GAIA-CLIM Report / Deliverable D3.2

Gap Analysis for Integrated Atmospheric ECV CLImate Monitoring:

Generic metrology aspects of an atmospheric composition measurement and of data comparisons

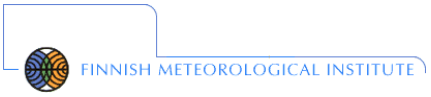
A Horizon 2020 project; Grant agreement: 640276

Date: 1 March 2016

Lead Beneficiary: BIRA-IASB

Nature: R

Dissemination level: PU



Work Package	WP 3 – Comparison error budget closure – Quantifying metrology related uncertainties of data comparisons
Deliverable	D3.2
Title	Generic metrology aspects of an atmospheric composition measurement and of a data comparison
Nature	R
Dissemination	PU
Lead Beneficiary	Royal Belgian Institute for Space Aeronomy (BIRA-IASB), Brussels, Belgium
Date	1 March 2016
Status	Version 1
Editors	Tijl Verhoelst (BIRA-IASB) and Jean-Christopher Lambert (BIRA-IASB)
Contributors	Tijl Verhoelst (BIRA-IASB), Jean-Christopher Lambert (BIRA-IASB) , Fabio Madonna (CNR), Alessandro Fassò (UniBergamo), and Peter Thorne (NUIM)
Contacts	tijl.verhoelst @ aeronomie.be, j-c.lambert @ aeronomie.be
URL	http://www.gaia-clim.eu/

This document has been produced in the context of the GAIA-CLIM project. The research leading to these results has received funding from the European Union's Horizon 2020 Programme under grant agreement n° 640276. All information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission has no liability in respect of this document, which is merely representing the authors' view.

Table of Contents

Executive summary

GAIA-CLIM aims at supporting the European Commission's Copernicus Programme by assessing and improving the fitness-for-purpose of sub-orbital (ground- and balloon-based) reference measurements in the validation of observational data sets from satellites. In particular, the project aims at improved traceability and uncertainty characterization, of the individual sub-orbital measurement (systems) and of the comparison with satellite data.

A key issue in the geophysical validation of satellite data sets with respect to sub-orbital reference measurements is the interpretation of their differences in terms of known, quantified, uncertainties. This aspect includes not only the measurement uncertainties associated with the individual measurements, but also the additional uncertainties that appear when comparing different perceptions of the inhomogeneous and variable atmosphere, that is, when comparing data sets characterized with different sampling and smoothing properties, both in space and time. Those "comparison uncertainties" are the main topic of investigation for GAIA-CLIM Work Package 3, and this document, Deliverable 3.2, describes the principle of co-location mismatch, and how the resulting uncertainties can be decomposed through a careful metrological analysis of the measurements and their comparison. Moreover, it provides an overview of available methods to quantify these uncertainties so that they can be taken into account when interpreting the results of a data validation (data comparison).

In Section 1, the general issue of co-location mismatch is introduced, followed by an introduction to the principles of metrology, and a first, conceptual, decomposition of the uncertainty budget of a data comparison. This decomposition reveals the need for three types of metrological analyses:

1) The metrological traceability of the measurement, described in Section 2 and required to properly estimate the measurement uncertainty,
2) The spatiotemporal properties of the measurement, described in Section 3 and required to estimate the additional uncertainties resulting from the fact that measurements are never point-like and the atmosphere never fully homogeneous and constant.
3) The metrology of a data comparison, described in Section 4, and required to estimate the uncertainties due to co-location mismatch in the broadest sense, due to point-like spatiotemporal differences but also smoothing differences. Only at this point can the uncertainty budget of a comparison be closed and interpreted in terms of data quality.

Section 5 mentions some key methods to quantify the uncertainties due to spatiotemporal properties and mismatches. As such, this document is meant to serve as a theoretical background and guidance document in support of several upcoming deliverables from WP3, such as D3.4, the TN on measurement mismatch studies, D3.6, the library of smoothing and sampling error estimates, and D3.7, the set of tools aimed at integrating WP3 work into the Virtual Observatory developed in WP5.

1. Introduction

High-quality observational datasets from satellites constitute a key component of the European Commission's Copernicus programme, which aims at providing users (mainly policy makers and public authorities) with reliable and up-to-date information related to environmental and security issues. The climate change and atmosphere monitoring services in particular rely heavily on observations from current and future satellite instruments measuring both key meteorological variables such as temperature and humidity, and atmospheric composition, including greenhouse gases and health-endangering pollutants.

For these services to be reliable and effective, it is a prerequisite that the underlying data sets are fit-for-purpose, i.e. that their quality is assured and that they meet user requirements. Quality assurance in the context of satellite remote sensing has been defined in the context of the CEOS- and GEO-endorsed Quality Assurance for Earth Observation (QA4EO) framework, as the need *for fully traceable Quality Indicators*. In practice, the extent to which the satellite measurements agree with ground-based reference measurements is an essential such quality indicator. Clearly, this agreement needs to be assessed in the context of the reported uncertainties, both those on the satellite and on the reference measurements. As such, traceability of the data production and of the associated uncertainties is another crucial quality indicator. The H2020 project GAIA-CLIM aims specifically at improving the traceability and uncertainty characterization of the sub-orbital reference measurements used to assess the quality of the satellite data sets. Moreover, it also addresses the uncertainty budget of these crucial satellite-to-reference comparisons from which several quality indicators are derived, but which require careful consideration of the additional errors due to co-location mismatch, i.e. the unavoidable differences in measurement times, locations and spatiotemporal smoothing.

1.1. Context and aims of this document

Two key scientific objectives of GAIA-CLIM concern the uncertainties on sub-orbital reference measurements and the additional uncertainties related to the intercomparison of measurements from different instruments and obtained at different locations and measurement times: (1) objective S3, mainly targeted within WP2, is to provide *reference quality measurement uncertainties that are traceable to recognized measurement standards*, and (2) objective S4, mainly targeted within WP3, is to understand and quantify *the metrology of a data comparison, including the additional uncertainties that arise from spatiotemporal mismatches between both observing systems*. As such, these objectives target the uncertainty budget of a comparison between a satellite measurement and a sub-orbital reference measurement, which, in the ideal case of perfect co-location, can be represented mathematically as:

$$\left|m_1 - m_2\right| \leq k\sqrt{u_1^2 + u_2^2} \quad (1)$$

where the left-hand side represents the observed difference between both measurements, $\boldsymbol{k}$ is a coverage factor, and $\boldsymbol{u_1}$ and $\boldsymbol{u_2}$ represent the measurement uncertainties (e.g. Immler et al., 2010). However, in case of spatiotemporal mismatch, i.e. non perfect co-location, an additional uncertainty term, $\boldsymbol{\sigma^2}$, must be included:

$$\left|m_1 - m_2\right| \leq k\sqrt{\sigma^2 + u_1^2 + u_2^2} \quad (2)$$

The present document, Deliverable 3.2, represents key output from WP3, and targets mainly objective S4, i.e. σ^2 in Eq.(2). In particular, it describes the spatiotemporal metrology of atmospheric composition measurements and their intercomparison at a generic level, illustrated with some practical examples. It sets the stage and will serve as a theoretical reference document for three future deliverables: D3.4, reporting on measurement mismatch studies and their impact on data comparisons (due M24), D3.5, the set of tools that will be integrated into the Virtual Observatory (also due M24), and D3.6, a library of spatiotemporal uncertainty estimates (due M30). Figure 1 visualizes the relation between the current deliverable, the different tasks within WP3, and the corresponding future deliverables.

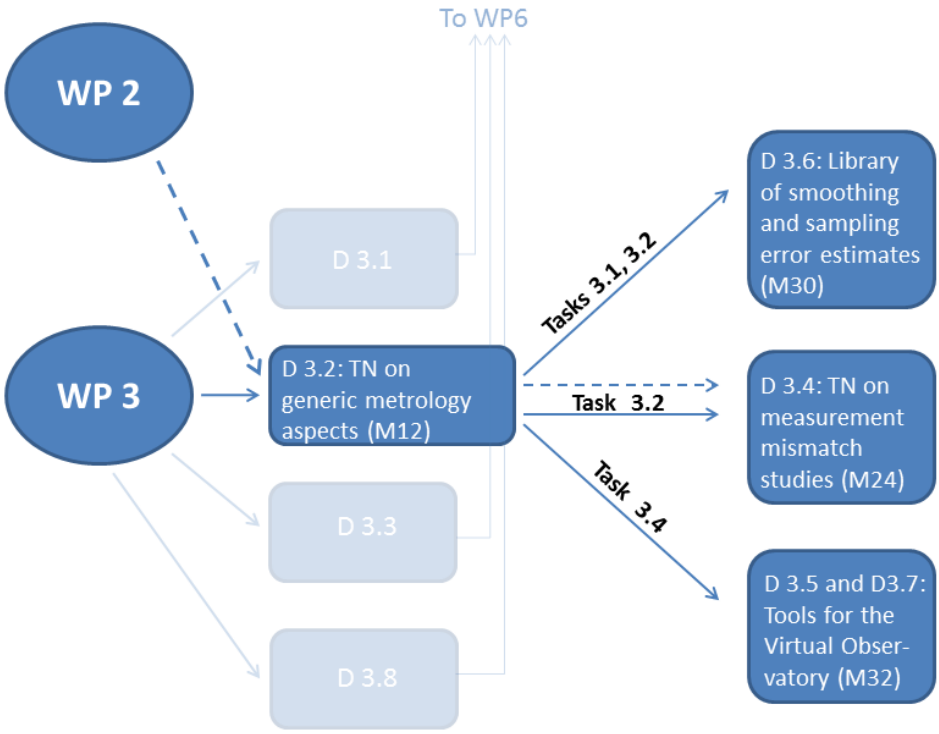


Figure 1: Inter-linkages between D3.2 and other deliverables from WP3. Because WP2 also deals with some metrology aspects and provides the measurement uncertainties required for uncertainty budget closure aimed for in Task 3.2 and deliverable D3.4, it is included in this graph.

Since the ultimate aim is to close the uncertainty budget of a comparison between satellite and sub-orbital reference measurements (cfr. Eq. 2), the availability of reliable and traceable measurement uncertainties (S3, WP2) is essential to the work performed within WP3. Consequently, the present document will also briefly cover those metrology aspects that are primarily dealt with by WP2, in this way further clarifying which metrological aspects are (or should be) part of the reported measurement uncertainties, and which are to be taken into account in addition to the measurement uncertainties when performing intercomparisons.

Note that the current document deals with Level-2 data only, i.e.(retrieved) columns and profiles, as this is the focus of WP3 within GAIA-CLIM. For Level-1 (ir)radiance data, similar considerations may apply, but this is beyond the scope of the work planned here. Level-3 data, i.e. gridded averages of Level-2 data, add another layer of metrological issues, in particular regarding representativeness of the averages due to the particular sampling patterns of the sounders and networks, but also this topic is largely beyond the scope of the current document.

1.2. Principles of metrology

In its broadest definition, metrology concerns the "science of measurement and its application" (BIPM, 2012). This includes in particular (1) the definition of internationally accepted units of measurement, (2) the realisation of these units of measurement in practice, and (3) the application of chains of traceability linking measurements made in practice to reference standards, i.e. making these measurements "metrologically traceable". In practice, these aims give rise to a large field of research focused on characterizing measurements and measurement systems in exhaustive detail, covering a wide variety in measurement (system) properties, as illustrated by Figure 2.

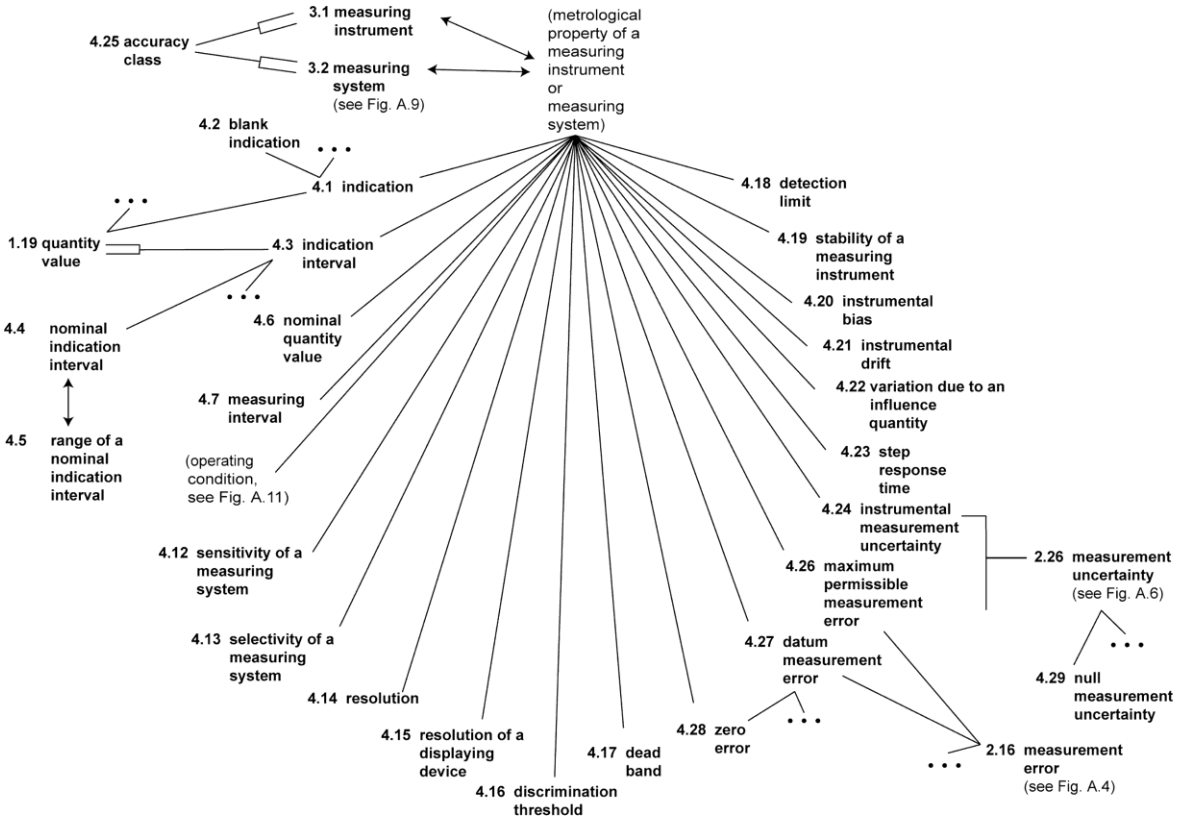


Figure 2: A (limited) overview of measurement (system) properties typically investigated and quantified in metrological research. This graph is Fig. A.10 from the VIM (BIPM, 2012).

In the context of remote sensing for earth observation (EO), metrological studies usually aim for an explicit description linking the measurement (often a retrieval of a physical parameter from an (ir)radiance measurement) to an SI standard through an unbroken chain of calibrations and documented processing steps. These *traceability chains* are described in more detail in Section 2, including the role these chains play in assessing the uncertainty on the measurement or retrieved quantity. However, there are additional properties of measurements that are not as commonly studied but still of great importance when comparing the measurement with other measurements or with model output, which are also to be considered part of the field of metrology. These include the sampling and smoothing properties, in the vertical, horizontal, and temporal dimension. This is further elaborated in Section 3. As already touched upon in Section 1.1, when comparing measurements additional uncertainties arise due to imperfect co-location (i.e. everything making up σ^2 in Eq. (2), and their quantification is the aim of what can be called " comparison metrology". Section 4 expands the different aspects to be treated for a detailed understanding of these

additional uncertainties, and several practical methods to quantify them are presented in Section 5. A first –conceptual- decomposition of σ^2 is already described in the following section.

In summary, the current document deals with the following metrological aspects of remote sensing measurements and their (inter-)comparison:

- Traceability to SI standards (Section 2),
- Spatiotemporal properties of a single measurement (Section 3),
- Measurement comparisons and co-location mismatch (Section 4).

1.3. Metrology of a data comparison and associated errors

In most data comparison endeavors, such as ground-based validation exercises, a compromise must be made between on the one hand abundance of comparisons pairs, and on the other hand additional comparison errors, not related to the actual measurements uncertainties but due to non-perfect co-location in space and time. This non-perfect co-location is a consequence of both a difference in spatio-temporal sampling, i.e. a satellite pixel centre never coincides exactly with the ground station location, and a difference in resolution, i.e. in the way each instrument has a smoothed perception of the inhomogeneous and variable atmosphere. This is visualized conceptually in Figure 3.

