics such as Monte Carlo Methods for uncertainty analysis and an extension of the Law of Propagation of Uncertainties to multiple output quantities.

2.3 The difference between error and uncertainty

The GUM is based on 'uncertainty analysis' which replaced an earlier concept of 'error analysis'. These two approaches differ in philosophy in a couple of important ways. Error analysis considered errors with respect to the unknown 'true' value and separately considered 'random' and 'systematic' effects. Uncertainty analysis considers uncertainty as a distribution around the known measured value of quantity values that could be attributed to the measurand. It also considers 'random' and 'systematic' effects within one single framework. Even though the GUM was internationally accepted nearly 30 years ago based on two decades of work by all the world's standardisation bodies, many scientific papers relating to different environmental observations still confuse the terms 'error' and 'uncertainty'. Furthermore, some metrologists overcompensate for this confusion and avoid the word 'error' altogether. However, understanding covariance in environmental observations fully, requires a robust understanding of the two concepts.

Uncertainty is a

> 'non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand based on the information used.'

While error is the

> 'measured quantity value minus a reference quantity value.'

The VIM adds a note to say that

> 'The concept of 'measurement error' can be used both a) when there is a single reference quantity value to refer to, which occurs if a calibration is made by means of a measurement standard with a measured quantity value having a negligible measurement uncertainty or if a conventional quantity value is given, in which case the measurement error is known, and b) if a measurand is supposed to be represented by a unique true quantity value or a set of true

> quantity values of negligible range, in which case the measurement error is not known.'

This document only considers the (b) case, where the measurement error cannot be known. Where it is known, a correction should be applied before uncertainty analysis. It is important to realise that although the sign and magnitude of the measurement error cannot be known, there are properties of the error that can be known. First, it is possible to know some information about the probability distribution function from which the error has been drawn. The standard uncertainty provides information about the standard deviation of that distribution[2] and it may additionally be possible to describe the shape of the distribution. Furthermore, it may be possible to describe how the error in one measured value relates to the error in another measured value (at a different location, time or in a different spectral band) through the error correlation.

2.4 The measurement model

The measurand — the quantity intended to be measured (see JCGM 200) — is in most cases indirectly determined from other quantities to which it is related by a measurement model. The measurement model is a mathematical expression(s) or an algorithm. It comprises all quantities known to be involved in a measurement. The measurement model enables an estimate of the measurand to be provided and an associated standard uncertainty evaluated.

For many measurements, the measurement model involves a real functional relationship f between N real--valued input quantities $X_1, \dots, X_N$ and a single real-valued output quantity (or measurand) Y:

$$Y = f(X_1, \dots, X_N).$$

This simple form does not apply for all measurements. Sometimes, particularly for the complex processing involved in generating a TDP, the measurement model cannot be written analytically. Instead, it can only be solved through iterative processes, or through machine learning methods that are only described through a coded algorithm. Often, especially in Earth observation, the measurement model can be multivariate, with more than one measurand (e.g. measured values at different locations, times or for different observation frequencies/wavelengths) determined simultaneously.

The measurement model should include all quantities that affect the measurement result. These include quantities that are measured as input quantities, and also quantities that represent corrections – e.g. a term to describe an instrument's temperature sensitivity. In general, it is also valuable to write the measurement model with an additional Δ term:

$$Y = f(X_1, \dots ,X_N) + \Delta.$$

The Δ term represents the extent to which the function f approximates reality. For example, the function f may be a linear function of a signal times a gain plus an offset, while the instrument could have some underlying non-linearity. Or the function f involves a trapezium rule summation used to allow a numerical calculation of an underlying integral. In these cases, our best estimate of the quantity Δ is the value $\delta = 0$. In [Mittaz et al 2019], the equation for measured values was written with a "plus zero" term:

$$y = f(x_1, \dots, x_N) + 0.$$

[2] Remember that the error is considered relative to the true value, while the uncertainty is a distribution around the measured value, so these are not entirely identical.

2.5 The Law of Propagation of Uncertainties and Monte Carlo Analysis

The purpose of uncertainty analysis is to propagate uncertainties associated with the input quantities, including the "plus zero" quantity $\delta = 0$, through the measurement model to give an uncertainty associated with the measured output quantity y.

