# Metrology of ground-based satellite validation: co-location mismatch and smoothing issues of total ozone comparisons

Tiji Verhoelst, José Granville, François Hendrick, U. Köhler, Christophe Lerot, Jean-Pierre Pommereau, A. Redondas, Michel van Roozendael, Jean-Christopher Lambert

Atmos. Meas. Tech., 8, 5039–5062, 2015
www.atmos-meas-tech.net/8/5039/2015/
doi:10.5194/amt-8-5039-2015


Atmospheric Measurement Techniques
Open Access
EGU

# Metrology of ground-based satellite validation: co-location mismatch and smoothing issues of total ozone comparisons

**T. Verhoelst[1], J. Granville[1], F. Hendrick[1], U. Köhler[2], C. Lerot[1], J.-P. Pommereau[3], A. Redondas[4], M. Van Roozendael[1], and J.-C. Lambert[1]**

[1]Belgian Institute for Space Aeronomy (BIRA-IASB), Ringlaan 3, 1180 Uccle, Belgium
[2]Meteorological Observatory at Hohenpeißenberg, Deutscher Wetterdienst (DWD-MOHp), Hohenpeißenberg, Germany
[3]Laboratoire Atmosphères, Milieux, Observations Spatiales (LATMOS), CNRS/UVSQ, Guyancourt, France
[4]Izaña Atmospheric Research Center, AEMET, Santa Cruz de Tenerife, Spain

*Correspondence to:* T. Verhoelst (tijl.verhoelst@aeronomie.be)

Received: 23 May 2015 – Published in Atmos. Meas. Tech. Discuss.: 4 August 2015
Revised: 6 November 2015 – Accepted: 11 November 2015 – Published: 2 December 2015

**Abstract.** Comparisons with ground-based correlative measurements constitute a key component in the validation of satellite data on atmospheric composition. The error budget of these comparisons contains not only the measurement errors but also several terms related to differences in sampling and smoothing of the inhomogeneous and variable atmospheric field. A versatile system for Observing System Simulation Experiments (OSSEs), named OSSSMOSE, is used here to quantify these terms. Based on the application of pragmatic observation operators onto high-resolution atmospheric fields, it allows a simulation of each individual measurement, and consequently, also of the differences to be expected from spatial and temporal field variations between both measurements making up a comparison pair. As a topical case study, the system is used to evaluate the error budget of total ozone column (TOC) comparisons between GOME-type direct fitting (GODFITv3) satellite retrievals from GOME/ERS2, SCIAMACHY/Envisat, and GOME-2/MetOp-A, and ground-based direct-sun and zenith–sky reference measurements such as those from Dobsons, Brewers, and zenith-scattered light (ZSL-)DOAS instruments, respectively. In particular, the focus is placed on the GODFITv3 reprocessed GOME-2A data record vs. the ground-based instruments contributing to the Network for the Detection of Atmospheric Composition Change (NDACC). The simulations are found to reproduce the actual measurements almost to within the measurement uncertainties, confirming that the OSSE approach and its technical implementation are appropriate. This work reveals that many features of the comparison spread and median difference can be understood as due to metrological differences, even when using strict co-location criteria. In particular, sampling difference errors exceed measurement uncertainties regularly at most mid- and high-latitude stations, with values up to 10 % and more in extreme cases. Smoothing difference errors only play a role in the comparisons with ZSL-DOAS instruments at high latitudes, especially in the presence of a polar vortex due to the strong TOC gradient it induces. At tropical latitudes, where TOC variability is lower, both types of errors remain below about 1 % and consequently do not contribute significantly to the comparison error budget. The detailed analysis of the comparison results, including the metrological errors, suggests that the published random measurement uncertainties for GODFITv3 reprocessed satellite data are potentially overestimated, and adjustments are proposed here. This successful application of the OSSSMOSE system to close for the first time the error budget of TOC comparisons, bodes well for potential future applications, which are briefly touched upon.

## 1 Introduction

Compliance of essential climate variable (ECV) records obtained from satellite platforms with user requirements such as those formulated within the Global Climate Observing System (GCOS) framework, is usually assessed through validation studies. These include as a key component the compari-