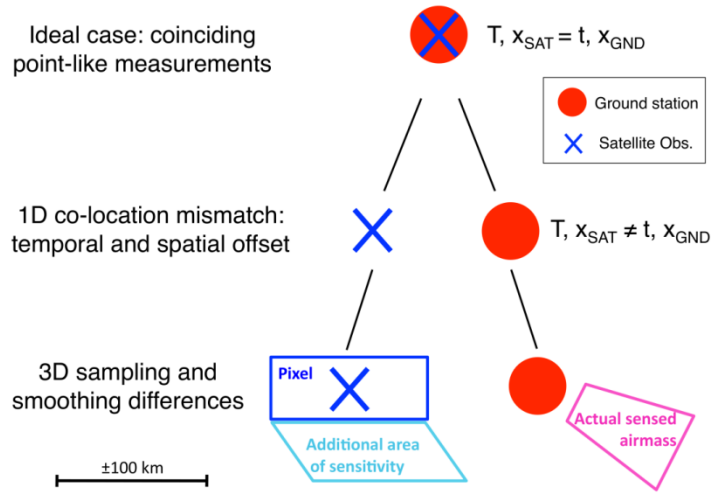


Figure 3: Conceptual visualization of the metrology of a satellite-to-ground measurement comparison. Ideally, both measurements are point-like in space and time, and coincide perfectly. In practice, the sampling pattern of the satellite sounder and the fixed locations of the ground network induce sampling difference errors. Furthermore, differences in resolution, or, more broadly, in area of actual measurement sensitivity, induce additional smoothing difference errors. Figure reproduced from Verhoelst et al. (2015).

A formal representation of the uncertainty budget of a comparison was already explored by Rodgers (1990,2000) and by Rodgers and Connor (2003), and further elaborated by von Clarmann (2006). Lambert et al. (2012) includes the multi-dimensional perspective, dealing with horizontal smoothing errors and errors due to less than perfect co-incidence. While the above-mentioned papers deal with uncertainties in terms of covariance matrices, and thus assume the errors to have a Normal distribution, it is instructive to focus first on the decomposition of individual differences (i.e. per

comparison pair) in terms of the different error sources, before deriving an uncertainty budget. A pair of co-located measurements, $\boldsymbol{x}_{SAT}$ and $\boldsymbol{x}_{GND}$, can be related to each other as:

$$x_{SAT} = x_{GND} + \varepsilon_{total} \quad , (3)$$

with

$$\varepsilon_{total} = -\varepsilon_{GND,rand} + \varepsilon_{SAT,rand} - \varepsilon_{GND,sys} + \varepsilon_{SAT,sys} + \varepsilon_{S4D} + \varepsilon_{dx/d4D} \quad , (4)$$

where

- $\varepsilon_{GND,rand}$ and $\varepsilon_{SAT,rand}$ represent the random errors related to the measurement uncertainty of both sensors,
- $\varepsilon_{SAT,sys}$ and $\varepsilon_{SAT,sys}$ represent the systematic errors related to the measurement uncertainty of both sensors,
- ε_{S4D} represents the so-called smoothing difference error, which contains horizontal, vertical and temporal components, and
- $\varepsilon_{dx/d4D}$ represents the so-called sampling difference error, which also contains horizontal, vertical and temporal components.

The sampling difference error is the error that would occur even if the measurements were point-like, but not perfectly coinciding. Note that this is not the same as the sampling error resulting from an incomplete sampling of a signal (e.g. von Clarmann, 2006). The smoothing difference error is the error that would occur even if the measurements have coinciding nominal locations (e.g. station location and pixel centre), but different resolutions.
From the error budget described in Eq. (4), it is in principle possible to derive an uncertainty budget in terms of variances (column measurements) or covariance matrices (profile measurements), e.g. to calculate $\boldsymbol{\sigma}^2$ in Eq. (2), but that would implicitly assume a Normal distribution of the errors, and the absence of correlations between the different terms. As shown in Verhoelst et al. (2015), these assumptions are not always valid, and it is therefore advised to investigate errors, and their probability density functions (PDFs), instead of uncertainties, whenever possible.

1.4. Definitions

The nomenclature followed throughout this document is based as far as possible on the international conventions published by the *Bureau International des Poids et Mesures* (BIPM) in the form of two key documents: the *Vocabulaire International de Métrologie* (VIM), and the Guide to the expression of Uncertainty in a Measurement (GUM), see also Sect. 1.5. Further definitions were taken from a list of authoritative documents, and compiled into a reference table in the framework of the EC FP7 project QA4ECV (http://qa4ecv.eu). These conventions are applied strictly in the current document, and they are provided as Annex A for reference. Frequently used terms and specific concepts not defined in the QA4ECV table are listed below. Note that these are consistent with the summary on terminology compiled within GAIA-CLIM as the "Guide to Uncertainty in Measurement and its Nomenclature".

- **Metrology**: Definition 2.2 in the VIM: "science of measurement and its application (NOTE: Metrology includes all theoretical and practical aspects of measurement, whatever the measurement uncertainty and field of application)."

- **Measurement error**: Definition 2.16 in the VIM: "Measured quantity value minus a reference quantity value".
- **Systematic measurement error**: Definition 2.17 in the VIM: "component of measurement error that in replicate measurements remains constant or varies in a predictable manner"
- **Random measurement error**: Definition 2.19: "component of measurement error that in replicate measurements varies in an unpredictable manner"
- **Measurement uncertainty**: Definition 2.26 in the VIM: "non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used"
- **Measurement bias**: Definition 2.18 in the VIM: "estimate of a systematic measurement error"
- **Uncertainty budget**: Definition 2.33 in the VIM: "statement of a measurement uncertainty, of the components of that measurement uncertainty, and of their calculation and combination"
- **Error budget**: Undefined in the VIM, but easily derived from the definition of the uncertainty budget: "statement of a measurement error, of the components of that measurement error, and of their calculation and combination"
- **Smoothing error**: the difference between the measurement and the truth at the nominal measurement location due to the smoothing properties of the instrument
- **Smoothing difference error**: Not to be confused with the smoothing error, the smoothing difference error represents the difference between two measurements, due to the differences in smoothing of the truth. E.g. for measurements with very similar smoothing properties, the smoothing difference error may be much smaller than the individual smoothing errors. See also Section 1.3.
- **Sampling error:** the difference between the measurement and the truth due to an incomplete sampling of the signal
- **Sampling difference error**: Not to be confused with the sampling error, the sampling difference error represents the difference between two measurements due to differences in sampling of the truth. See also Section 1.3.
- **Co-location mismatch**: Generic term implying the mismatch between two co-located measurements in spatiotemporal smoothing and sampling. It causes smoothing and sampling difference errors in the horizontal, vertical and temporal domains.

1.5. Related documents and projects

Within GAIA-CLIM

- The **Gap Assessment and Impacts Document** (GAID), in particular the gaps identified by WP2 and WP3, identified as 2.xx and 3.xx
- Deliverable D3.1, the **initial input from WP3 to the GAID**, which includes a literature review
- The GAIA-CLIM Guidance Note **Guide to Uncertainty in Measurement & its Nomenclature**

- Deliverable D3.4 (upcoming): **Measurement mismatch studies and their impact on data comparisons**
- Deliverable D3.6 (upcoming): **Library of smoothing/sampling error estimates for key atmospheric composition measurement systems, and smoothing/sampling error estimates for key data comparisons**

Nomenclature and metrology principles

- VIM, 3rd edition: ***International Vocabulary of Metrology – Basic and General Concepts and Associated Terms*** (VIM 3rd edition) JCGM 200:2012
- GUM: **Evaluation of measurement data – Guide to the expression of uncertainty in measurement** , JCGM, JCGM 100:2008, 2008, http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf
- **Data Modeling for Metrology and Testing in Measurement Science**, Pavese, F. & Forbes, A. B. (Eds.), Springer Science + Business Media, 2009
- **Measurement Uncertainty Analysis Principles and Methods**, NASA Measurement Quality Assurance Handbook – ANNEX 3, 2010
- **Annex A of the current document: the QA4ECV terms and definitions**

Related projects

- QA4ECV (www.qa4ecv.eu): Quality Assurance for Essential Climate Variables, aims at "*developing an internationally acceptable Quality Assurance (QA) framework that provides understandable and traceable quality information for satellite data used in currently evolving climate and air quality services.*" It serves in particular as preparation for the Copernicus Climate Change Service.
- FIDUCEO (www.fiduceo.eu): Fidelity and uncertainty in climate data records from Earth Observations aims at building "*nine new climate datasets from Earth Observation using a rigorous treatment of uncertainty, informed from the discipline of metrology*."
- MetEOC/MetEOC-2 (http://www.emceoc.org) aims at improving the metrology in Earth Observation (EO), and includes WPs on satellite calibration test sites, climate indicators, SI traceability of biophysical parameters, solar irradiance, and ECV measurements.
- GeoMON (website no longer active), was an EC FP6 project focusing on atmospheric composition and with multiple aims, ranging from better data production to validation and integration in models. It dealt extensively with the characterization of the horizontal smoothing properties of key atmospheric composition measuring instruments.
- NORS (nors.aeronomie.be), the "demonstration Network Of ground-based Remote Sensing observations in support of the Copernicus Atmospheric Service" was an EU FP7 project aimed at demonstrating the value of ground – based remote sensing data from the Network for the Detection of Atmospheric Composition Change (NDACC) for quality assessment and improvement of the Copernicus Atmospheric Service products.

2. Metrology, part I: Traceability chains and uncertainty estimates

For a (sub-orbital) measurement to be a reliable reference, it is crucial that it can be linked to an authoritative standard (e.g., *Système International*, SI) through an unbroken and documented chain of well-characterized calibrations and operations. In other words, it requires *metrological traceability*. Such traceability is achieved for selected reference measurements such as the temperature and humidity profiles from radiosondes contributing to the GCOS Reference Upper Air Network, GRUAN (Immler et al., 2010). Work in that direction is ongoing for many other instrument types and networks, e.g., with a focus on the Network for the Detection of Atmospheric Composition Change (NDACC) in WP2 of GAIA-CLIM, in other EU-supported projects such as QA4ECV, and in other initiatives such as the SPARC/IGACO-O3/IOC Ozonesonde Data Quality Assessment (O3S-DQA). Similar efforts exist to improve the traceability of satellite data records, e.g. within the EU H2020 projects FIDUCEO and MetEOC. Section 2.1 briefly illustrates what is meant by a traceability chain for data production.

A key motivation for such (metrological) traceability is the determination of the uncertainties on the final measured or retrieved geophysical quantities by proper propagation and addition of the uncertainties introduced by the different processing steps into a total uncertainty budget. This process is described further in Section 0.

It must be noted though that these uncertainties are at least partly theoretical and/or approximate in nature: they are based on measurement models, uncertainty propagation methods (e.g., linearized analytical propagation), and auxiliary data which are often affected by the same limitations as the measurement and retrieval process themselves. Moreover, for non-reference measurements, the chain itself usually is incomplete and includes steps for which uncertainty information is virtually impossible to obtain, for instance due to manufacturers not sharing such information. On top of that, uncertainty propagation usually concerns only the effect of random errors, and little can be done to quantify systematic errors when these have not all been calibrated out. Consequently, measurements and their uncertainties require to be validated by independent means. This is the topic of Section 0.

2.1. Metrological traceability of data production

As an example, **Figure 4** presents a high-level traceability chain for an O_3 column products derived from nadir satellite measurements of UV-Vis scattered light. It covers the processing from calibrated radiance measurements to retrieved ozone column. A similar chain can be constructed which links the radiance measurement to both the detector output and the calibration files characterizing the detector properties in direct relation to the SI units. Most of the items in this chain can be decomposed further, down to the level of (auxiliary) data sources, equations, and elemental processing algorithms, including version numbers and references to essential documentation such as the Algorithm Theoretical Basis Document (ATBD). In practice, different maturity levels of traceability can be distinguished. Within the EU FP7 project QA4ECV, these are:

A. Basic processing information,
B. Detailed processing and uncertainty information,
C. Metrological traceability.

For most of the ECVs and instruments addressed by GAIA-CLIM, traceability is still limited (typically around level B), and progress on this is a key aim of WP2.

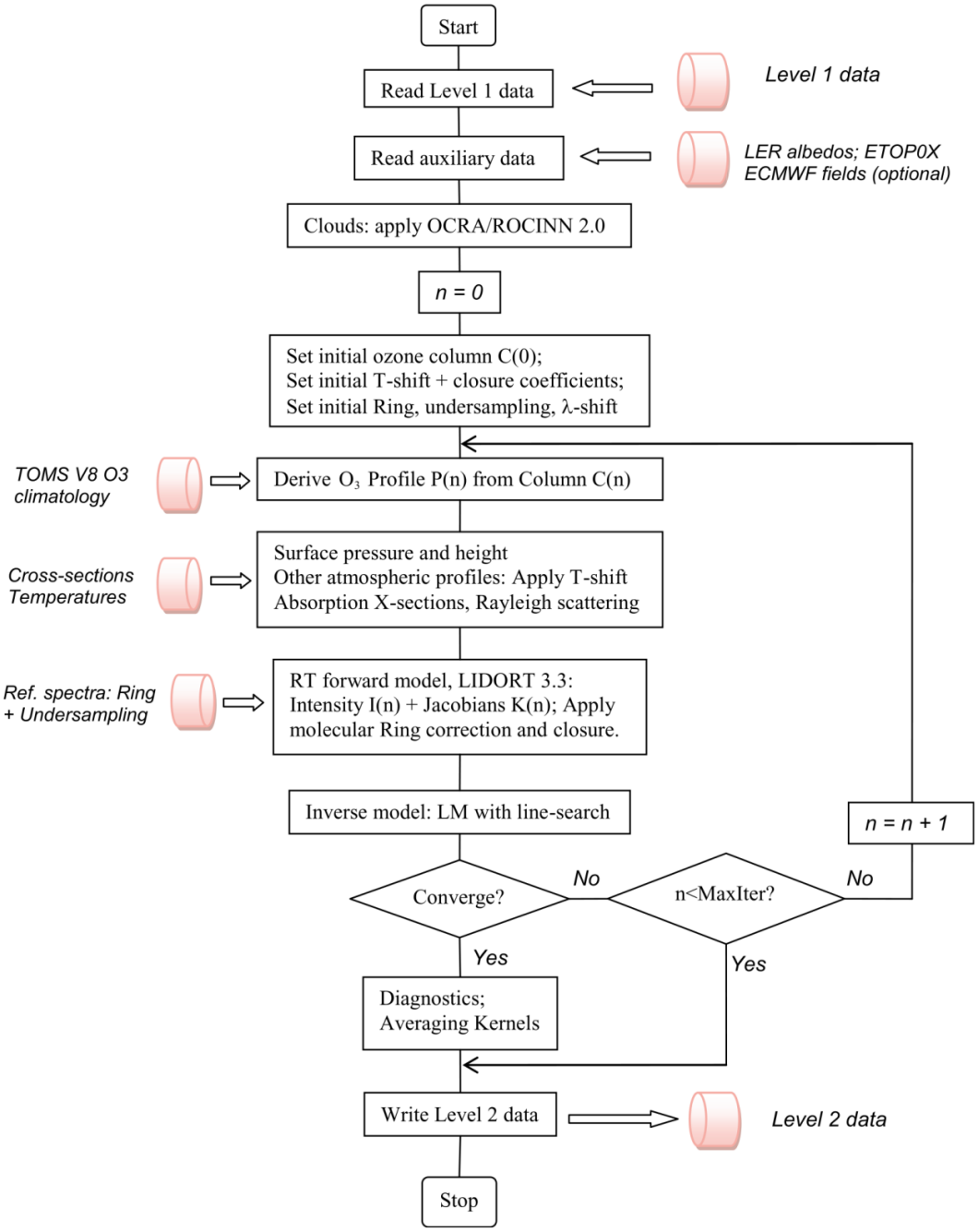


Figure 4: Example of a metrological traceability chain, in this case for the derivation of O_3 column products from nadir satellite measurements of UV-Vis scattered light, using the GDP5 processor. Reproduced from Van Roozendael et al. (2012).

2.2. Propagation of uncertainties along the traceability chain of data production

While the most straightforward method of assessing uncertainty is through repetition of the measurement under identical conditions, i.e. Type-A uncertainty determination following the GUM nomenclature, this approach is rarely possible when observing (atmospheric) ECVs. Indeed, because of atmospheric variability, a large set of measurements would have to taken simultaneously, probing exactly the same air masses, which is most often neither technically nor financially feasible.

Instead, the total uncertainty must be reconstructed from known uncertainties on the different processing steps and auxiliary data sets, which can be derived either within laboratory settings or through dedicated sensitivity and intercomparison tests. Clearly, the traceability chain plays a crucial role in collecting and combining all the uncertainty information into a total uncertainty budget.