The GUM gives two methods to propagate uncertainties through a measurement model – using the Law of Propagation of Uncertainties (LPU) or using Monte Carlo Analysis (MCA). The LPU uses a locally-linear approximation to combine such uncertainties, using the expression:

$$u^2(y) = \sum_{i=1}^{N} c_i^2 u^2(x_i) + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} c_i c_j u(x_i, x_j). \tag{1}$$

where,

$u(y)$ is the combined uncertainty associated with the measured value y.

$u(x_i)$ is the uncertainty associated with the input quantity x_i.

$u(x_i, x_j)$ is the covariance between x_i and x_j (see also section 2.7)

$c_i = \frac{\partial y}{\partial x_i}$ is the sensitivity coefficient, the 'translation' from an uncertainty associated with the input quantity to an uncertainty associated with the measured quantity.

N is the number of input quantities

This equation is written in these two parts as the first term $\sum_{i=1}^{N} c_i^2 u^2(x_i)$ represents the uncertainty if the x_i are fully independent (no error correlation) and the second term represents the effect of error correlation. The first term is the well-known "uncertainties add in quadrature" rule.

The LPU (for a single output quantity[3]) can also be written in matrix form as:

$$u^2(y) = \boldsymbol{c}\boldsymbol{S}(\boldsymbol{x})\boldsymbol{c}^{\mathrm{T}} \tag{2}$$

where $\boldsymbol{c} = [\partial y/\partial x_1, \dots, \partial y/\partial x_N]^{\mathrm{T}}$ is a column vector of sensitivity coefficients and $\boldsymbol{S}(\boldsymbol{x})$ is the error covariance matrix for the input quantities (discussed further in Section 2.9), with diagonals $u^2(x_i)$ and off-diagonals $u(x_i, x_j)$. It can be helpful to write

$$\boldsymbol{S}(\boldsymbol{x}) = \boldsymbol{URU} \tag{3}$$

where $\boldsymbol{U}$ is a square diagonal matrix with the standard uncertainty associated with each input quantity, $u(x_i)$ along the diagonal and zeroes elsewhere, and $\boldsymbol{R}$ is a square correlation matrix giving the correlation coefficient (between -1 and +1) for each pair of input quantities.

Monte Carlo Analysis approximates the input probability distributions by finite sets of random draws from those distributions and propagate the sets of input values through the measurement function to obtain a set of output values regarded as random draws from the probability distribution of the measurand. The output values are then analysed statistically, for example to obtain standard deviations and error covariances. The measurement function in this case need not be linear nor written algebraically. Steps such as inverse retrievals and iterative processes can be addressed in this way. The input probability distributions can be as complex as needed.

[3] Extensions to multiple output quantities are given in Section 2.6

Monte Carlo methods can provide information about the shape of the output probability distribution for the measurand, deal better with highly non-linear measurement functions and with more complex probability distributions, and can be the only option for models that cannot be written algebraically. However, they are computationally more expensive, which is an important consideration with the very high data volumes of EO. Often uncertainty analyses will use a combination of Monte Carlo methods and the LPU, for example by using Monte Carlo methods to determine the uncertainty for a particular quantity, which is then used as an input to LPU in a subsequent uncertainty analysis.

2.6 Multivariant measurands: EO always gives a dataset rather than an individual measurement

The concepts of correlation and error correlation are extremely important in analysing FDRs, TDPs and FRMs. FDRs, TDPs and FRMs do not provide one measured value. Instead, they provide observations at different locations and/or times. For radiometric measurements, such observations may also be measured simultaneously in different spectral bands and/or at different viewing angles. For active radar sensors there may be measurements in multiple frequency bands as well as at different locations and times.

In higher-level processing (to TDPs and beyond), multiple observations may be combined – for example, by combining data in different spectral bands during a retrieval or obtaining the ionosphere correction for an altimeter by combining data from two frequency bands. Measurements in different locations and times are also commonly combined, by gridding data spatially and temporally, or compared (e.g., in time series).

For that higher-level processing, measured values from different spectral/frequency bands and/or at different positions, times, and/or from different viewing angles, will be different input quantities in calculating the measurand at that level. Therefore, for equation (2), the higher-level processing will need a covariance matrix between these measured values. Because in Earth observation higher-level processing is often performed by a different team of scientists to those who provide the lower-level product, it is important for lower-level processing to provide error covariance information for the set of observations which later processing can use.