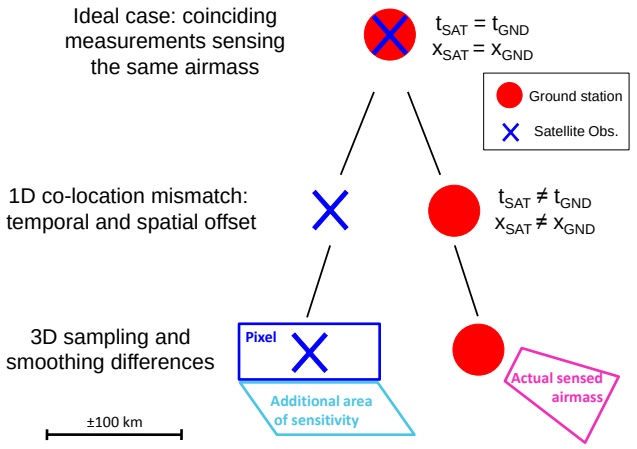


**Figure 1.** Conceptual visualization of the metrology of a satellite to ground measurement comparison. In the ideal case, ground and satellite-sensed air masses coincide in space and time. In practice, spatiotemporal sampling mismatches are inevitable, and the extent of the actually sensed air masses around the nominal locations depends on measurement types and atmospheric conditions.

son with reference measurements from ground-based instruments (see, e.g. Keppens et al., 2015, this issue, for a detailed protocol). In these validation exercises, a compromise must be made between, on the one hand, abundance of comparison pairs, and on the other hand, non-instrumental comparison errors due to non-perfect co-location in space and time between satellite and ground-based measurements. This non-perfect co-location is a consequence of both a difference in sampling, i.e. a satellite pixel centre generally does not coincide exactly with a ground station, and a difference in the way each instrument has a smoothed perception of the real, non-homogeneous, atmospheric field. Indeed, the actual air mass to which the measurement is sensitive has a 4-D extent, determined by the interplay between measurement principle and atmosphere. Figure 1 visualizes this problem of different sampling and smoothing properties of the instruments that are being compared.

While pioneering literature exists on these metrology aspects of a comparison for meteorological variables (see, e.g. Ridolfi et al., 2007; Lambert et al., 2012; Ignaccolo et al., 2015) and for ozone profiles (Sparling et al., 2006; Cortesi et al., 2007), they remain to be quantified for total ozone column (TOC) comparisons. This is the objective of the current paper. Ultimately, the aim is full error budget closure, a prerequisite for proper interpretation of the comparison results in terms of data quality.

Regarding the correct use of the terms "error" and "uncertainty", the VIM (Vocabulaire International de Métrologie, BIPM, 2012) defines an error as the (measured) quantity value minus a reference quantity value. Taking a ground-based measurement as the reference, the difference between a co-located satellite measurement and said reference measurement can thus be considered an error. This error contains several components such as for instance a measurement error and a co-location error, and it can be either positive or negative, expressed in absolute units or relative to the reference quantity value.

Uncertainty is defined as a non-negative parameter characterizing the dispersion of the quantity values attributed to a measurand. Hence, the uncertainty quantifies the statistical properties of an ensemble of errors. For instance, the random errors between measurement and truth often follow a normal probability distribution, the width of which can be considered the (random) measurement uncertainty.

In the following, the term error is therefore used for the deviation between a single value and the corresponding reference, while the term uncertainty covers the statistical properties of these errors. For instance, in Sect. 3.5, a measurement error will be simulated by a random draw from a normal distribution with a width determined by the measurement uncertainty provided with the data product.

### 1.1 Error budget of a data comparison

As an extension of the pioneering work by Rodgers (1990, 2000) and Rodgers and Connor (2003) to assess the error budget of retrieval-type remote sensing data comparisons, von Clarmann (2006) presents a unified formalism and Lambert et al. (2012) a multi-dimensional perspective including horizontal smoothing errors and errors due to less than perfect coincidence. The same error budget decomposition is followed here, and can be used as follows to relate a satellite measurement ($x_{SAT}$) with a ground-based reference measurement ($x_{GND}$): $x_{SAT} = x_{GND} + \epsilon_{total}$, with

$$\epsilon_{total} = -\epsilon_{1N} + \epsilon_{2N} - \epsilon_{1M} + \epsilon_{2M} + \epsilon_{SH} + \epsilon_{ST} + \epsilon_{dO_3/dH} + \epsilon_{dO_3/dt}, \quad (1)$$

where

- $\epsilon_{1N}$ and $\epsilon_{2N}$ represent the random errors related to the measurement uncertainty of the different sensors,
- $\epsilon_{1M}$ and $\epsilon_{2M}$ represent the systematic errors related to the measurement uncertainty,
- $\epsilon_{SH}$ represents the so-called horizontal smoothing difference error, due to differences in smoothing of horizontal structures in the atmospheric field,
- $\epsilon_{ST}$ represents the temporal smoothing difference error, due to differences in temporal averaging of atmospheric variability,
- $\epsilon_{dO_3/dH}$ represents the error due to differences in the horizontal sampling of the field, and
- $\epsilon_{dO_3/dt}$ represents the error due to differences in temporal sampling of the field.