Propagation of uncertainties can be performed in multiple ways. If the end product can be linked to a more fundamental measurement with an explicit mathematical formulation, it is possible to propagate uncertainty due to random errors in an analytical way, under the assumption of local linearity, by using the first-order derivatives of the different processing steps (e.g. Boersma et al. 2004, De Smedt et al. 2012). Alternatively, full error probability distribution functions (PDFs) could be propagated with Monte Carlo methods, avoiding both the assumption of Gaussian error distributions and (locally) linear processing steps. Finally, this approach can be extended further to more general ensemble-type studies and sensitivity analyses in case of highly complex traceability chains (e.g. Hendrick et al., 2011, Dirksen et al., 2014). The end result is a detailed decomposition of the uncertainty on the derived geophysical quantity as a function of key features of the measurement (and of the retrieval in case of remote sensing), as illustrated in Figure 5 for a radiosonde measurement of an atmospheric temperature profile.

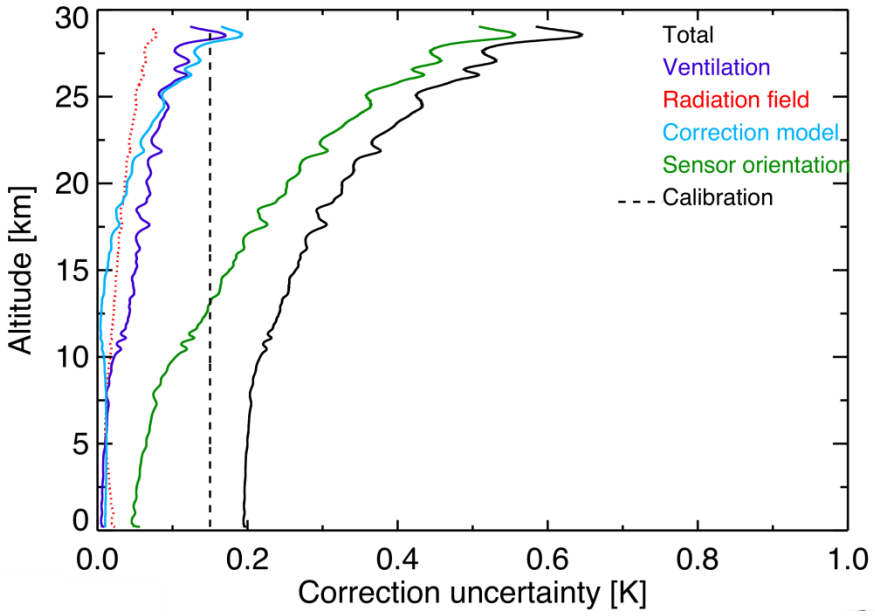


Figure 5: Total uncertainty budget for a Vaisala RS92 radiosonde temperature profile measurement, reproduced from Dirksen et al. (2014).

2.3. The need for independent validation

The introduction of Section 2 already touched upon the limitations of uncertainty propagation through traceability chains, in particular for most satellite atmospheric sounders. Consequently, such data sets may have poorly quantified uncertainties, which can be random or depend on several influence quantities such as solar zenith angle, surface albedo, cloud cover, instrumental degradation, etc. To reliably assess the quality and fitness-for-purpose of satellite data sets, the validation with well-controlled and documented (ground-, balloon- and aircraft-based) reference measurements is therefore of utmost importance.

The process of ground-based validation of atmospheric ECVs itself can be described in the form of a chain of processing and analysis steps, for which detailed protocols have already been established, e.g., within the FP7 projects MACC, PASODOBLE, and QA4ECV, and in the context of ESA's GMES Service Element (GSE) and Climate Change Initiative (CCI). A detailed overview of the validation process is described and applied by Keppens et al. (2015) to the validation of nadir ozone profile data retrieved from MetOp-A GOME-2 measurements with ozonesonde and lidar measurements. In short, the scheme as it was elaborated further for the validation server under development in QA4ECV, consists of the following steps:

1. Translation of user requirements into validation requirements
2. Satellite data selection, filtering and post-processing
3. Data content study (DCS) of satellite dataset
4. Information content study (ICS) of satellite dataset
5. Selection and characterisation of correlative data
6. Identification and characterisation of co-located data pairs
7. Homogenization: Resampling, smoothing, and conversions of representation systems and units
8. Data content and information content studies of co-located satellite and correlative datasets only
9. Data comparisons
10. Derivation of statistical Quality Indicators: bias, spread, stability, dependence on influence quantities
11. Production of a user-oriented report
12. Verification of compliance with user requirements

Step 6 includes the delicate task of choosing appropriate co-location criteria, which requires a thorough understanding of the smoothing and sampling properties of the different instruments, and of atmospheric variability. These are the main topics of Section 3 of the current document. Steps 9 and 10 are addressed in Section 4, which deals with the metrology of a data comparison. Also covered in that section are some aspects of step 7: the harmonization, e.g., in terms of vertical resolution and smoothing.

3. Metrology, part II: Effects of spatiotemporal inhomogeneities

As described in the previous section, metrology of an atmospheric ECV measurement is first of all focused on the traceability to a community-agreed standard, and on the determination of the measurement uncertainty through proper propagation of different sources of known errors and uncertainties along the traceability chain. In such studies, little attention is paid to the impact of inhomogeneities and variability in the atmospheric field on the measurement: the measurement is either treated as a point-like sample of the atmosphere, or the atmosphere is assumed to be homogeneous and constant over the measurement area and integration time. The purpose of the current section is to discuss the actual spatiotemporal properties of different measurement types, and to combine these with estimates of the natural variability to get a quantitative grip on the additional errors and uncertainties to be taken into account when making comparisons with models or with other measurements with different spatiotemporal properties.

3.1. The inhomogeneous atmosphere

An inhomogeneous and variable atmosphere will lead to additional errors and uncertainties if variations in the field occur at scales smaller than (or similar to) the scale of the measurement, i.e. if the measurement system does not satisfy the Nyquist criterion. In the EO case, the scale of the measurement system refers to both its sampling pattern and spatio-temporal resolution. For a balloon-borne sonde, this includes for instance the balloon trajectory, the detector read-out rate, and the detector response time. For a retrieval based on satellite-measured (ir)radiance, this includes the geospatial sampling pattern, the averaging of the field over a pixel footprint, and more complex radiative transfer and retrieval sensitivity along the entire line-of-sight between photon source and detector. Some advanced, multi-dimensional retrieval schemes taking into account horizontal inhomogeneity of the retrieved quantity do exist (e.g., Dinelli et al. 2010, Kiefer et al., 2010), but in general the atmosphere is assumed to be homogeneous and constant. The scales of atmospheric variability depend strongly on ECV and atmospheric regime. For instance, the upper panel of Figure 6 illustrates that the total ozone column (TOC) field contains significant variability at a scale of a few degrees, which is a typical measurement scale for both ground-based and satellite observations of TOC. While the variability is almost negligible w.r.t. typical measurement uncertainties in equatorial regions, this is definitely not the case at middle latitudes and near the polar vortex (Verhoelst et al. 2015). Part of the network providing ground-based reference measurements of TOC is indicated, illustrating how spatial variability will affect the measurements at different stations in different degrees, and it suggests the possibility of taking this into account in future network design. The bottom panel of Figure 6 presents a similar simulations but for stratospheric water vapour concentrations. Information on natural variability is to some extent available for most of the ECVs targeted by GAIA-CLIM (see the Appendices of D3.1), but not necessarily down to the scale of the measurements, and the lack of knowledge on those scales is therefore identified as a gap in the GAID.

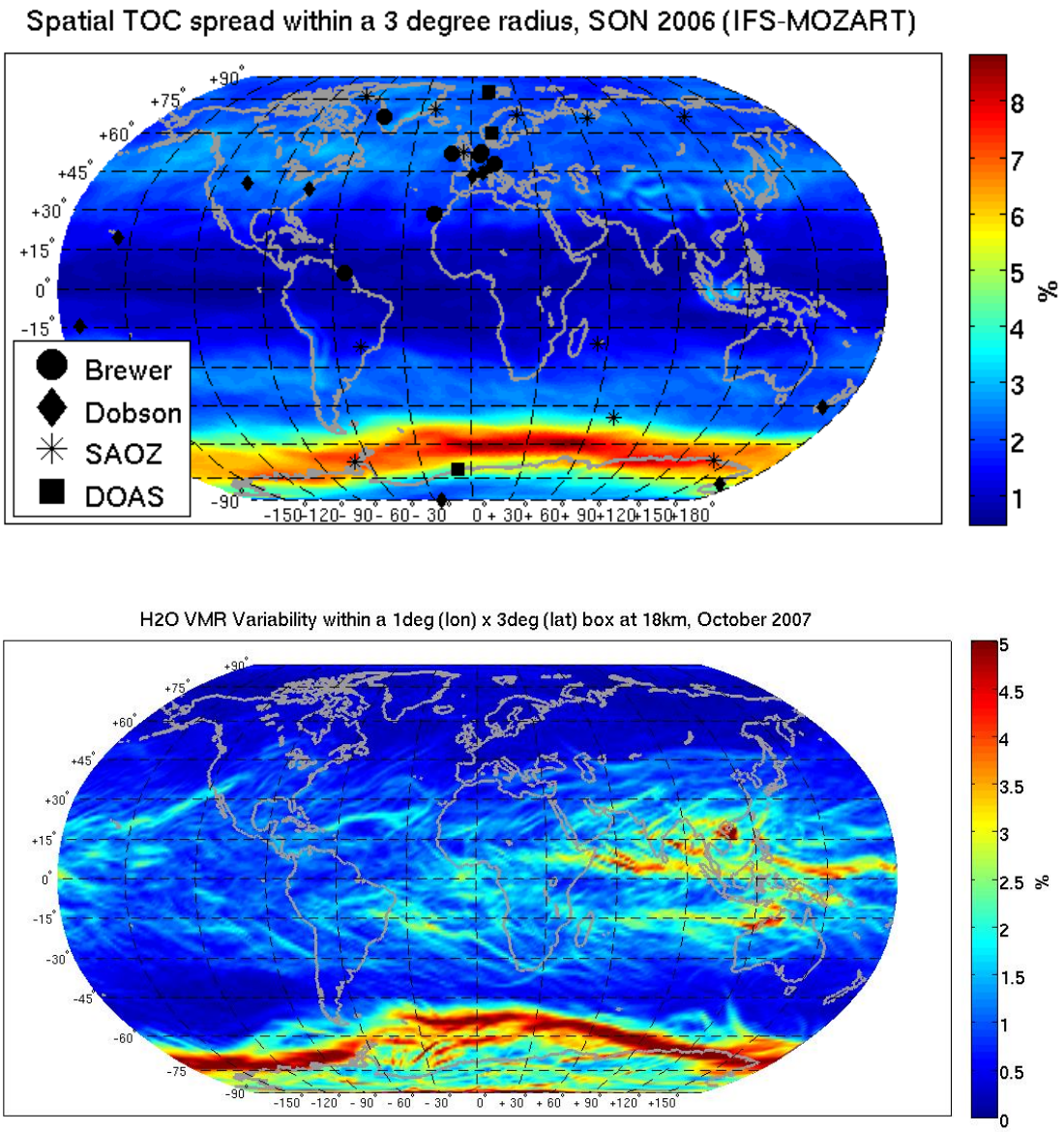


Figure 6: ***Upper panel*****:** Variability of the total ozone column (TOC) field within a 3 degree radius as estimated with an OSSSMOSE simulation using IFS-MOZART model fields. Spatial variability of the TOC is low at equatorial latitudes, non-negligible w.r.t. typical measurement uncertainties at mid-latitudes, and exceptionally large near the polar vortex at the time of the simulation (austral spring, i.e. ozone hole conditions*).* ***Lower panel*****:** similar results for stratospheric water vapour on a scale typical for an IR limb measurement from a polar-orbiting satellite (e.g. MIPAS on ENVISAT).

3.2. Multi-dimensional nature of an atmospheric measurement

The aim here is to introduce in general aspects of the multi-dimensional nature of an atmospheric measurement and available methods to quantify its multi-dimensional extent. Examples of the resulting errors and uncertainties are provided for illustration purposes.

3.2.1. Vertical sampling, resolution and sensitivity

In the vertical dimension, sampling, resolution, and sensitivity are for most instruments well characterized. This section is meant as a brief summary of common practices regarding the characterization of these properties. Moreover, the possibility to quantify vertical smoothing errors is highlighted since such an effort is rarely undertaken, even though it could add valuable information to a data product.

Caveat: Sampling rate ≠ resolution

While vertical sampling rates are often chosen to be similar to the intrinsic resolution of the measurement system, it is important to realize that those are not the same metrological property: the sampling grid only represents the altitudes/pressures to which a measurement is attributed, or retrieved, with no guarantee that one level is independent from another, nor that two subsequent levels represent everything in between, e.g. in the sense of an average. The vertical resolution of the measurement may thus be different from that sampling rate. For instance, ozonesondes usually have a read-out rate of one measurement every 2 seconds, corresponding to a sampling grid step of roughly 10 meter elevation gain when assuming a typical ascent rate of 5 m/s. However, the detector response time is of the order of 20-30 seconds, causing a significant smoothing of the atmospheric profile and corresponding to an actual vertical resolution of only 100-150 meters. Vice versa, in case of a very low sampling rate, part of the profile may remain unobserved. Similarly, retrieved vertical profiles from satellite measurements or ground-based remote sensing (e.g. FTIR, LIDAR) may be sampled at a relatively high number of vertical grid points, while the actual degrees-of-freedom of the retrieval (DOF) is much lower, indicating a much lower vertical resolution, and potentially a large contribution from an a priori constraint in the case of an optimal estimation retrieval. In the case of LIDAR measurements of aerosol properties, the actual vertical resolution is determined by a low-pass filter applied in the data processing to reduce measurement noise, leading to the introduction of the concept of *effective resolution* (Iarlori et al. 2015).

The sampling properties of an atmospheric ECV measurement are usually clearly defined in the data product: the vertical grid is provided in either pressure levels or geopotential heights. The vertical resolution, i.e. the vertical smoothing of the true profile, is sometimes harder to assess. Numbers may not be provided, or, if they are, they are very generic in nature. Nevertheless, some quantification can be done, either from known instrumental characteristics (e.g. the detector response time of an ozonesonde), or from specific retrieval diagnostics in the case of remote sensing data, as detailed below.

The vertical averaging kernel as a key diagnostic

In the case of a remote sensing measurement, a key such diagnostic is the vertical averaging kernel, or AK in short. Formally, the AK is the product of the Jacobians of both the forward model (radiative transfer) and the retrieval or inverse model. In the context of optimal estimation (OE), one can state that the AK represents how the retrieval system smooths or amplifies departures of the true profile from the a priori. Every level of the retrieved profile has an associated AK vector, quantifying the sensitivity to the different levels of the true profile. Together, these AK vectors make up an AK matrix. Examples of AK matrices are shown in Figure 7. Averaging kernels have been evaluated since the early years of remote sensing, in particular in geological remote sensing in the 60s (e.g. Backus and Gilbert, 1968), and since the 70s also in atmospheric remote sensing (e.g. Rodgers, 1976).

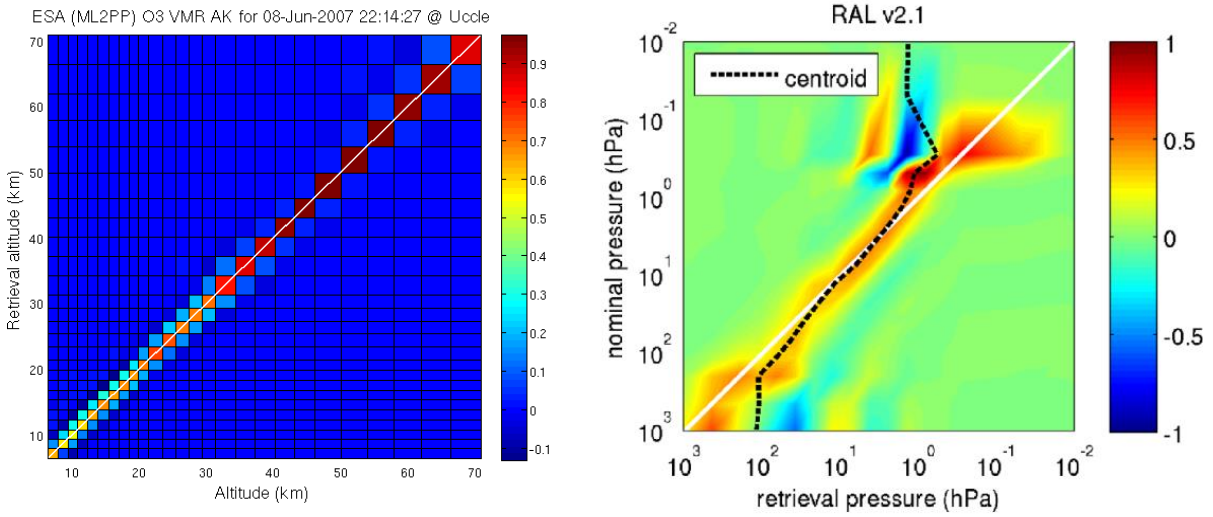


Figure 7: ***Left-hand panel:*** Vertical averaging kernel matrix (AKM) for an ozone profile retrieval from a limb IR emission sounder.. This example corresponds to a well-behaved retrieval with most of the sensitivity on or near to the diagonal, indicating good vertical resolution. ***Right-hand panel:*** Vertical AKM for a nadir ozone profile retrieval from a backscattered UV-Vis sounder. This corresponds to a more challenging retrieval, resulting in sensitivity far from the diagonal.