Thus, instead of a single output quantity, y, we have a vector of output quantities $(\boldsymbol{y})$ and the univariate equation (2), needs to be written as a multivariate equation. Including the expansion (3), this gives the multivariate form of the LPU as:

$$\boldsymbol{S}(\boldsymbol{y}) = \boldsymbol{CS}(\boldsymbol{x})\boldsymbol{C} = \boldsymbol{CURUC} \qquad (4)$$

where $\boldsymbol{S}(\boldsymbol{y})$ is the covariance matrix for the different output measured values, $\boldsymbol{C}$ is a square matrix of sensitivity coefficients for each component of $\boldsymbol{y}$ for each input quantity x_i.

In practice, this formulation may be difficult to handle, particularly when there are very large numbers of observations in a satellite level 1 record. It may not be possible to treat all observations by a radiometric sensor, at all times, for all spatial pixels, in all spectral bands as a single set of output quantities. Instead, methods are needed to parametrise the covariance matrix. These methods are discussed in Section 2.9, below.

2.7 Error correlation and natural variability

It is possible to calculate correlation between individual observations statistically. However, such an approach does not distinguish natural variability correlation from error correlation. This distinction is very important in environmental observations. In any set of environmental observations, whether they are in the form of a time series of observations or a spatial distribution of observations over a network or in the different pixels of an array (or both), there will be a variation due to underlying natural effects. In some cases, this variation will be the desired signal, in other cases it will take the form of 'noise'.

For example, consider the measurement of the 'global mean surface temperature' trend. The Paris Climate Agreement aims to limit the rise in temperature to "well below 2 °C with an aim of below 1.5 °C". However, it is also clear that temperatures vary by many tens of degrees between day and night, between summer and winter, depending on the weather and between the poles and the equator. On the scale of the natural variability, the climate signal cannot be seen. It is only seen by averaging spatially and temporally or by actively removing the diurnal and seasonal effects and by measurements spanning several decades. If two nearby thermometers are compared using the common equations for calculating correlation, then they will be seen to be highly correlated, because they have a common variability around the mean due to the natural variability (diurnal and seasonal cycles and weather).

For some applications, this common variability will be relevant. If we were designing a network of sensors to understand a global mean temperature, we may wish to choose sensors sufficiently far apart that they are not subject to the same weather conditions, for example.

However, in many applications we want to distinguish covariance due to natural effects from covariance due to instrumental effects. If we consider the natural variability to be part of the 'real signal', then we need to evaluate the covariance only due to instrumental effects, sometimes called the 'error covariance'.

In the analysis discussed in this document, we use the terms 'error covariance' (scaled by the uncertainties) and 'error correlation' (normalised to range -1 to +1) to emphasise the distinction from covariance (scaled by the standard distributions) and correlation (normalised) due to natural variability.

2.8 Vocabulary and notation rules

2.8.1 Errors and uncertainties

The distinction between 'error' and 'uncertainty' has been discussed in Section 2.3. In general it is important to note that the only adjectival modifier phrase to be used with 'uncertainty' is 'standard uncertainty' (for an uncertainty that represents the standard deviation of the distribution around the measured value that could be attributed to the measurand), and 'expanded uncertainty' (for an uncertainty that represents multiples of the standard deviation, the multiple described by the coverage factor k, such as $k = 2$ for an approximately 95 % confidence interval for a Gaussian distribution). Because 'uncertainty' describes the spread of the distribution, uncertainties cannot be 'systematic' or 'random' or 'correlated' or 'independent'. These modifiers describe properties of the 'error'.

It is possible to describe effects as 'sources of uncertainty' or 'sources of error'. In general, using the phrase 'sources of uncertainty' is preferable because it encourages an 'uncertainty analysis' way of thinking, rather than the outdated 'error analysis'. An exception is when there is an effect that is, or could be, corrected. In these cases, 'sources of error' may be more appropriate.

The JCGM generally talks about 'the uncertainty associated with […]', for example 'the uncertainty associated with temperature', or 'the uncertainty associated with atmospheric conditions'. This is, however, a rather long phrase and 'the temperature uncertainty' is acceptable, especially where the full phrase would otherwise be repeated in a paragraph. 'The uncertainty in temperature' is less desirable.

2.8.2 Types of uncertainty and types of error

There are three different ways that uncertainties and errors are often qualified, using the pairs random/systematic, Type A / Type B, independent/common. These are often confused, but they do mean different things. These distinctions are not always helpful – it is better to think through the specifics of a particular analysis, but because they are used, and often muddled, it's important to understand the differences.