Rodgers (2000) shows how different quantities derived from the AK characterize the vertical resolution, sampling and sensitivity. These include:

- The **DFS** (Degree of Freedom of the Signal), quantifies the number of independent pieces of information contained within the retrieved atmospheric state vector. It is computed as the trace of the AK.
- **Vertical resolution** can be quantified from the AK in various ways, but the aim is to characterize the spread of the sensitivity around the diagonal for each retrieval level. This measure thus captures to what extent the retrieval at a certain level is sensitive to the true state at lower and higher levels. Typical metrics include full-width-at-half-maximum (FWHM) measures, or the Backus-Gilbert spread.
- A **centroid offset**, computed as the difference between the actual centroid of the sensitivity for a certain level and nominal altitude of that level, indicates whether the retrieval sensitivity is centered on the attributed retrieval level. If this is not the case, one can expect significant measurement bias, depending on the vertical structure of the atmospheric field.
- The **sensitivity** of the retrieval at a given level is defined as the sum of the elements of the AK vector corresponding to that level. It is thus an integrated sensitivity to the entire true profile. If its value is significantly below unity, the measurement is at that level strongly influenced by the regularization, e.g. the a priori profile in case of OE.

An illustration of vertical resolution as estimated from an AK matrix is provided in Figure 8.

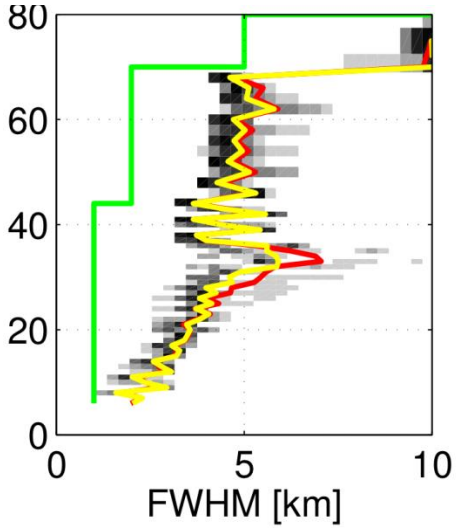


Figure 8: FWHM (yellow line) of the averaging kernels of MIPAS ozone profile retrievals. This FWHM is a proxy of the vertical resolution of the measurement. The green line represents the grid step size, i.e. the sampling rate, which is well below the actual vertical resolution. Reproduced from Laeng et al. (2015).

Quantifying vertical smoothing errors

Nowadays, AKs are widely used to quantify and verify the vertical metrological properties of a remote sensing measurement. Additionally, If a high vertical resolution estimate of the true profile is available, the AK can be used to estimate the errors and uncertainties that the smoothing properties of the measurement system induce: one only has to compare the high-resolution profile with the same profile after application of the AK matrix on the difference with the a priori profile:

$$\varepsilon_{smoothing} = x_a - x_{highres} + A_V (x_{highres} - x_a) , (5)$$

where x_a is the a priori profile, A_V is the vertical averaging kernel, and $x_{highres}$ is high-resolution estimate of the true profile, e.g. from a sonde measurement. However, such a quantification of the vertical smoothing errors and uncertainties is not often undertaken, and for the ECVs dealt with in GAIA-CLIM, no such information is made available in the data files.

Methods to minimize errors due to differences in vertical sampling and smoothing properties when comparing different types of measurements are discussed in Sect. 4.1

3.2.2. Horizontal resolution and sampling

Historically, horizontal resolution properties of atmospheric composition measurements have not been given the same level of attention as the vertical domain. Nevertheless, in the presence of atmospheric inhomogeneities and gradients at a scale smaller or similar to the horizontal resolution of the measurement , horizontal smoothing errors are just as likely to occur as vertical smoothing errors. Pioneering work in the computation and use of horizontal averaging kernels exists (von Clarmann et al. 2009, Ridolfi et al. 2007, Cortesi et al. 2007), see e.g. Figure 9, and more pragmatic estimates of the horizontal resolution of several types of ground-based and satellite instruments have been derived in the EU FP6 project GEOmon (e.g. Lambert et al., 2011), and in the EU FP7 project NORS. Still, such a quantification of horizontal smoothing properties is far from common practice, and the resulting errors are rarely quantified.

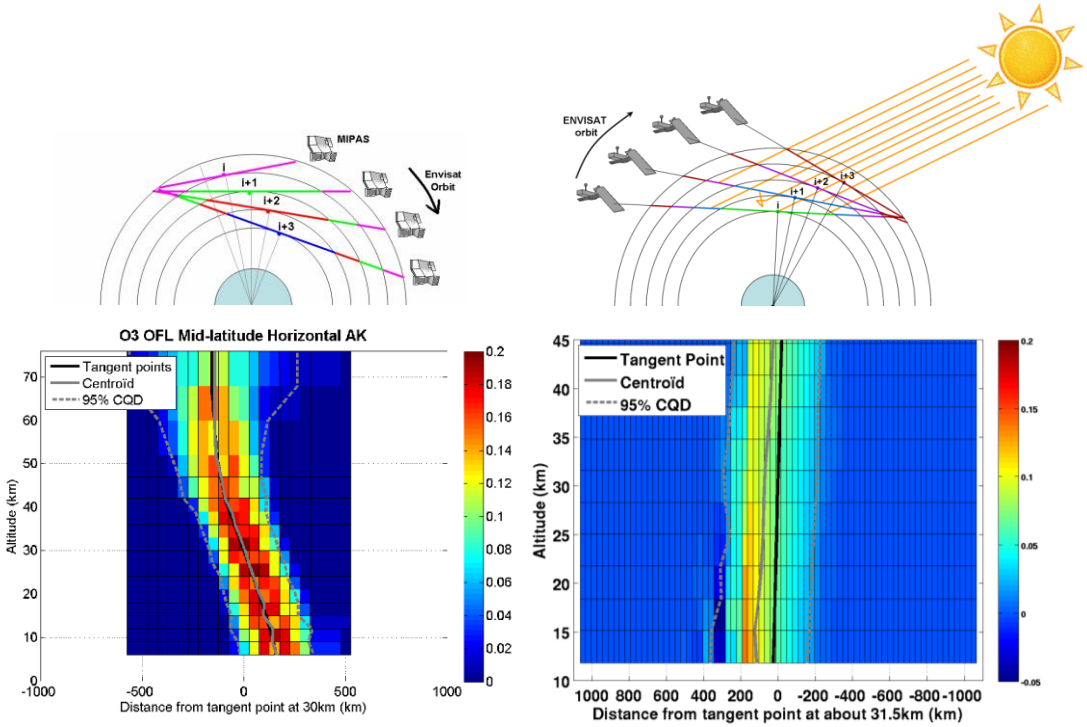


Figure 9: ***Upper row:*** Conceptual illustration of a scanning sequence for a limb IR emission sounder (left) and for a limb UV-Vis scattered light sounder (right). It can be seen that the orbital progression during the measurement sequence causes a shift in tangent point, and in orientation of the light paths. ***Lower row:*** The corresponding horizontal averaging kernels for ozone profile retrievals. From von Clarmann et al. (2009) and Vandenbussche et al. (2010).

A key point is that *horizontal resolution* not only depends on obvious parameters such as for instance the pixel footprint of a nadir satellite observation, but on the entire observation geometry, on the retrieval scheme in the case of remote sensing, and on potential data post-processing such as measurement averaging. A pertinent case illustrating the importance of a metrological evaluation of the entire measurement and retrieval chain when estimating horizontal smoothing properties is that of zenith-sky observations of scattered UV-Visible light during low-sun conditions. While the instrument is pointed at zenith, the bulk of the absorption processes determining the signal from which the trace gas concentrations are retrieved occurs along the line-of-sight between the sun and the scatterer, and consequently covers a large horizontal extent in the twilight setting. This is illustrated in Figure 10.

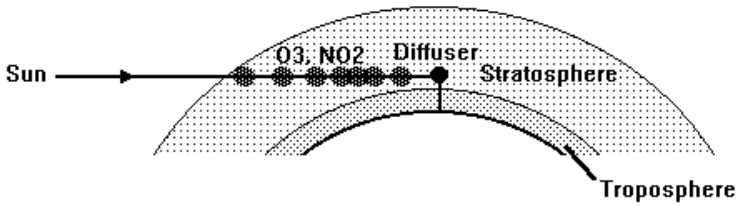


Figure 10: Twilight observations of zenith-scattered light are sensitive to trace gas concentrations along a large horizontal line-of-sight.

Another case of "hidden" horizontal smoothing is that of observations relying on integration times which allow large airmasses to be blown through the field-of-view by the local wind field. This is for instance the case for LIDAR observations of stratospheric ozone, which can require up to 4 hours of integration time, corresponding to several hundred km of horizontal smoothing assuming a wind speed of the order of 100km/h at stratospheric altitudes.

The resulting *horizontal smoothing errors* are particularly large when the measurement is sensitive over a large horizontal area, and the atmosphere is highly variable across this region. This is the case for instance for solar occultation measurements of the O_3 volume mixing ratio in the UT/LS, as illustrated with an OSSSMOSE simulation (cfr. Section 5.4) in Figure 11.

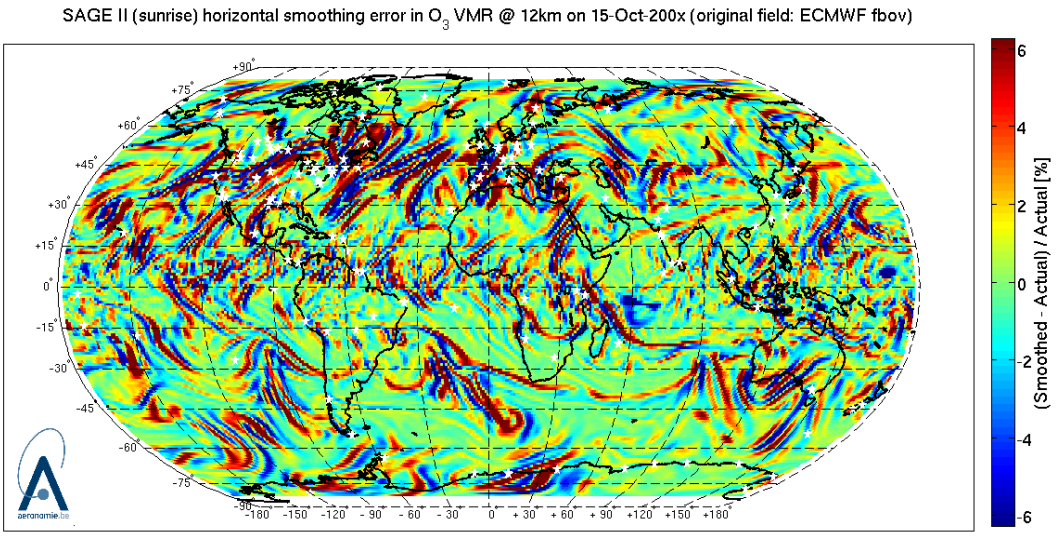


Figure 11: OSSSMOSE simulated horizontal smoothing errors in O_3 VMR measurements with SAGE II at sunrise, assuming global coverage, and based on ECMWF reanalysis fields and a pragmatic observation operator for SAGE II measurements derived from an analysis of horizontal averaging kernels (Vandenbussche et al. 2010).

Based on a similar OSSSMOSE simulation, **Figure 12** presents a time series of horizontal smoothing errors in ozone volume mixing ratio measurements retrieved from limb IR emission observed with MIPAS on ENVISAT above the Antarctic station of Dumont d'Urville. A clear feature are the large smoothing errors in austral spring, which are a consequence of the large variability in TOC at the edge of the polar vortex, as illustrated already in the lower panel of **Figure 6**.

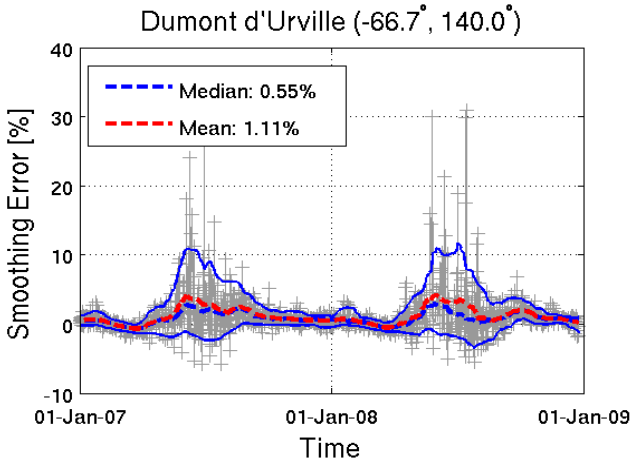


Figure 12: Simulated horizontal smoothing errors for MIPAS/ENVISAT measurements of ozone volume mixing ratio (VMR) at 45km altitude above Dumont d'Urville (Antarctica). The smoothing error is calculated here as the relative difference between the horizontally smoothed O3 VMR and the point-like value at the nominal location of a MIPAS profile, i.e. the location of the tangent point at 30km altitude. The dashed lines represent the 1-month running mean and median smoothing error. The solid lines represent running 1-month 16% and 84% quantiles. The smoothing errors are seen to be extremely large when a strong polar vortex is present and ozone is depleted, i.e. in Austral spring. Moreover, the non-zero mean (and median) indicate that also a systematic error is introduced. The simulation was performed with the OSSSMOSE system, described in more detail in Section 5.4.

Also the *horizontal sampling* properties depend strongly on the measurement type and platform. The sampling can be relatively complete and at high spatial frequency, e.g. from nadir satellite sounders with small pixel sizes and large swaths, and consequently short revisit times, such as the upcoming TROPOMI instrument on Sentinel-4p. On the other hand, for (stellar) occultation and limb-emission measurements, the sampling is usually much sparser and often highly inhomogeneous. Also for most ground-based instruments, both the network sampling and the sampling properties of a single measurement are limited and inhomogeneous. This is the case for instance for balloon-borne sondes. Balloon drift over an entire ascent, which takes roughly 1.5 hours, depends on wind strength and direction, but >100km is not uncommon. A detailed analysis of radiosonde drift statistics on the global level was performed by Seidel et al. (2011), see the left-hand panel of Figure 13 for an illustration. Because the wind has a preferred direction at many sonde-launching stations, the sampling of a sonde dataset is often not homogeneous around the station location (right-hand panel of Figure 13). Most atmospheric variables measured by balloon sondes vary significantly on these scales, and balloon drift will therefore lead to non-negligible differences w.r.t. the launch site.

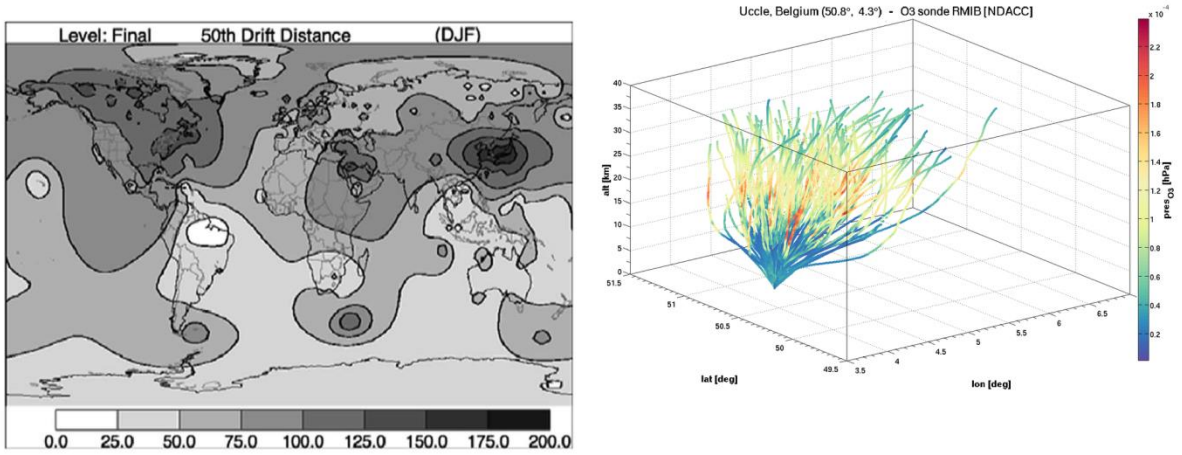


Figure 13: ***Left-hand panel***: Contour plot of balloon drift distance for the global radiosonde network, for December, January, and February, as derived by Seidel et al. (2011). ***Right-hand panel***: Balloon trajectories are not distributed homogeneously around the NDACC station of Uccle, B., due to predominantly Westerly winds (Figure courtesy of D. Hubert, BIRA-IASB).