Type A / Type B are terms that were introduced in the GUM (JCGM 100:2008) to describe methods of evaluating uncertainty. The VIM defines these as:

> Type A evaluation of measurement uncertainty: evaluation of a component of measurement uncertainty by a statistical analysis of measured quantity values obtained under defined measurement conditions.
>
> Type B evaluation of measurement uncertainty: evaluation of a component of measurement uncertainty determined by means other than a Type A evaluation of measurement uncertainty.

This distinction relates to how the magnitude of the uncertainty was evaluated: i.e. whether it evaluated through a statistical method (e.g. calculating the standard deviation of repeat readings), or through any other method (e.g. an understanding of how the instrument works, tests of sensitivity to environmental conditions, using the output of someone else's uncertainty analysis (e.g. through a calibration certificate), etc. This classification can also be applied to how we evaluate the correlation between two quantities, or between measured values taken at different times, locations or in different spectral bands. We can determine such correlations statistically (Type A) or through our understanding of the instrument and measurement methods (Type B).

This classification has no impact on any subsequent uncertainty analysis – uncertainties evaluated by Type A methods are treated identically to uncertainties evaluated by Type B methods in subsequent analysis. Both Type A and Type B methods can be used to evaluate both random and systematic effects, although care is needed to evaluate systematic effects with Type A methods. It is therefore not necessary in any uncertainty analysis to classify uncertainty components as based on Type A or Type B methods, and due to the common, erroneous, confusion that Type A implies random and Type B implies systematic, it may be better not to use these terms.

It is, however, worth using this classification to understand the discussion about the difference between natural variability (natural variance/covariance) and instrument effects. When using Type A methods to evaluate either an uncertainty or a covariance it is important to consider carefully whether the instrument effect can be sufficiently isolated from the natural variability. There are studies that have successfully used Type A methods in the presence of natural variability, e.g., Holl et al 2019, but these are rare.

The random/systematic and independent/correlated distinctions for errors are more useful because they provide information on error correlation structures. In many cases these terms are used interchangeably, with 'random' implying 'independent' and 'systematic' implying 'correlated', though there are authors who intentionally distinguish these further. To understand this, we can consider that there are two aspects to consider when categorising the correlation structure of the errors. The first relates to the origin of the uncertainty, and the second to how the error plays out across multiple observations at different times, locations or in different spectral bands.

The origin of an uncertainty may be a stochastic process, which could never be corrected, even in principle. This is a random effect and includes concepts such as electrical noise, quantum effects, etc. Alternatively, the origin of uncertainty may be a systematic effect that in principle could be corrected for, if we could only characterise it better, and the uncertainty relates to our lack of knowledge of that correction. For example, an uncertainty caused by limits in the calibration process, perhaps due to alignment accuracies or knowledge of aperture sizes, or amplifier gains. Thus, we can have 'systematic effects' and 'random effects'.

How such effects create correlation structures in multiple measurements at different locations, times or in different spectral bands, depends on how the measurement system is set up. For example, a random electrical noise in a measurement of an on-board/on-site calibration target that is measured only occasionally will lead to an error that is common (correlated) for all measurements between two such calibrations, but independent from one calibration to the next. Similarly, a systematic effect such as knowledge of the area of an aperture on the system, would lead to a common error for all observations using that aperture, but if a different aperture is used, e.g., for some spectral channels, then the effect would be independent between channels.

However, such distinctions (between random effects and independent errors) are rarely used, and therefore if they are to be used, need clear explanation. And, as with the Type A/Type B distinction, distinctions based on the origin of the effect do not help inform uncertainty analysis. The only distinction that is useful for uncertainty analysis is the distinction between effects that lead to errors that are common from observation to observation and effects that lead to errors that are independent from observation to observation. The terms 'systematic error' and 'random error' have been used traditionally to make these distinctions. As discussed below (Section 2.9.3), an additional concept of 'structured errors' is also helpful. 'Structured errors' fall between 'random' and 'systematic'.

The terms 'noise' and 'bias' are frequently used in environmental observations to describe random and systematic errors respectively. Again, these terms differ subtly from this common usage. The most common application of these terms is when a data set is compared to another data set considered a reference. The differences will generally be distributed as a spread around a mean. The spread is considered the 'noise' and the offset of the mean difference is considered the 'bias'. This differs subtly from the 'random' and 'systematic' categorisation. The noise/bias distinction is generally calculated from a 'top-down' approach, that is based on comparison with a referenced sensor to quantify the noise or bias, while the random/systematic categorisation comes from a 'bottom-up' uncertainty analysis, where each source of uncertainty is considered separately, and then combined using the GUM methods. With a bottom-up uncertainty analysis, comparisons are used to validate, rather than estimate, uncertainties.