3.2.3. Temporal sampling and resolution

While the integration times of measurements of meteorological variables are usually short enough to consider the atmosphere constant during the measurement, this is not always the case when observing trace gas concentrations. For instance, LIDAR measurements of stratospheric ozone require several hours of integration time, which is long enough for the measurement to be affected by variations in the ozone profile due to transport. Similarly, zenith-sky observations of NO_2 columns during twilight are affected by the rapid photolysis (or formation) of NO_2 with increasing (decreasing) solar radiation during the measurement sequence.

Besides long integration times, measurements are sometimes averaged into e.g. daily means to improve the signal-to-noise ratio and to increase the spatial representativeness of the measurement, which can be beneficial when comparing to satellite measurements with large ground

footprints. Still, this also constitutes a form of temporal smoothing and one must be aware that processes other than transport can potentially alter the atmospheric field, e.g. photo-chemistry or local sources and sinks. This approach of using daily means is for instance common in total ozone column measurements with direct-sun instruments such as Dobsons and Brewers.

Finally, one must be aware of the possibility that large horizontal smoothing introduces a hidden form of temporal smoothing through local solar time variation over the area of measurement sensitivity. The solar zenith angle therefore changes over the measurement footprint, and this must be taken into account when studying ECVs with a pronounced diurnal cycle, whether the latter is due to photo-chemistry of other diurnal phenomena such as condensation and evaporation.

4. Metrology, part III: Data comparisons

In this section, the characterization of individual measurements as described in the previous sections is integrated into the metrology of a comparison between satellite and ground-based reference measurements. Returning to Eq. (4) in Section 1.3, Section 2 provides the random and systematic errors related to the measurement uncertainties (ε_{rand} and ε_{sys}) and the concepts explored in Section 3 allow a quantification of the smoothing difference errors (ε_{S4D}). In the current section, the total error budget is explored by also including the sampling difference errors ($\varepsilon_{dx/d4D}$), which depend on the way the co-location is performed.

It is well outside the scope of this document to cover all possible combinations of satellite and ground-based sounders, for all possible atmospheric variables, and for all methods of co-locating measurements. Rather, one case study is elaborated in Section 4.3, illustrating the metrology of a comparison and the different error terms that have to be quantified in order to close the uncertainty budget and to derive meaningful conclusions on the agreement between satellite and reference measurements. Two essential topics are touched upon before reporting on that case study: (1) the methods that are available to associate satellite with reference measurements, i.e. ways of generating co-located measurement pairs, and (2) harmonization of data sets in terms of vertical resolution and sensitivity. While mismatches in vertical resolution are in principle a part of the more general spatiotemporal mismatches, they are dealt with separately in the validation process because methods exist to harmonize the data to a common vertical resolution. This is possible in the vertical dimension because most measurement techniques offer a continuous vertical sampling within their measurement range, i.e. there are no unobserved gaps between subsequent altitude levels. This allows the use of regridding techniques without introducing large interpolation uncertainties. That is not the case in the horizontal and temporal dimensions, which are characterized by poor sampling of the atmospheric signals: the distance between two stations is much larger, and the revisit time of the satellite sounder much longer, than the small scales of variability of the atmosphere. Consequently, horizontal or temporal interpolation is not advised, and the errors and uncertainties due to co-location mismatch must be quantified a posteriori, e.g. using auxiliary data.

4.1. Co-location criteria

Several methods for co-locating satellite measurements with ground-based reference measurements exist, ranging from basic constraints on geometrical distance and time, to more elaborate criteria including information on atmospheric dynamics. Which method is most appropriate depends on a multitude of factors, including the concerned variable and the application. For instance, near-real-time (NRT) validation services of a stratospheric variable with low spatiotemporal variability can be performed with basic geometric co-location methods, while a more advanced assessment of a tropospheric variable, susceptible to local sources and sinks, may require detailed airmass matching including back-trajectory calculations and a full estimation of the actual area of sensitivity of the measurement.

While it is outside of the scope of the current document to cover an extensive review of all methods with their (dis)advantages, several often-used methods are nevertheless listed below, without the ambition of exhaustive. We refer to the annexes of D3.1 for a literature review covering applications of most of these criteria.

- **Great-circle distance and time**: d < D and t < T, where d and t are usually the differences between the nominal locations and measurement times. The constraints D and T depend heavily on the variable under study, ranging for instance from a few km and minutes (humidity validation) to 1000km and 12h (stratospheric CH_4 validation).
- **Latitude, longitude, and time constraints**: many atmospheric variables display less variability along parallels than they do along meridians. Separate constraints on latitude and longitude differences allow to take this into account.
- **Static airmass matching**: For several types of measurement, pragmatic observation operators have been constructed which quantify the actual area of measurement sensitivity as e.g. a multi-dimensional polygon (see e.g. Lambert et al., 2011). From these polygons, several optimized co-location can be derived, formalizing the requirement that both measurements must to some extent sample the same airmass. For instance, one can require a certain intersection between both polygons. This is illustrated in **Figure 15**.
- **Potential vorticity**: A constraint can be placed on the difference in potential vorticity between satellite and reference measurement, *PV*. This is a constraint on the dynamical properties of the air parcels, ensuring that they have similar origins and hence similar trace gas concentrations.
- **Back trajectories**: A more elaborated way of taking into account the dynamical history of air parcels is by calculating back trajectories using wind fields from, e.g., meteorological analyses. This technique is particularly useful when dealing with large time differences between both measurements, which are sometimes unavoidable, for instance when validating a day-time satellite overpass with a night-only reference measurement (e.g. from a LIDAR). Note that this technique assumes that the tracked molecule or parameter acts as a pure dynamical tracer.

After applying these constraints, one may either have no co-locations, in which case a more relaxed criterion is adopted out of necessity, or multiple pairs satisfying the criterion in which case a "closest" match is usually extracted, if necessary through a weighted combination of the different criteria. For instance, distance and time can be combined into an "effective" distance by assuming a certain wind velocity.
In practice, even the more sophisticated criteria are usually not sufficient to fully avoid additional errors in the comparison due to co-location mismatch. This is illustrated in Section 4.3.

4.2. Differences in vertical resolution and sensitivity

The issue of comparing profile measurements with different vertical resolution has been a topic of investigation already since the early days of remote sensing, and the use of vertical averaging kernels to study vertical resolution properties of retrieved profiles was already pioneered in the 1970s (e.g. Rodgers, 1976). The possibility to estimate vertical smoothing errors using the averaging kernels was presented in Eq. (5) in Section 3.2.1. Moreover, methods exist to regrid data to a common vertical sampling, taking into account differences in vertical resolution (Rodgers and Connor, 2003; Calisesi et al. 2005). In the context of validation with ground-based reference measurements, this usually implies degrading the resolution of the reference profile to that of the profile retrieved from the satellite observations. The profile at reduced resolution is computed as:

$$x_{2,lowres} = x_a + A_{V1}(x_2 - x_a), \quad (6)$$

where x_1 is the low resolution (retrieved) profile, x_2 is the high-resolution reference profile, x_a is the a priori used in the retrieval of x_1, and A_{V1} the corresponding averaging kernel.

As described above, these methods rely on the availability (and accuracy) of vertical averaging kernels, but they can be extended to a more general method using generic functions with a similar bandwidth, such as a Gaussian with a FWHM corresponding to the vertical resolution. In such an application the vertical resolution may be estimated from means other than the AKs. In practice, these procedures require care on several technical aspects, described in detail in Keppens et al. (2015). Figure 14 illustrates the regridding, including smoothing to a lower resolution, of a high-resolution ozonesonde measurement. This approach of bringing satellite measurement and reference measurement to the same vertical grid and resolution is common place now in most validation exercises.

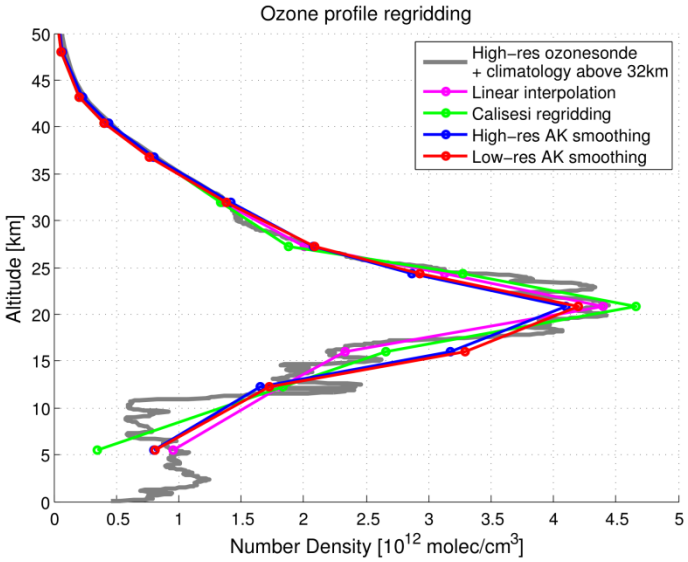


Figure 14: An illustration of different smoothing and regridding approaches applied to a high-resolution ozonesonde profile (extended with a climatology above 32km). They range from a simple linear interpolation, not taking into account the smoothing properties of the system that is emulated (here a GOME-2 nadir ozone profile retrieval), to a more advanced approach including the use of the appropriate GOME-2 vertical averaging kernel. Figure courtesy of S. Compernolle (BIRA-IASB), following recipes provided by Keppens et al. (2015).

4.3. Case study

To illustrate the role of the different error terms in Eq. 4, and their impact on the total uncertainty budget, a case study is developed here which focusses on total ozone column (TOC) comparisons between on the one hand ground-based measurements obtained with the (SAOZ) ZLS-DOAS instrument at the NDACC station of Dumont d'Urville on Antarctica, and on the other hand GOME/ERS-2 nadir satellite measurements. These comparisons were originally performed for a validation exercise in the context of ESA's Ozone CCI (Koukouli et al., 2015), but without a detailed assessment of the comparison metrology and related co-location errors. Verhoelst et al. (2015)

revisit these comparisons and use the OSSSMOSE system (Section 5.4) to simulate the measurement comparisons and quantify the impact of the different co-location mismatch errors.

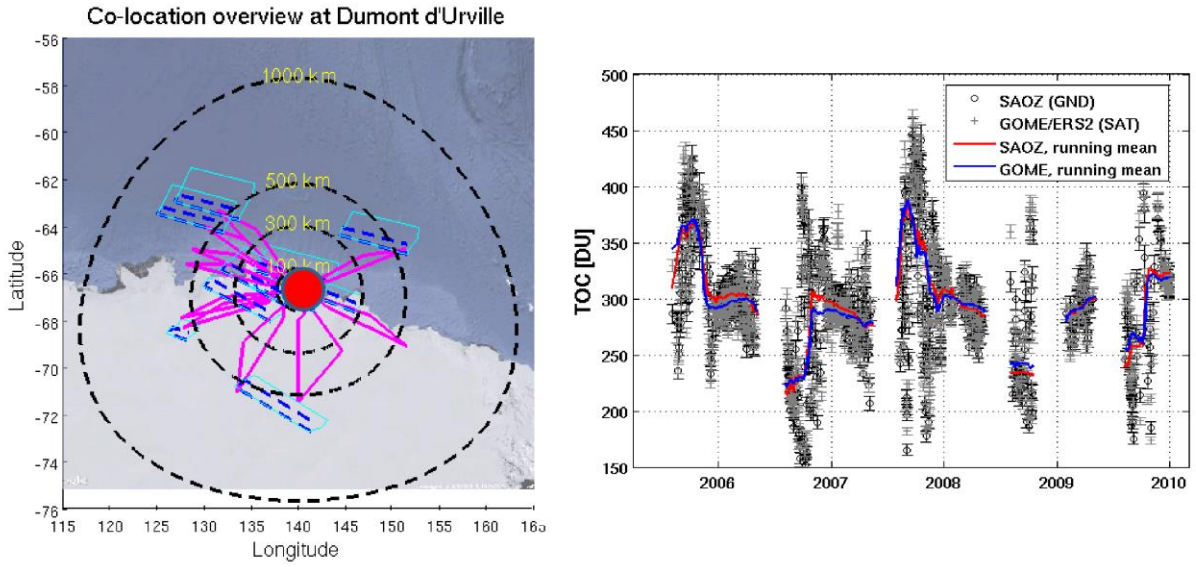


Figure 15: ***Left hand panel****:* Illustration of a few co-located total ozone column (TOC) measurement pairs made up of (1) ground-based ZLS-DOAS measurements obtained with the CNRS/LATMOS SAOZ instrument at the NDACC station of Dumont d'Urville in Antarctica (red disk) and represented by the magenta polygons which describe the actual area of measurement sensitivity, and (2) GOME/ERS-2 nadir UV-Vis measurements represented by their ground pixels (blue dashed lines) and the additional sensitivity towards the sun (cyan). The co-location criterion used here required a non-zero intersection between the measured airmasses. ***Right-hand panel:*** Time series of TOC measurements from these co-located measurement pairs, including 1-month running means.

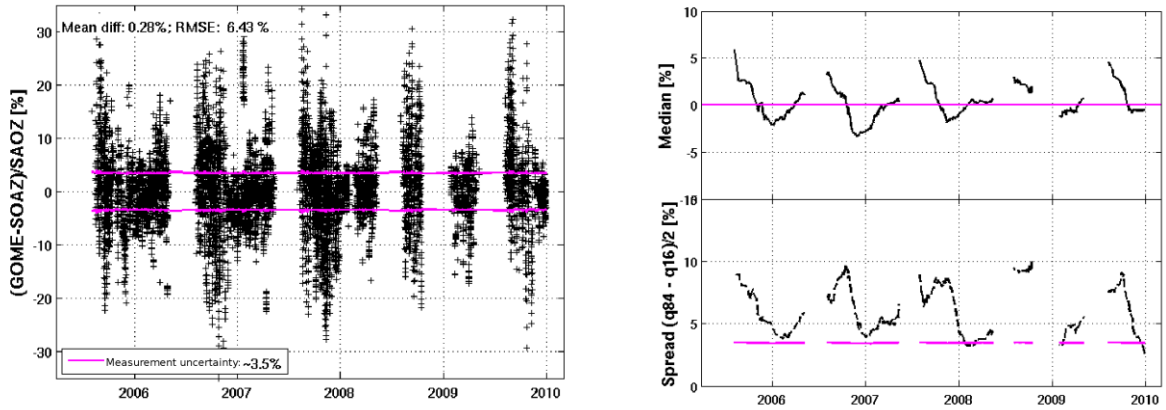


Figure 16: ***Left-hand panel:*** Time series of the relative differences between the measurements co-located in Figure 15. The RMSE of these differences is almost twice as large as the combined measurement uncertainty (magenta lines). ***Right-hand panel:*** 3-month running median (upper panel) and spread (lower panel) of the differences shown in the left-hand panel. The median difference is used here as a proxy for systematic errors, while the spread captures those errors that are predominantly random in nature. These curves reveal clear seasonality in the systematic and random components, well above what can be accounted for by the combined measurement uncertainty (magenta line). It will be shown here that this is not necessarily due to shortcomings in the data or their uncertainties, but that these discrepancies mostly result from co-location mismatch.

Figure 15 presents a visualization of a small subset of the co-locations, and the time series of 5 years of co-locations. From these time series, it is obvious that the TOC field is highly variable at this station, which is often close to or within the ozone hole (less than 220DU by definition), depending on the exact shape of the polar vortex.

Figure 16 reveals that the differences between co-located measurements are well above what could be expected from the combined measurement uncertainties, and that these show clear seasonal structures, both in their median value over several months and in the spread. Looking at the separation of co-located pairs, of the order of several 100km, and keeping in mind the spatial variability of the TOC field near the polar vortex quantified in the upper panel of **Figure 6**, one can expect large co-location errors and any conclusion on data quality would at this point be premature. Indeed, Verhoelst et al. (2015) simulated these comparisons with the OSSSMOSE system, which set-up here to apply appropriate observation operators for each measurement onto high-resolution TOC reanalysis fields to simulate each individual measurement (see also Section 5.4 for more details). These simulations can be done with the different error terms switched on or off, e.g. one can assume negligible measurement uncertainty, which produces differences solely due to co-location mismatch. Or, one can assume the measurements to be point-like and the comparisons affected solely by sampling mismatch, i.e. the nominal coordinates do not coincide. With all error terms from Eq. (4) included, the aim is to reproduce the curves observed in the right-hand panel of **Figure 16**. **Figure 17** presents the results for this particular comparison exercise.