2.8.3 Notation

There are rules in the ISO 80000 standard series about how quantities should be written.

- In equations, italics represent a variable/quantity. Upright text is used for labels, e.g., L_{Earth} for radiance of the Earth, or λ_i for the ith wavelength ('Earth' is a label, i is a counting variable)
- Spaces are inserted before and between units, with a non-breaking space to prevent the unit being on a different line to the measured value. E.g., 300 K and 3 mW m^{-2} sr^{-1} nm^{-1}. The use of / in units should be avoided unless the expression is very simple, e.g., m/s.

- Units should be repeated in ranges (e.g., 300 K – 330 K), or parentheses should be used, e.g., (3 – 5) mW m^{-2} sr^{-1} nm^{-1}.
- A space should be used between a number and the % sign, e.g., 2.5 %. A space is not used for the angular degrees, e.g., 45°, but is for temperatures, e.g., 30 °C.
- Units are always written with a small letter when written out in full (e.g., 300 kelvin) except for "degrees Celsius"
- Unit symbols (e.g., K) are always in upright type.

For uncertainty expressions explicitly, the standard uncertainty is represented by the symbol u. Many environmental observation papers will use the symbol σ to represent a standard deviation of the errors. Such expressions are based on 'error analysis' and can lead to confusion when used for 'uncertainty analysis'. The symbol σ should only be used for a standard deviation. Of course, for Type A methods of evaluating uncertainty, a standard deviation may be calculated in order to evaluate a standard uncertainty; then $u(X) = \sigma(x_i)$.

The uncertainty associated with a quantity X would normally be written $u(X)$, with the variance written as $u^2(X)$; however, to avoid these expressions taking too much space in a complicated equation, u_X and u_X^2 can be used.

When a numerical uncertainty is quoted in addition to a value of a quantity, the GUM recommends expressing the uncertainty in parentheses, e.g., $T = 305.23(15)\ \mathrm{K}$. However, this notation is not well understood outside metrology circles and readers can be confused whether, e.g., here (correctly) the '15' represents 0.15 K or (incorrectly) represents two additional digits. The notation $305.23 \pm 0.15\ \mathrm{K}$ is incorrect, because uncertainty is always positive. To limit confusion, uncertainty should be written explicitly, for example, in a column in a table of results or with phrases such as "with an associated uncertainty of 0.15 K".

2.8.4 Pragmatism in speech and formality in writing

This section has discussed the formal vocabulary and notation to use when stating uncertainties and covariance. It is important that statements in writing are expressed unambiguously, even though some phrases may appear 'long-winded'. In speech, however, it is common to use shortcuts and, in this regard, it is important to remain pragmatic.

2.9 Parametrising covariance matrices

2.9.1 Correlation information needed to propagate uncertainties

As discussed in Section 2.6 above, most Earth observation data are provided as 'datasets' – collections of different observations taken at different times and/or locations and perhaps also at different spectral bands. At one level of processing, it is common for uncertainty analysis to be considered for a single observation, but at higher levels of processing, data from different individual observations are combined in some way. Therefore, scientists doing higher level processing need information about the error covariance structures in the individual observations.

Conceptually, this is achieved by providing an error covariance matrix, which can be summarised by an uncertainty matrix and an error correlation matrix as in Equation (3). Practically, an error covariance matrix may not be possible, as discussed in section 2.6. It can also be difficult to understand an error covariance matrix intuitively, particularly when multiple sources of uncertainty have different correlation structures and when correlation structures are complex. Parametrising the covariance matrix into building blocks helps with thinking about the correlation structures, with generating a covariance matrix and with storing and using a covariance matrix.

2.9.2 Dimensions in correlation

The first step towards parametrising the covariance matrix is to consider what the different correlation 'dimensions of interest' are. The dimensions of interest are the different dimensions for which different observations will be combined or compared at later levels. For an imaging radiometric satellite sensor, these will include along-track and cross-track spatial dimensions, a time dimension and a spectral dimension (wavelengths of a hyperspectral instrument, or bands of a multispectral instrument). There may also be a dimension relating to the viewing angle of the surface (whether the sensor views at nadir or at different angles) or the position of the satellite within the orbit.

For a radar altimeter, the dimensions of interest will include the 'fast time' (the time within a single pulse), the 'slow time' (all longer timescales from the pulse-to-pulse frequency to years and decades), and spatial dimensions. The spatial dimensions are important to understand corrections for how the radar propagates through the atmosphere, or how the waves on the ocean surface interact with the measured waveforms.

The 'dimensions of interest' may not strictly be 'dimensions'. This term represents any structure that helps understand correlations. For example, in a network of air temperature sensors, the dimensions will include time and space, but may also include the type of instrument used or the type of housing the instrument is placed within. For example, all instruments within a Stevenson screen will have common errors, and all instruments that are on automatic weather stations will have different common errors, therefore 'type of instrument' acts as a 'dimension' that gives information about correlation structures. Another 'dimension' could be where an instrument was calibrated if, for example, in an international network, there are groups of instruments all calibrated in one institute and other groups calibrated at other institutes.

2.9.3 Beyond systematic and random: types of structure in a covariance matrix

For each source of uncertainty, the error correlation structure along each relevant dimension of interest needs to be evaluated and documented, recognising that the error correlation structures may be different in different dimensions.

The concepts of 'systematic' and 'random' are insufficient to explain the different correlation structures in Earth observation data sets. There is also a need to consider 'structured' correlations, for quantities whose errors vary in ways between 'systematic' and 'random'. As an example, consider a thermal infrared multiband sensor that makes measurements in different spatial pixels along track and across track. Each along-track row of pixels is a 'scanline'. Thermal infrared sensors are highly sensitive to the onboard temperature, and therefore the instrument gain is monitored in orbit, say every scanline, with a measurement of an onboard blackbody. There will be an uncertainty associated with the temperature of that blackbody. That uncertainty leads to an unknown error in any one calibration. The uncertainty gives us information about the spread of the distribution of such errors, but we cannot know the absolute magnitude of the error for any one calibration. What we do know is that if all spectral channels are calibrated against the same blackbody, the error in blackbody temperature is a systematic effect from channel to channel[4]. Similarly, for all cross-track pixels, the same gain is used, and therefore the same error applies to all pixels, it is a systematic effect across

[4] Note that the error in the blackbody temperature will be the same from channel to channel so the correlation coefficient is 1, but the uncertainty in Earth radiance due to this temperature uncertainty will vary from wavelength to wavelength as shorter wavelengths are more sensitive to blackbody temperature than longer wavelengths. This is handled, in equations (2) and (4) by the sensitivity coefficients in $\boldsymbol{C}$.

track. Along the track, though, a new gain calculation (and therefore new blackbody temperature measurement) is determined every scanline, so the effect is random in that dimension.

It is common with such sensors for the gain determined every scanline to be averaged in a rolling average. In this case, the channel-to-channel correlation is still systematic, as is the cross-track correlation. But the along-track correlation is now triangular, with neighbouring scanlines sharing all-but-one measurements in common in the rolling average, and a triangular drop off in correlation with a half base equal to the averaging window.

To assist with thinking about correlation structures, the FIDUCEO project defined some standard correlation forms. These forms have two purposes – they create a standard way of defining the shape of a correlation matrix, which can easily be encoded in digital form, and they help instrument experts who are new to uncertainty analysis recognise patterns intuitively.

The correlation forms **random** and **rectangular_absolute** relate to random and systematic effects respectively. A random correlation form implies that for different observations within that dimension, the correlation matrix is diagonal. The **rectangular_absolute** form implies that for different observations within that dimension, the correlation matrix is all ones within the range of the rectangle (which may be all observations, or a set, e.g., if there is a block of time when one calibration is used, and then a second block when a new calibration is used, then for the time dimension, the covariance matrix will consist of two squares of ones). The term 'rectangular' refers to the fact that the correlation relative to any one observation will be a rectangular shape on a graph and the 'absolute' refers to the fact that the correlation structure is for points within an absolute range, rather than defined relative to any one point (e.g. an observation just before the calibration is changed is fully correlated with all earlier points, and not correlated with any later points, with a very asymmetrical correlation shape).

More complex structures can also be defined. For example, if there is a simple rolling average of a noisy quantity, then the error correlation relative to any one point takes, as described above, the form of a triangle, dropping from 1 at the peak to zero at the base over the width that the rolling average is calculated over. This creates a correlation matrix banded down the diagonal. FIDUCEO described this as a **triangular_relative**, with 'triangular' representing the shape of the correlation relative to any one point, and 'relative' representing the fact that all observations have a symmetrical correlation structure with their neighbours. It may have been challenging for an instrument expert to have recognised that the correlation structure was triangular, but easy to realise that 'this is a rolling average, and that means triangular'.