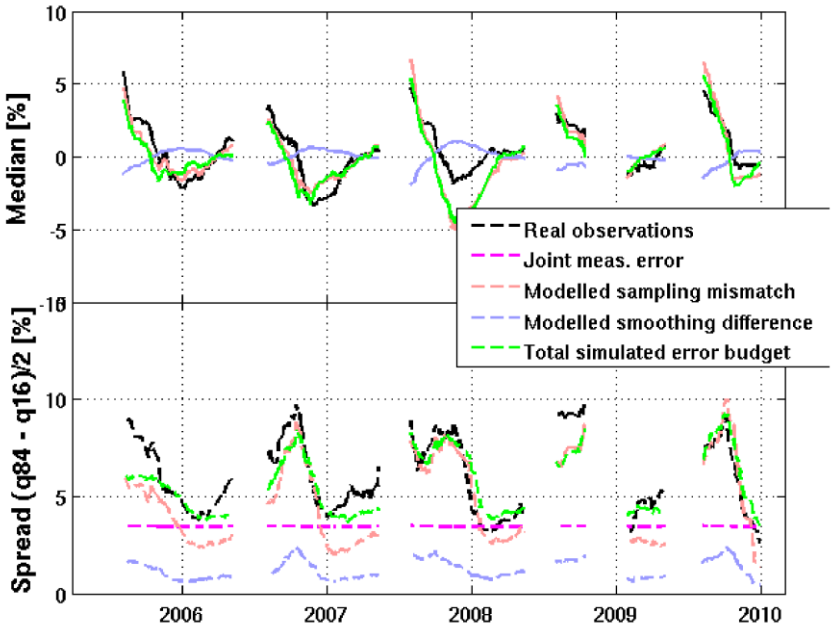


Figure 17: Similar to the right-hand panel of Figure 16, but with the results from the OSSSMOSE comparison metrology simulator overlaid as colored lines. The blue lines represent the simulated errors due to differences in horizontal resolution, but not taking into account the offset (in space and time) in actual measurement locations. The red curves represent the errors due to sampling differences, i.e. the differences in measurement location and time, but not taking into account that the measurements also have different horizontal resolutions. The green curves include all error terms: smoothing differences, sampling differences and measurement errors (the latter simulated as random draws from a Normal distribution determined by the measurement uncertainty provided with the data)., The simulation matches the observed differences remarkably well, indicating that the data and their uncertainties are in fact reliable, and that the large observed differences are almost entirely due to co-location mismatch.

The good agreement between the total simulated error budget (in green) and the statistics of the observed differences (black) indicates that all error sources are accounted for. Indeed, it appears

that the sampling difference error (red) plays a major part in the total error budget, both in the random and systematic components. Smoothing difference errors up to a few percent occur as well, but they do not exceed the combined measurement uncertainty. Figure 18 contains the full difference probability distribution function (PDF), offering a more comprehensive view than the median and interquantiles presented in **Figure 17**. The good agreement between the observed and simulated difference PDFs is further evidence that all error sources are properly represented in the simulation. The conclusion is therefore that the large observed differences between satellite and ground-based measurements, which exceed the combined measurement uncertainty, do not indicate any issues with data quality or reported uncertainties: they can be explained entirely as due to spatiotemporal mismatch, in spite of the fairly sophisticated co-location criterion which required intersecting area's of measurement sensitivity.

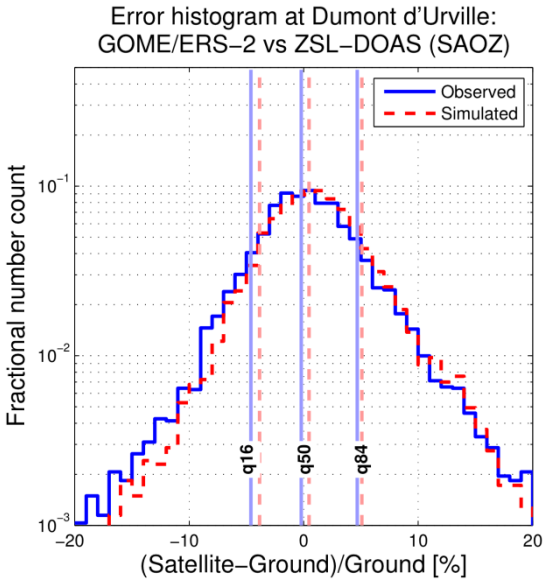


Figure 18: Observed and OSSSMOSE-simulated probability distribution function of the differences between co-located GOME/ERS-2 and CNRS/LATMOS SAOZ measurements at the NDACC station of Dumont d'Urville. The good agreement indicates that the simulation has taken into account all significant error terms.

Further results on total ozone column validation, including several different satellite and ground-based instruments, are reported in Verhoelst et al. (2015), and the extension of this approach to other ECVs is one of the main aims of WP3 in GAIA-CLIM.

5. Methods to quantify spatiotemporal mismatch uncertainties

The sparseness of the ground-based reference data and/or of the satellite data set, implies that co-location criteria can often not be taken strict enough to avoid additional errors and uncertainties in the comparisons due to differences in sampling and smoothing. In such a case, *a posteriori* estimates of these uncertainty budget terms can still be made, with a variety of methods. Those methods related to work performed within GAIA-CLIM are outlined below. They range from very much data driven to explicit modelling of the physics and measurement metrology. Figure 19 summarizes these different methods.

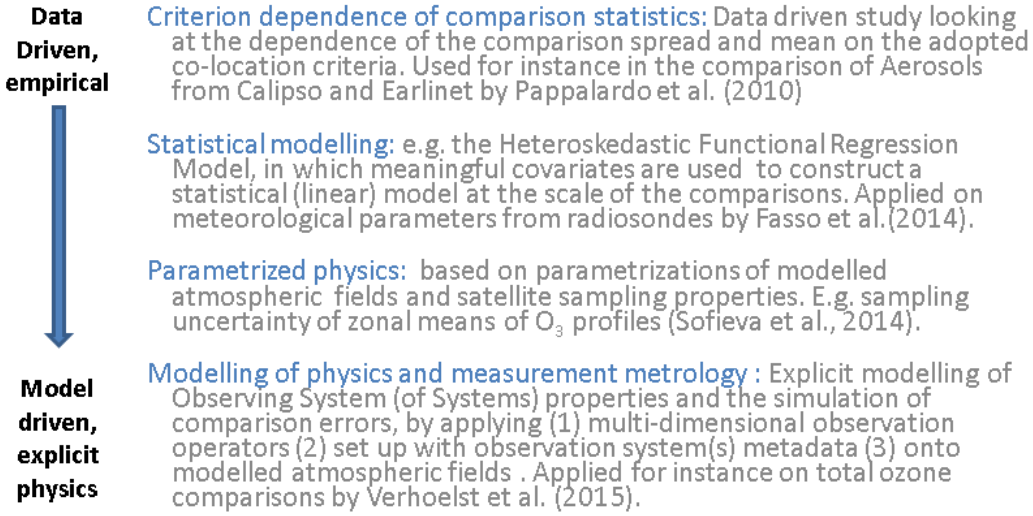


Figure 19: Overview of different methods for a posteriori estimates of co-location mismatch errors and/or uncertainties.

5.1. Criterion dependence of comparison statistics

If a sufficient number of comparison pairs are available, it is possible to study the effect of relaxing or tightening of the co-location criterion on the comparison statistics. The latter can be diagnostics such as the comparison spread (variance or standard deviation), the mean (or median), or a correlation coefficient. Figure 20 illustrates this approach for the correlation between aerosol backscatter coefficient counts from Calipso and EARLINET, as analyzed by Pappalardo et al. (2010). For this graph, only simultaneous measurements (i.e. within 10 minutes of each other) were used, but a large range of allowed horizontal separations was explored. Already at a separation of a few hundred kilometer, the correlation is much smaller than for the smallest separation for which a meaningful correlation could be computed (D <100km), indicating strong inhomogeneity in the aerosol field at these scales.

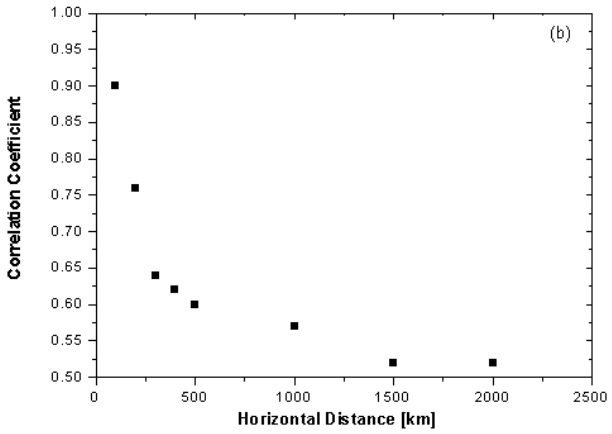


Figure 20: Correlation between Calipso and EARLINET backscatter coefficient counts as a function of horizontal separation. Reproduced from Pappalardo et al. (2010), their Fig. 15.

Likewise, Figure 21 illustrates the dependence of two diagnostics of an intercomparison of total ozone column measurements on the maximum allowed horizontal separation between the satellite observation and the ground-based reference measurement. The spread on the differences is seen to increase linearly with increased maximum horizontal separation, and already exceeds the combined measurement uncertainty when allowing separations up to 250km. Even the median of the differences is found to depend on the adopted co-location criterion, contradicting the often-made assumption that atmospheric variability "*averages out*" if sufficient co-locations are available. It is clear that in this case the particular sampling characteristics of the measurement systems combined with structural atmospheric gradients lead to systematic errors in the comparisons that depend on the adopted co-location criterion.

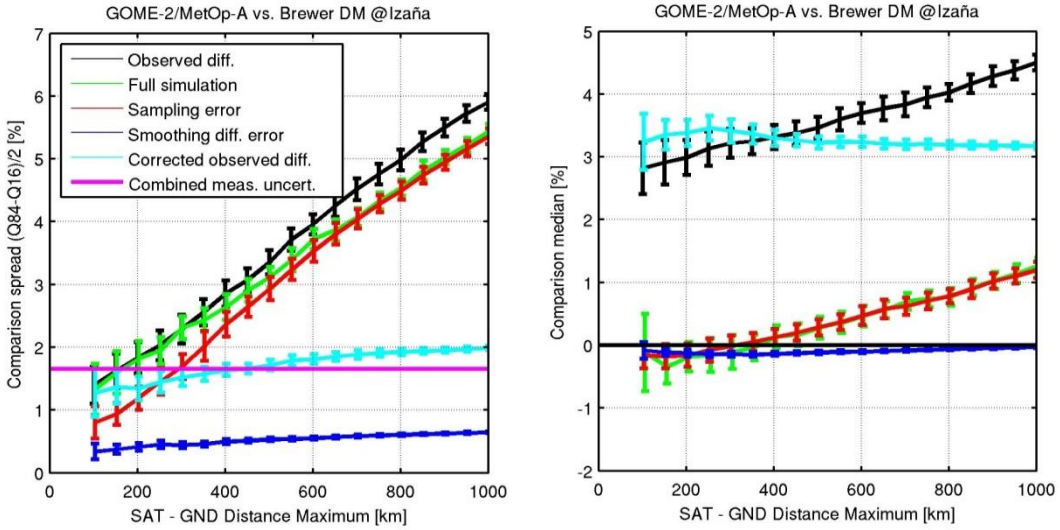


Figure 21: Dependence of the spread (left-hand panel) and median (right-hand panel) of the differences between GOME-2/MetOp-A and AEMet Brewer total ozone column measurements at the Izana station (Tenerife Island). In the current context, only the black and magenta solid lines are of interest. The other lines correspond to an OSSE approach detailed further in Section 5.4. Reproduced from Verhoelst et al. (2015).

The advantage of this method is that it is entirely data-driven and does not rely on a quantified understanding of the underlying physics. As such, it is applicable to species and atmospheric regimes for which little is known about the scales and amplitudes of spatio-temporal variability. It is

also fairly straightforward to implement and low on computation cost. The disadvantage is that it requires fairly large numbers of co-locations to be able to have reliable statistics even for tight co-location criteria.

5.2. Statistical modelling

A more advanced data-driven analysis of a set of co-located measurements uses the Heteroskedastic Functional Regression Model (HFRM) to disentangle environmental (co-location mismatch) and measurement uncertainties. It was demonstrated by Fassò et al. (2014) on thermodynamic profiles obtained from radiosondes launched from two nearby GRUAN sites. The basic idea is that the differences between co-located measurements are modelled using a locally linear function of measured covariates, and allowing for measurement and height-dependent environmental uncertainties. In this way, the model can differentiate between measurement uncertainty, reducible environmental uncertainty and irreducible environmental uncertainty, see Figure 22.

A clear added value of the modelling component is the possibility to decompose the uncertainty budget and consequently to assess the accuracy of the measurement uncertainties provided with a data product. Drawbacks of this approach are the necessity of meaningful covariate measurements, the limitation to uncertainties, rather than individual errors, and the coding effort required to implement the model fitting.

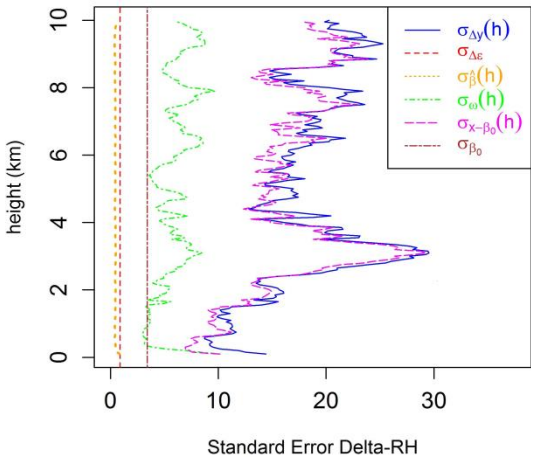


Table 1. Budget of total collocation uncertainty for relative humidity (Δrh) in Beltsville–Sterling, using the HFR model of Eqs. (14) and (15).

Source of uncertainty		$\bar{\sigma}^2$	$\bar{\sigma}^2$%	$\bar{\sigma}$
Total uncertainty	Δ_y	343.8	100 %	18.54
Collocation drift	Δ_μ	343.0	–	18.52
Bias (adjustable)	β_0	11.6	3.4 %	3.40
Environ. error (reducible)	$x \sim \beta_0$	293.2	85.5 %	17.12
Environ. error (irreducible)	ω	37.1	10.8 %	6.09
Sampling error	$\hat{\beta}$	0.21	0.1 %	0.46
Measurement error	Δ_ε	0.81	0.2 %	0.90

Figure 22: Uncertainty budget decomposition of relative humidity comparisons between co-located radiosonde launches, as derived using the HFRM statistical approach. From Fassò et al. (2014).

5.3. Parametrized physics

Another potential approach combines a parametrization of natural variability and of measurement system properties. For instance, with the aim of quantifying the sampling uncertainty σ_{sample} of gridded monthly means of ozone profiles from the HARMonized data set of Ozone profiles (HARMOZ), created in the framework of the ESA Ozone_cci project, Sofieva et al. (2014) developed the following approach: First, the measurement inhomogeneity H is defined and computed as a combination of the measurement asymmetry A and entropy E (see Sofieva et al.,2014, for the definition of these properties):

$$H = ½(A + (1 – E))$$

For the HARMOZ data set, the inhomogeneity in longitude is negligible, and only the latitude and time dimension must be taken into account. Then, the natural variability is parametrized as a climatological variance σ_{var}. Lastly, the coefficient α in the equation $\sigma_{sample} = (\sigma_{var} . H)^{\alpha}$ is determined by fitting this relation to a more detailed determination of the sampling error with an OSSE approach (cfr. the next section). From Figure 23 it was concluded that α is approximately equal to unity, and thus:

$$\sigma_{sample} = \sigma_{var} . H$$

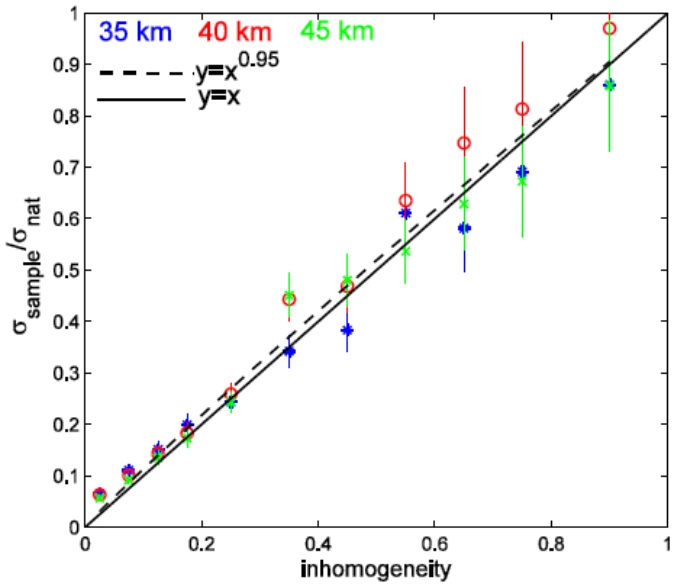


Figure 23: Relation between measurement inhomogeneity and sampling uncertainty normalized to the natural variability, for the HARMOZ ozone profile data set developed within Ozone_cci. The data points represent detailed estimates of the sampling uncertainty based on an OSSE approach, while the solid and dashed line represent a parametrized relation. From Sofieva et al. (2014).