FIDUCEO also developed yet more complex structures, e.g., **stepped_triangular** for a situation where a rolling average calibration is not performed every scan line, but only occasionally, and a **bellshaped_relative** for where the rolling average is weighted and the correlation drops off in a more 'bell-shape' than triangle. There are possible periodic correlation shapes as well, and every new application may require new forms. These terms have no inherent priority – any method for defining the error correlation structures in the different dimensions of interest is useful, but the FIDUCEO terms have been helpful for linking physical processes (such as a periodic calibration or rolling averages) to create a correlation matrix.

A list of such correlation forms is given in the Processing document.

2.9.4 Using correlation information

The different error correlation structures, define the shape of the correlation matrix $\boldsymbol{R}$. For example, for a correlation matrix representing a **random** effect, the correlation matrix for different observations in that dimension would be a diagonal matrix, with 1s down the diagonal and 0s elsewhere. For a **rectangular_absolute** effect, the correlation matrix would be a block of ones within the absolute range of the systematic correlation. For example, for six observations with correlation blocks of three (e.g., a calibration is done every three observations), the correlation matrix would be a block diagonal matrix:

$$\boldsymbol{R} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

For a **triangular_relative** correlation form, with a base of three, the correlation matrix would be a banded diagonal matrix:

$$\boldsymbol{R} = \begin{bmatrix} 1 & 2/3 & 1/3 & 0 & 0 & 0 \\ 2/3 & 1 & 2/3 & 1/3 & 0 & 0 \\ 1/3 & 2/3 & 1 & 2/3 & 1/3 & 0 \\ 0 & 1/3 & 2/3 & 1 & 2/3 & 1/3 \\ 0 & 0 & 1/3 & 2/3 & 1 & 2/3 \\ 0 & 0 & 0 & 1/3 & 2/3 & 1 \end{bmatrix}.$$

Thus, for each effect (i.e., each source of uncertainty) such classifications provide a means to define the correlation matrix for the different observations along any dimension of interest (the rows and columns representing e.g., different spatial pixels across a scanline, or different spectral bands, or observations at different times, depending on the application). The covariance matrix is then calculated, for this effect, using equation (4).

The combined covariance matrix due to multiple independent effects (e.g., combining uncertainties associated with different terms in the measurement model that do not have an error correlation between them), is given as a sum of the covariance matrices for each effect, i.e.:

$$\boldsymbol{S}_{\mathrm{com}}(\boldsymbol{y}) = \boldsymbol{C}\boldsymbol{S}_i(\boldsymbol{x})\boldsymbol{C} = \sum_i \boldsymbol{C}_i\boldsymbol{U}_i\boldsymbol{R}_i\boldsymbol{U}_i\boldsymbol{C}_i \qquad (5)$$

Where i represents each effect in turn, and $\boldsymbol{S}_{\mathrm{com}}$ is the combined covariance matrix due to all the different effects.

2.9.5 Simplifying covariance structures

Section 2.9.4 describes how covariance matrices can be built from different effects. Such an analysis may not be desirable because covariance matrices require large data storage and are computationally expensive to manipulate. However, the information stored within them is very sparse and the correlation forms also allow us to parametrise the correlation structures with a much smaller number of quantities. This enables covariance structures to be stored, and multiplied and inverted, in simpler ways. Merchant et al 2019 describes one way such parametrisation can be done. And the NPL CoMET tools (see below), give python modules that can be used to propagate error covariance information in a computationally-efficient manner.

2.9.6 The CoMET tools to support covariance-based analysis

The CoMET toolkit is an NPL-led open-source software project that provides Python tools for the easy handling and processing of dataset error-covariance information. The toolkit aims to abstract away the complexity dealing with measurement uncertainties. More information is available at qa4oe.org/tools.html.

3 Comparisons

3.1 How metrologists use comparisons

National Metrology Institutes (NMIs) have always used comparisons for scientific purposes, to test their methods and, especially, to test their uncertainty analysis. In the early stages of research into new measurement methods, these scientific comparisons show up the unknown unknowns – the differences between participants that are not (yet) considered in the uncertainty analysis. At this point, comparisons tend to be informal and performed, for example, through participants visiting each other's facilities. As a field matures and the technical approaches move from research to operational services, comparisons show increasing agreement between participants. At this point the purpose of comparisons changes from aiding research into auditing and peer review.