This sampling uncertainty should then be added to the formal uncertainty on the monthly ozone profile averages to yield the total uncertainty on the gridded monthly means:

$$\sigma_{tot}^2 = \sigma_{mean}^2 + \sigma_{samp}^2$$

Such a parametrization allows for a fast estimate of sampling uncertainties of a large set of high-level data. The downside is that in the current shape, the output of this approach is limited to uncertainties, i.e. the width of the error distribution, and does not allow an estimate of individual errors and of their mean (i.e. a systematic error).

5.4. Physical modelling

The most informative, but also the most elaborate, option is the Observing System Simulation Experiment. This concept can be applied here as follows: first, it consists in the creation of (multi-dimensional) observation operators constrained by the metadata of the real observing systems, then, this is followed by the application of those observation operators onto high-resolution atmospheric fields, either from models or from gridded measurement data sets and campaigns. Provided that both the observation operators and the fields are realistic, this method allows a simulation of each individual measurement, with or without smoothing properties taken into

account, and as such it allows a quantified estimate of the error terms due to smoothing and sampling differences, and of the combined mismatch errors. This approach is followed for instance in Verhoelst et al. (2015), based on a suite of software tools named OSSSMOSE, and used to generate illustrations of smoothing and sampling issues throughout Sections 3 and 4 of this document. The general setup of the OSSSMOSE system is shown in Figure 24 , but the reader is referred to Verhoelst et al. (2015) for more details.

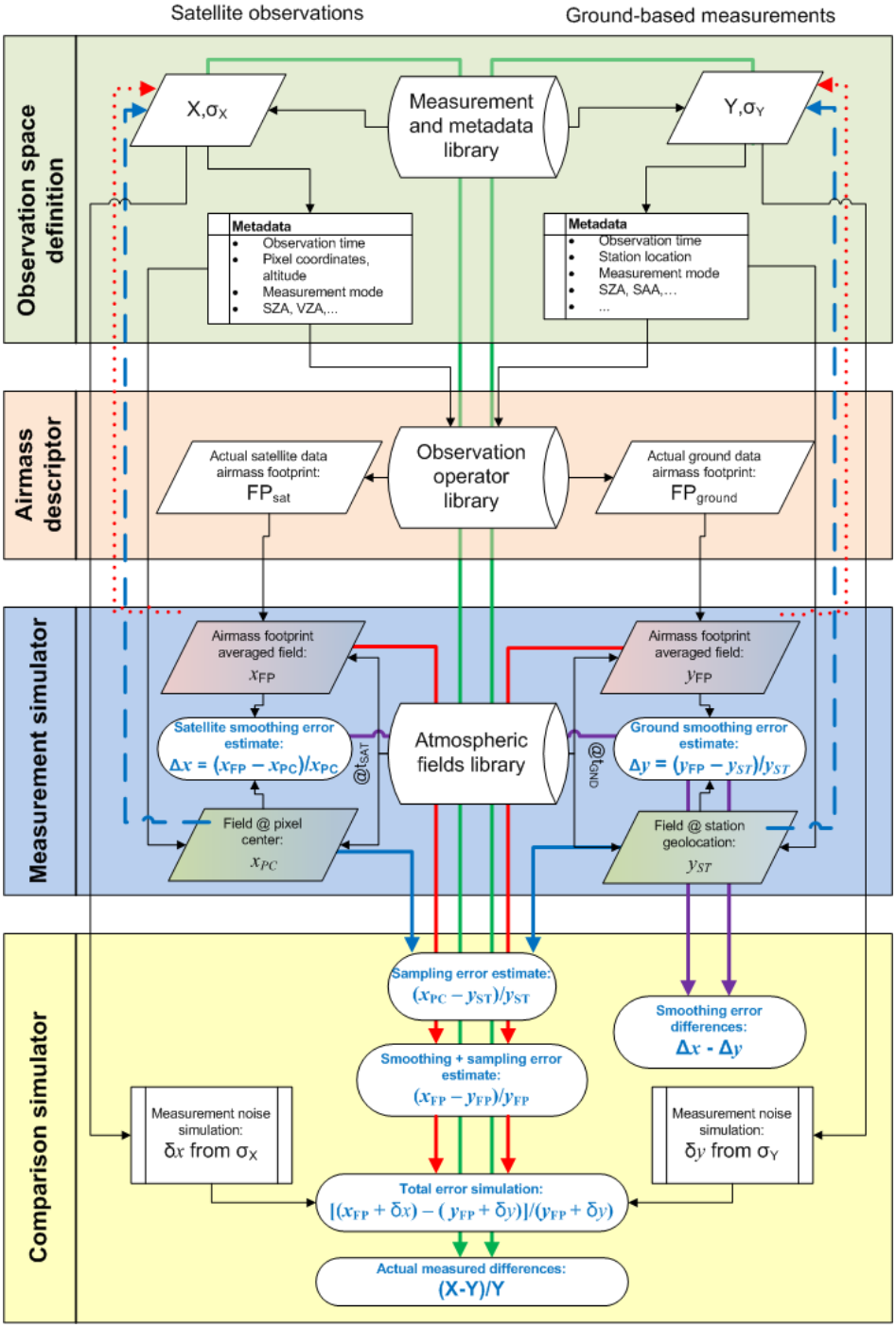


Figure 24: Architecture of the OSSSMOSE system for Observing System Simulation Experiments. See Verhoelst et al. (2015) for further details.

The advantage of this approach is the possibility to estimate individual errors, and hence to analyze the entire error probability distribution function, rather than only its width (i.e. the uncertainty). On

the downside, it requires reliable auxiliary data (model or other high-resolution gridded data) and some computational effort. This system was used for the case study reported on in Section 4.3, and an additional example is provided in Figure 25. Again, the large comparison spread and median, with very strong seasonal features, do not indicate data quality issues but can be traced back entirely to spatio-temporal mismatch errors.

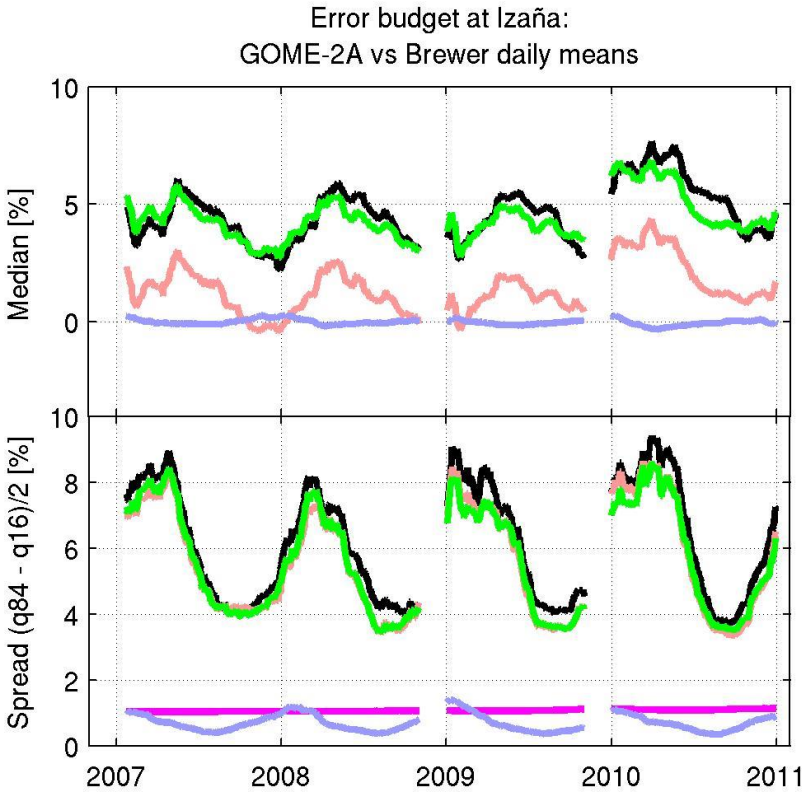


Figure 25: Median of (upper panel) and spread on (lower panel) the differences between GOME-2A and AEMet Brewer (daily mean) measurements of the total ozone column above the NDACC sub-tropical station of Izaña (Tenerife Island). The black curves correspond to the actual measured differences for co-locations up to 1000km and 12hours. The magenta line indicates the combined measurement uncertainty. The red and blue curves represent the OSSSMOSE simulated horizontal sampling and smoothing differences respectively. The green curves represent the total simulated differences, including also the errors due to measurement uncertainty and the impact of the mountain-top location of the Brewer, which doesn't see the ozone column below 2600m. From Verhoelst et al. (2015).

6. Conclusion and prospects

The metrology of an atmospheric measurement contains several aspects that are crucial when assuring the quality of the data and assessing their fitness-for-purpose with respect to user requirements, in particular those requirements expressed by the Copernicus Climate Change and Atmosphere Monitoring services(C3S and CAMS). Important aspects are (1) traceability of the data production, (2) uncertainty assessment, e.g., through propagation along the corresponding traceability diagrams, and (3) full characterization of the spatiotemporal properties of the measurements in terms of sampling, smoothing, and sensitivity. Moreover, quality assurance of (satellite) data sets involves comparison with ground-based reference measurements in order to obtain an independent indicator of the quality of the data and their uncertainties. These comparisons are affected by additional errors and uncertainties due to spatiotemporal co-location mismatch: different instruments and measurement methods have different sampling and smoothing properties, in the horizontal, vertical, and temporal domain. As the atmosphere is variable and inhomogeneous on a variety of scales, these differences can impact significantly the comparison results, depending on the co-location criteria and harmonization methods that were used to minimize or limit spatiotemporal mismatch. Consequently, proper interpretation of the comparison results requires a corresponding effort to characterize the metrology of the data comparisons.

The current document has provided a general introduction to both the metrology of a single measurement of an atmospheric variable, and that of a comparison. Generic concepts were formulated and illustrated with specific examples from data sets and comparisons playing an important role in GAIA-CLIM and related projects. This includes, among others, traceability diagrams, uncertainty propagation methods, co-location criteria, horizontal smoothing properties, and the decomposition of the error budget of a specific comparison exercise in terms of measurement and co-location mismatch errors.

Several future deliverables from GAIA-CLIM's WP3 will build upon the concepts and examples provided in the current document:

- D3.4, a technical note reporting in detail on several measurement mismatch studies and their impact on data comparisons, i.e. work performed in Task 3.2
- D3.6, a library of smoothing and sampling error estimates for key atmospheric composition measurement systems (i.e. results from Task 3.1), and for key data comparisons (results from Task 3.2).
- D3.5 and D3.7, tools for the quantification of co-location mismatch and smoothing uncertainties to be integrated into the Virtual Observatory developed in WP5.

This document is to be advertised in related projects (e.g., QA4ECV, FIDUCEO and CCI) and to the larger community dealing with atmospheric measurements and their intercomparisons, with the aim of raising awareness of the importance of a careful assessment of the measurement and of metrological aspects of a data comparison, and to spread knowledge on methods that allow mitigation and quantification of spatiotemporal smoothing and sampling errors.

References

Backus, G. & Gilbert, F., The Resolving Power of Gross Earth Data, Geophys. J. Int., 16, 169-205, 1968

BIPM, IEC, IFCC, ILAC, IUPAC, IUPAP, ISO, OIML (2012) The international vocabulary of metrology—basic and general concepts and associated terms (VIM), 3rd edn. JCGM 200:2012.

Boersma, K. F., H. J. Eskes, and E. J. Brinksma , Error analysis for tropospheric NO2 retrieval from space, J. Geophys. Res., 109, D04311, doi:10.1029/2003JD003962, 2004.

Calisesi, Y., V. T. Soebijanta, and R. van Oss, Regridding of remote soundings: Formulation and application to ozone profile comparison, J. Geophys. Res., 110, D23306, doi:10.1029/2005JD006122, 2005

von Clarmann, T.: Validation of remotely sensed profiles of atmospheric state variables: strategies and terminology, Atmos. Chem. Phys., 6, 4311-4320, doi:10.5194/acp-6-4311-2006, 2006.

von Clarmann, T., De Clercq, C., Ridolfi, M., Höpfner, M., and Lambert, J.-C.: The horizontal resolution of MIPAS, Atmos. Meas. Tech., 2, 47-54, doi:10.5194/amt-2-47-2009, 2009.

Cortesi, U., Lambert, J. C., De Clercq, C., Bianchini, G., Blumenstock, T., Bracher, A., Castelli, E., Catoire, V., Chance, K. V., De Mazière, M., Demoulin, P., Godin-Beekmann, S., Jones, N., Jucks, K., Keim, C., Kerzenmacher, T., Kuellmann, H., Kuttippurath, J., Iarlori, M., Liu, G. Y., Liu, Y., McDermid, I. S., Meijer, Y. J., Mencaraglia, F., Mikuteit, S., Oelhaf, H., Piccolo, C., Pirre, M., Raspollini, P., Ravegnani, F., Reburn, W. J., Redaelli, G., Remedios, J. J., Sembhi, H., Smale, D., Steck, T., Taddei, A., Varotsos, C., Vigouroux, C., Waterfall, A., Wetzel, G., and Wood, S.: Geophysical validation of MIPAS-ENVISAT operational ozone data, Atmos. Chem. Phys., 7, 4807-4867, doi:10.5194/acp-7-4807-2007, 2007.

De Smedt, I., Van Roozendael, M., Stavrakou, T., Müller, J.-F., Lerot, C., Theys, N., Valks, P., Hao, N., and van der A, R.: Improved retrieval of global tropospheric formaldehyde columns from GOME-2/MetOp-A addressing noise reduction and instrumental degradation issues, Atmos. Meas. Tech., 5, 2933-2949, doi:10.5194/amt-5-2933-2012, 2012.

Dinelli, B. M., Arnone, E., Brizzi, G., Carlotti, M., Castelli, E., Magnani, L., Papandrea, E., Prevedelli, M., and Ridolfi, M.: The MIPAS2D database of MIPAS/ENVISAT measurements retrieved with a multi-target 2-dimensional tomographic approach, Atmos. Meas. Tech., 3, 355-374, doi:10.5194/amt-3-355-2010, 2010.

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

Fassò, A., Ignaccolo, R., Madonna, F., Demoz, B. B., and Franco-Villoria, M.: Statistical modelling of collocation uncertainty in atmospheric thermodynamic profiles, Atmos. Meas. Tech., 7, 1803-1816, doi:10.5194/amt-7-1803-2014, 2014.

Hendrick, F., Pommereau, J.-P., Goutail, F., Evans, R. D., Ionov, D., Pazmino, A., Kyrö, E., Held, G., Eriksen, P., Dorokhov, V., Gil, M., and Van Roozendael, M.: NDACC/SAOZ UV-visible total ozone measurements: improved retrieval and comparison with correlative ground-based and satellite observations, Atmos. Chem. Phys., 11, 5975-5995, doi:10.5194/acp-11-5975-2011, 2011.

Iarlori, M., Madonna, F., Rizi, V., Trickl, T., and Amodeo, A.: Effective resolution concepts for lidar observations, Atmos. Meas. Tech., 8, 5157-5176, doi:10.5194/amt-8-5157-2015, 2015.

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.

Keppens, A., Lambert, J.-C., Granville, J., Miles, G., Siddans, R., van Peet, J. C. A., van der A, R. J., Hubert, D., Verhoelst, T., Delcloo, A., Godin-Beekmann, S., Kivi, R., Stübi, R., and Zehner, C.: Round-robin evaluation of nadir ozone profile retrievals: methodology and application to MetOp-A GOME-2, Atmos. Meas. Tech., 8, 2093-2120, doi:10.5194/amt-8-2093-2015, 2015.

Koukouli, M.-E., Lerot, C., Granville, J., Goutail, F., Lambert, J.-C., Pommereau, J.-P., Balis, D., Zyrichidou, I., Van Roozendael, M., Coldewey-Egbers, M., Loyola, D., Labow, G., Firth, S., Spurr, R., and Zehner, C.: Evaluating a new homogeneous total ozone climate data record from GOME/ERS-2, SCIAMACHY/Envisat and GOME-2/MetOp-A, J. Geophys. Res.-Atmos., 120, doi:10.1002/2015JD023699, 2015.

Kiefer, M., Arnone, E., Dudhia, A., Carlotti, M., Castelli, E., von Clarmann, T., Dinelli, B. M., Kleinert, A., Linden, A., Milz, M., Papandrea, E., and Stiller, G.: Impact of temperature field inhomogeneities on the retrieval of atmospheric species from MIPAS IR limb emission spectra, Atmos. Meas. Tech., 3, 1487-1507, doi:10.5194/amt-3-1487-2010, 2010.

Laeng, A., D. Hubert, T. Verhoelst, T. von Clarmann, B.M. Dinelli, A. Dudhia, P. Raspollini, G. Stiller, U. Grabowski, A. Keppens, M. Kiefer, V. Sofieva, L. Froidevaux, K.A. Walker, J.-C. Lambert, C. Zehner, The ozone climate change initiative: Comparison of four Level-2 processors for the Michelson Interferometer for Passive Atmospheric Sounding (MIPAS), Remote Sensing of Environment, Volume 162, Pages 316-343, doi:10.1016/j.rse.2014.12.013, 2015.

Lambert, J.-C. & Vandenbussche, S. EC FP6 GEOmon Technical Note D4.2.1 - Multi-dimensional characterisation of remotely sensed data - Chapter 1: Ground-based measurements BIRA-IASB, 2011

Lambert, J.-C., C. De Clercq, and T. von Clarmann, Comparing and merging water vapour observations: A multi-dimensional perspective on smoothing and sampling issues, Chapter 9 (p. 177-199) of book "Monitoring Atmospheric Water Vapour: Ground-Based Remote Sensing and In-situ Methods", N. Kämpfer (Ed.), ISSI Scientific Report Series, Vol. 10, Edition 1, 326 p., ISBN: 978-1-4614-3908-0, DOI 10.1007/978-1-4614-3909-7_2, © Springer New York 2012.

Lambert, J.-C., and S. Vandenbussche, "EC FP6 GEOmon Technical Note D4.2.1 - Multi-dimensional characterisation of remotely sensed data - Chapter 1: Ground-based measurements," GEOmon TN-IASB-OBSOP / Chapter 1, BIRA-IASB, April 2011.

Pappalardo, G., et al., EARLINET correlative measurements for CALIPSO: First intercomparison results, J. Geophys. Res., 115, D00H19, doi:10.1029/2009JD012147, 2010.

Ridolfi, M., Blum, U., Carli, B., Catoire, V., Ceccherini, S., Claude, H., De Clercq, C., Fricke, K. H., Friedl-Vallon, F., Iarlori, M., Keckhut, P., Kerridge, B., Lambert, J.-C., Meijer, Y. J., Mona, L., Oelhaf, H., Pappalardo, G., Pirre, M., Rizi, V., Robert, C., Swart, D., von Clarmann, T., Waterfall, A., and Wetzel, G.: Geophysical validation of temperature retrieved by the ESA processor from MIPAS/ENVISAT atmospheric limb-emission measurements, Atmos. Chem. Phys., 7, 4459-4487, doi:10.5194/acp-7-4459-2007, 2007.

Rodgers, C. D.:Retrieval of atmospheric temperature and composition from remote measurement of thermal radiation Rev. Geophys. and Space Phys., 14, 609-624, 1976.

Rodgers, C.D.: Characterization and error analysis of profiles retrieved from remote sounding measurements, J. Geophys. Res., 95, 5587–5595, doi:10.1029/JD095iD05p05587, 1990.

Rodgers, C.D.: Inverse Methods for Atmospheric Sounding, Vol. 2 of Series on Atmospheric, Oceanic and Planetary Physics, World Scientific, Singapore, 2000.

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

Seidel, D. J., B. Sun, M. Pettey, and A. Reale , Global radiosonde balloon drift statistics, J. Geophys. Res., 116, D07102, doi:10.1029/2010JD014891, 2011.

Sofieva, V. F., Kalakoski, N., Päivärinta, S.-M., Tamminen, J., Laine, M., and Froidevaux, L.: On sampling uncertainty of satellite ozone profile measurements, Atmos. Meas. Tech., 7, 1891-1900, doi:10.5194/amt-7-1891-2014, 2014.

Van Roozendael, M.; Spurr, R.; Loyola, D.; Lerot, C.; Balis, D.; Lambert, J.-C.; Zimmer, W.; van Gent, J.; van Geffen, J.; Koukouli, M.; Granville, J.; Doicu, A.; Fayt, C. & Zehner, C. , Sixteen years of GOME/ERS-2 total ozone data: The new direct-fitting GOME Data Processor (GDP) version 5—Algorithm description, J. Geophys. Res., 117, D03305, doi:10.1029/2011JD016471, 2012.

Vandenbussche, S.; De Clercq, C.; Lambert, J.-C.; Spurr, R. & von Clarmann, T. EC FP6 GEOmon Technical Note D4.2.1 - Multi-dimensional characterisation of remotely sensed data - Chapter 4: Satellite measurements of limb-scattered ultraviolet-visible light BIRA-IASB, 2010

Vandenbussche, S.; Lambert, J.-C. & Spurr, R. EC FP6 GEOmon Technical Note D4.2.1 - Multi-dimensional characterisation of remotely sensed data - Chapter 5: Satellite measurements of solar/stellar occultation BIRA-IASB, 2010

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

Annex A: QA4ECV terms and definitions

TERM	DEFINITION	SOURCE
accuracy	closeness of agreement between a measured quantity value and a true quantity value of a measurand; note that it is not a quantity and it is not given a numerical quantity value	VIM/ISO:99, GUM
area (volume) of representativeness	the area (volume) in which the concentration does not differ from the concentration at the station by more than a specific range	Larssen
bias	(1) systematic error of indication of a measuring system (2) estimate of a systematic measurement error (3) estimate of a systematic forecast error	(1) VIM/ISO:99 (2) VIM/ISO:99 (3) MACC
calibration	(1) the process of quantitatively defining the system responses to known, controlled signal inputs (2) operation that, under specified conditions, in a first step, establishes a relation between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication	(1) CEOS/ISO:19159 (2) VIM/ISO:99
dead band (or neutral zone)	maximum interval through which a value of a quantity being measured can be changed in both directions without producing a detectable change in the corresponding indication	VIM/ISO:99
detection limit	measured quantity value, obtained by a given measurement procedure, for which the probability of falsely claiming the absence of a component is β, given a probability α of falsely claiming its presence	VIM/ISO:99
error	(1) measured quantity value minus a reference quantity value (2) difference of quantity value obtained by measurement and true value of the measurand (3) difference of forecast value and a, estimate of the true value	(1) VIM/ISO:99 (2) CEOS/ISO:19159 (3) MACC
establish	define, document and implement	CDRH

instantaneous field of view (IFOV)	opening angle corresponding to one detector element	ISO:19130
fiducial	used as a fixed standard of reference for comparison or measurement (fiducial point)	WordNet
fiducial marker	refers to an object placed in the field of view of an imaging system which appears in the image produced, for use as a point of reference or a measure	
field-of-regard	an area of the object space scanned by the field-of-view of a scanning sensor	NIST
field-of-view	the solid angle from which the detector receives radiation	NIST
footprint	the area of a target encircled by the field-of-view of a detector of radiation, or irradiated by an active system	NIST
geometrical resolution	ability of a sensor system to record signals separately from neighboring object structures	DIN 18716-3: 1997-07
ground sampling distance (GSD)	linear distance between pixel centres on the ground	CEOS/ISO:19159
influence quantity	quantity that, in a direct measurement, does not affect the quantity that is actually measured, but affects the relation between the indication and the measurement result	VIM/ISO:99
in situ measurement	(1) a direct measurement of the measurand in its original place (2) any sub-orbital measurement of the measurand	(1) CEOS/ISO:19159 (2) GEOSS
measurand	quantity intended to be measured	VIM/ISO:99
metadata	data about the data; parameters that describe, characterise, and/or index the data	WMO
monitoring	(1) systematic evaluation over time of some quantity (2) by extension, evaluation over time of the performance of a system, of the occurrence of an event etc.	(1) NIST (2) MACC
point-to-area (point-to-volume) representativeness	the probability that a point measurement lies within a specific range of area-average (volume-average) concentration value	Nappo
positional accuracy	closeness of coordinate value to the true or accepted value in a specified reference system	ISO:19116

precision	(1) measure of the repeatability of a set of measurements. Note that precision is usually expressed as a statistical value based upon a set of repeated measurements such as the standard deviation from the sample mean (2) closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions	(1) ISO:19116 (2) VIM/ISO:99
procedure	specified way to carry out an activity or a process	ISO:9000
process	set of interrelated or interacting activities that use inputs to deliver an intended result	ISO:9000
process validation	establishing documented evidence of a high degree of assurance that a specific process will consistently produce a product meeting its pre-determined specifications and quality characteristics	CDRH
quality	degree to which a set of inherent characteristics of an object fulfils requirements	ISO:9000
quality assurance	part of quality management focused on providing confidence that quality requirements will be fulfilled	CEOS/ISO:19159, ISO:9000
quality assessment	term referring to the derivation of quality indicators providing sufficient information to assess whether quality requirements are fulfilled	CEOS
quality control (QC)	(1) QC refers to the activities undertaken to check and optimise accuracy and precision of the data after its collection (2) part of quality management focused on fulfilling quality requirements	(1) CEOS/ISO:19159 (2) ISO:9000
quality indicator (QI)	a means of providing a user of data or derived product with sufficient information to assess its suitability for a particular application. This information should be based on a quantitative assessment of its traceability to an agreed reference or measurement standard (ideally SI), but can be presented as a numeric or a text descriptor, provided the quantitative linkage is defined.	QA4EO
radiometric calibration	a determination of radiometric instrument performance in the spatial, spectral, and temporal domains in a series of measurements, in which its output is related to the true value of the measured radiometric quantity	NIST

random error	(1) component of measurement error that in replicate measurements varies in an unpredictable manner; note that random measurement error equals measurement error minus systematic measurement error (2) component of forecast error that varies in an unpredictable manner	(1) VIM/ISO:99 (2) MACC
relative standard uncertainty	standard measurement uncertainty divided by the absolute value of the measured quantity value	VIM/ISO:99
repeatability	measurement precision under set of conditions including the same measurement procedure, same operator, same measuring system, same operating conditions and same location, and replicated measurements over a short period of time	VIM/ISO:99
representativeness	the extent to which a set of measurements taken in a given space-time domain reflect the actual conditions in the same or different space-time domain taken on a scale appropriate for a specific application	Nappo
reproducibility	measurement precision under a set of conditions including different locations, operators, and measuring systems	VIM/ISO:99
resolution	(1) smallest change in a quantity being measured that causes a perceptible change in the corresponding indication (2) the least angular/linear/temporal/spectral distance between two identical point sources of radiation that can be distinguished according to a given criterion (3) the least vertical/geographical/temporal distance between two identical atmospheric features that can be distinguished in a gridded numerical product or in time series of measurements; resolution is equal to or coarser than vertical/geographical/temporal sampling of the grid or the measurement time series	(1) VIM/ISO:99 (2) NIST (3) MACC
stability	Property of a measuring instrument, whereby its metrological properties remain constant in time	VIM/ISO:99
systematic error	component of measurement error that in replicate measurements remains constant or varies in a predictable manner	VIM/ISO:99
system	set of interrelated or interacting elements	ISO:9000

traceability	(1) (*metrological traceability*) property of a measurement result relating the result to a stated metrological reference (free definition and not necessarily SI) through an unbroken chain of calibrations of a measuring system or comparisons, each contributing to the stated measurement uncertainty (2) ability to trace the history, application or location of an object, a product or a service	(1) VIM/ISO:99 (2) ISO:9000
traceability chain	sequence of measurement standards and calibrations that is used to relate a measurement result to a reference	VIM/ISO:99
uncertainty	non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used	VIM/ISO:99
validation	(1) the process of assessing, by independent means, the quality of the data products derived from the system outputs (2) verification, where the specified requirements are adequate for an intended use (3) confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled (4) the process of assessing, by independent means, the degree of correspondence between the value of the radiometric quantity derived from the output signal of a calibrated radiometric device and the actual value of this quantity. (5) confirmation by examination and provision of objective evidence that specifications conform to user needs and intended uses, and that the particular requirements implemented through software can be consistently fulfilled	(1) CEOS/ISO:19159 (2) VIM/ISO:99 (3) ISO:9000 (4) NIST (5) CDRH

verification	(1) provision of objective evidence that a given item fulfils specified requirements; note that, when applicable, measurement uncertainty should be taken into consideration. (2) confirmation, through the provision of objective evidence, that specified requirements have been fulfilled (3) the provision of objective evidence that the design outputs of a particular phase of the software development life cycle meet all of the specified requirements for that phase	(1) VIM/ISO:99 (2) ISO:9000 (3) CDRH
vicarious calibration	post-launch calibration of sensors that make use of natural or artificial sites on the surface of the Earth	CEOS/ISO:19159

References:

CDRH

Center for Devices and Radiological Health (CDRH), General Principles of Software Validation; Final Guidance for Industry and FDA Staff, CBER CDRH/OC Doc. N. 938, January 11, 2002. Publicly available via http://www.fda.gov/RegulatoryInformation/Guidances/ucm085281.htm

CEOS

CEOS Committee on Earth Observation Satellites (CEOS): Terms and Definitions and other documents and resources publicly available on http://calvalportal.ceos.org.

Deepika

S. R. Deepika and N. Avinash, in "Proceedings of the Fourth International Conference on Signal and Image processing", Volume 1, p. 576, 2012

DIN 18716-3

DIN 18716-3: 1997-07, Photogrammetry and remote sensing - Part 3: Remote sensing terms

GUM

Joint Committee for Guides in Metrology (JCGM/WG 1) 100:2008, Evaluation of measurement data – Guide to the expression of uncertainty in a measurement (GUM), http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf

ISO:9000

ISO 9000:2015(en), Quality management systems - Fundamentals and vocabulary

ISO:19116

ISO 19116:2004(en), Geographic information - Positioning services

ISO:19130

ISO/TS 19130-2:2014(en), Geographic information - Imagery sensor models for geopositioning - Part 2: SAR, InSAR, lidar and sonar

ISO:19159

ISO/TS 19159-1:2014(en), Geographic information - Calibration and validation of remote sensing imagery sensors and data — Part 1: Optical sensors

Larrsen

Larssen, S., R. Sluyter, and C. Helmis, Criteria for EUROAIRNET – The EEA Air Quality Monitoring and Information Network, 1999.

MACC

MACC II Service Validation Protocol, Deliverable D153.4, May 2013, http://www.gmes-atmosphere.eu/documents/maccii/deliverables/man/MACCII_MAN_DEL_D_153.1_20130528_Lambert_V2.pdf

Nappo

Nappo, C.J., Caneill J.Y., Furman R.W., Gifford F.A., Kaimal J.C., Kramer M.L., Lockhart T.J., Pendergast M.M, Pielke R.A., Randerson D., Shreffler J.H., and Wyngaard J.C., The Workshop on the Representativeness of Meteorological Observations, June 1981, Boulder, CO, Bull. Am. Meteorol. Soc. 63, 761-764, 1982.

NIST

Prokhorov, A. V., R. U. Datla, V. P. Zakharenkov, V. Privalsky, T. W. Humpherys, and V. I. Sapritsky, Spaceborne Optoelectronic Sensors and their Radiometric Calibration. Terms and Definitions. Part 1. Calibration Techniques, Ed. by A. C. Parr and L. K. Issaev, NIST Technical Note NISTIR 7203, March 2005.

QA4EO

QA4EO – A Quality Assurance framework for Earth Observation, established by the CEOS. It consists of ten distinct key guidelines linked through an overarching document (the QA4EO Principles) and more community-specific QA4EO procedures, all available on http://qa4eo.org/documentation.html A short QA4EO "user" guide has been produced to provide background into QA4EO and how one would start implementing it (http://qa4eo.org/docs/QA4EO_guide.pdf)

VIM/ISO:99

Joint Committee for Guides in Metrology (JCGM/WG 2) 200:2012 & ISO/IEC Guide 99-12:2007, International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (VIM), http://www.bipm.org/en/publications/guides/vim.html

WMO

WMO Quality Management Framework (QMF), home page at http://www.bom.gov.au/wmo/quality_management.shtmll

WordNet

Princeton University "About WordNet." WordNet. Princeton University. 2010, http://wordnet.princeton.edu.