This second purpose was formalised in 1999 by the signing of the Mutual Recognition Arrangement (MRA) by the world's NMIs. The MRA says that 'within an appropriate degree of equivalence' the results of one NMI can be considered equivalent to the results of another NMI. In practice this enables world trade and the use of artefacts and instruments calibrated in another country. Being a legal process, the MRA relies on NMIs regularly reviewing each other's calibration and measurement capabilities through a combination of formal peer review and auditing and through formal 'key comparisons' that compare the measurement capability of laboratories – both at the international level (by a handful of laboratories with, generally, the lowest uncertainties) and at the regional level (e.g. within Europe or within Asia-Pacific).

The formal key comparisons are run with strict guidelines and are always blind comparisons (only one 'pilot' laboratory has access to the results before they are published). There is ongoing discussion about the best ways of analysing such comparisons, and in particular about the choice of the Key Comparison Reference Value (KCRV) against which all participants are compared. In very mature fields, where the differences between the measured values of the different participants and the KCRV are consistent with uncertainties, the most common KCRV is the weighted mean of the results of the different NMIs. In less mature fields where results show more spread, this may not be the appropriate choice and alternatives (including 'weighted mean with cut-off' which limits the weight assigned to the laboratories with the lowest uncertainties, or simply using a median value) are considered.

It is important to note that for metrologists the purpose of comparisons is to test and validate uncertainty claims. Comparisons are not performed to estimate uncertainties.

3.2 E_N ratio

Ideally in a bilateral comparison, the analysis is made between two independent observations, each with full uncertainty analysis, to calculate the equivalence ratio:

$$E_N = \frac{|\rho_1 - \rho_2|}{k\sqrt{u_1^2 + u_2^2 + u_{\text{comp}}^2}} \qquad (6)$$

where ρ_1 and ρ_2 are the two independent measured values and u_1 and u_2 are the two standard uncertainties associated with those measured values. u_{comp} is the standard uncertainty associated with the comparison itself (e.g., from a known difference between the observation conditions – matchup uncertainties) and k is the coverage factor for the appropriate confidence interval (usually $k = 2$ to have a confidence interval of ~95 %).

An equivalence ratio $E_N < 1$ suggests that the two measured values agree within their uncertainties, while a larger E_N ratio suggests that at least one uncertainty is underestimated.

For a multilateral comparison (with multiple independent observations, e.g., by several participants) then it is possible either to do every pair of bilateral comparison and to present data in a table showing whether the E_N ratio is greater or less than 1 for each pair of observations, or to determine a comparison reference and calculate the E_N ratio for each participant with respect to the reference, using:

$$E_{N,i} = \frac{\rho_i - \rho_{\mathrm{ref}}}{k\sqrt{u_i^2 + u_{\mathrm{ref}}^2 + u_{\mathrm{comp}}^2}} \tag{7}$$

Note that while the reference can be arbitrarily chosen, some care must be taken in choosing it as it is easy to interpret an $E_N > 1$ as implying “bad” data. The reference is the KCRV in metrological comparisons (Section 3.1).

3.3 Earth observation comparisons

The QA4EO guideline 4, “A guide to comparisons”, sets out best practices for EO comparisons. These guidelines are based on the guidelines established by metrology institutes for the MRA comparisons.

There have been several FRM projects that have hosted CEOS comparisons of the instruments used as part of FRM systems. The FRM4STS project performed comparisons of radiation thermometers both in the laboratory (against reference blackbodies) and in situ (over water and land surfaces). Similarly, the FRM4SOC project performed comparisons of ocean colour validation sensors, again including both laboratory and field (in water and above water) comparisons. Currently, FRM4VEG project plans to conduct and intercomparison exercise of drone mounted validation sensors over vegetation. Such FRM comparisons are similar in principle to metrological comparisons, and similar analysis methods are possible.

FDRs and TDPs from satellites can also be compared, both sensor-to-sensor (FDR to FDR; or TDP to TDP) and sensor-to-ground (FDR to FRM or TDP to FRM). As with metrological comparisons, such EO comparisons are best performed when all the compared observations have complete uncertainty analysis and the observational differences are well understood, allowing the E_{N} ratio to be calculated. They differ from metrological comparisons mostly in the sense that usually a very large number of comparisons are done (e.g., over multiple scenes / multiple match ups and at different times). Therefore, comparison analysis must also consider how to combine uncertainties from multiple comparisons. That requires a strong understanding of the error correlation structures in the data.

Further guidelines on EO comparisons are under development.