In the first two terms, the numeric indices refer to the two different sensors. The last two terms together will hereafter be called the errors due to sampling differences, which are not to be confused with the sampling errors related to quantities derived from an incomplete sampling of a signal (see, e.g. von Clarmann, 2006). The vertical domain is not included here, since for total columns it is not applicable in the sampling sense, and already taken into account by the air mass factors in the smoothing/sensitivity sense.

To derive a total uncertainty budget from these errors, correlations between the different terms must be taken into account. This means that it is not correct to sum quadratically the uncertainties corresponding to each error term separately. These correlations arise because, e.g. sampling and smoothing differences may be sensitive to the same gradient in the atmospheric field. The approach followed here takes these correlations into account as it is based on an explicit description of the errors throughout the entire comparison metrology, and not on a summing of uncertainty estimates. This is further detailed in the following section.

### 1.2 An Observing System Simulation Experiment

Sampling difference errors in co-located data comparisons or in the construction of level-3 data have been estimated in the past using purely statistical techniques (e.g. Fassò et al., 2014), or based on some level of parametrization of atmospheric variability (e.g. Sofieva et al., 2014, and references therein). While these methods have their advantages, e.g. in terms of required computing power and/or independence of model data, they can not address all statistical properties of both sampling and smoothing difference errors.

In recent years, significant progress has been achieved in the development of pragmatic observation operators describing the actual extent of the air masses probed by each measurement technique, and in the availability of reliable, high spatial resolution, global atmospheric fields such as the Modern-Era Retrospective analysis for Research and Applications (MERRA, Rienecker et al., 2011) and the reanalysis produced within the Monitoring Atmospheric Composition and Climate project (MACC, Inness et al., 2013). This constitutes the backbone of the approach followed here, in which we estimate the comparison errors due to metrological differences through an Observation System Simulation Experiment (OSSE, see, e.g. Arnold and Dey, 1986; Errico et al., 2013). Briefly summarized it consists in the creation of multi-dimensional observation operators constrained by the real observing system metadata, followed by the application of those observation operators onto the high-resolution atmospheric fields. Provided that both observation operators and fields are realistic, this simulation allows a quantified estimate of the error terms due to smoothing and sampling differences, and of the combined metrological error. The required tools make up our software suite OSSSMOSE (Observing System of Systems Simulator for Multi-mission Synergies Exploration). The general structure of this OSSE is visualized in the flowchart in Fig. 2, and described in detail in Sect. 3.

### 1.3 Total ozone column validation as a topical case study

Total ozone column measurements from satellites remain of prime scientific importance, both for the monitoring of tropospheric ozone pollution (e.g. Valks et al., 2014), and for the detection of stratospheric ozone recovery, including its impact or dependence on climate change (e.g. Weber et al., 2011). Consequently, satellite TOC records benefit from a long-lasting validation effort, in particular by comparison with direct-sun (Brewer and Dobson) and zenith-scattered light differential optical absorption spectroscopy (ZLS-DOAS) instruments (see, e.g. Lambert et al., 1999; Balis et al., 2007a, b; Loyola et al., 2011; Koukouli et al., 2012; Labow et al., 2013). Within the Global Climate Observing System (GCOS) framework, total uncertainty and stability requirements of 2 % and 1 % decade$^{-1}$, respectively, were formulated for the TOC essential climate variable (ECV) (GCOS, 2011).

Due to the highly structured and variable nature of the atmospheric ozone field, this validation work inevitably has to deal with the impact of metrological errors on the data comparisons, an aspect which has nevertheless not been given sufficient attention in the existing literature. As such, ground-based TOC validation represents a pertinent case study for a detailed OSSE to quantify the errors due to smoothing and sampling differences.

In this context, a key product is the reprocessed TOC data set based on ESA's GOME/ERS-2 (the Global Ozone Monitoring Experiment onboard the European Remote-Sensing Satellite), SCIAMACHY/Envisat (the SCanning Imaging Absorption spectroMeter for Atmospheric CHartographY onboard the ENVIronmental SATellite, a Belgian–Dutch–German contribution to ESA's Envisat), and EUMETSAT's GOME-2/MetOp-A (GOME-2 onboard the Meteorological Operational platform) observations, produced in ESA's Ozone Climate Change Initiative (CCI) project (Lerot et al., 2014). To assess the quality of these new products, extensive validation work was carried out by comparison with co-located ground-based reference measurements, obtained with direct-sun instruments such as Dobsons and Brewers, and with ZSL-DOAS instruments such as the Système d'Analyse par Observation Zénithale (SAOZ). This validation work is already published, Koukouli et al. (2015), and it is not the purpose of the present paper to reproduce these results.

Also these ground-based reference measurements have recently benefitted from harmonization and reprocessing efforts, e.g. in ESA's "i-Cal" intercalibration project for the Dobsons and Brewers (which was a contribution to the Committee on Earth Observation Satellites, CEOS), and follow-

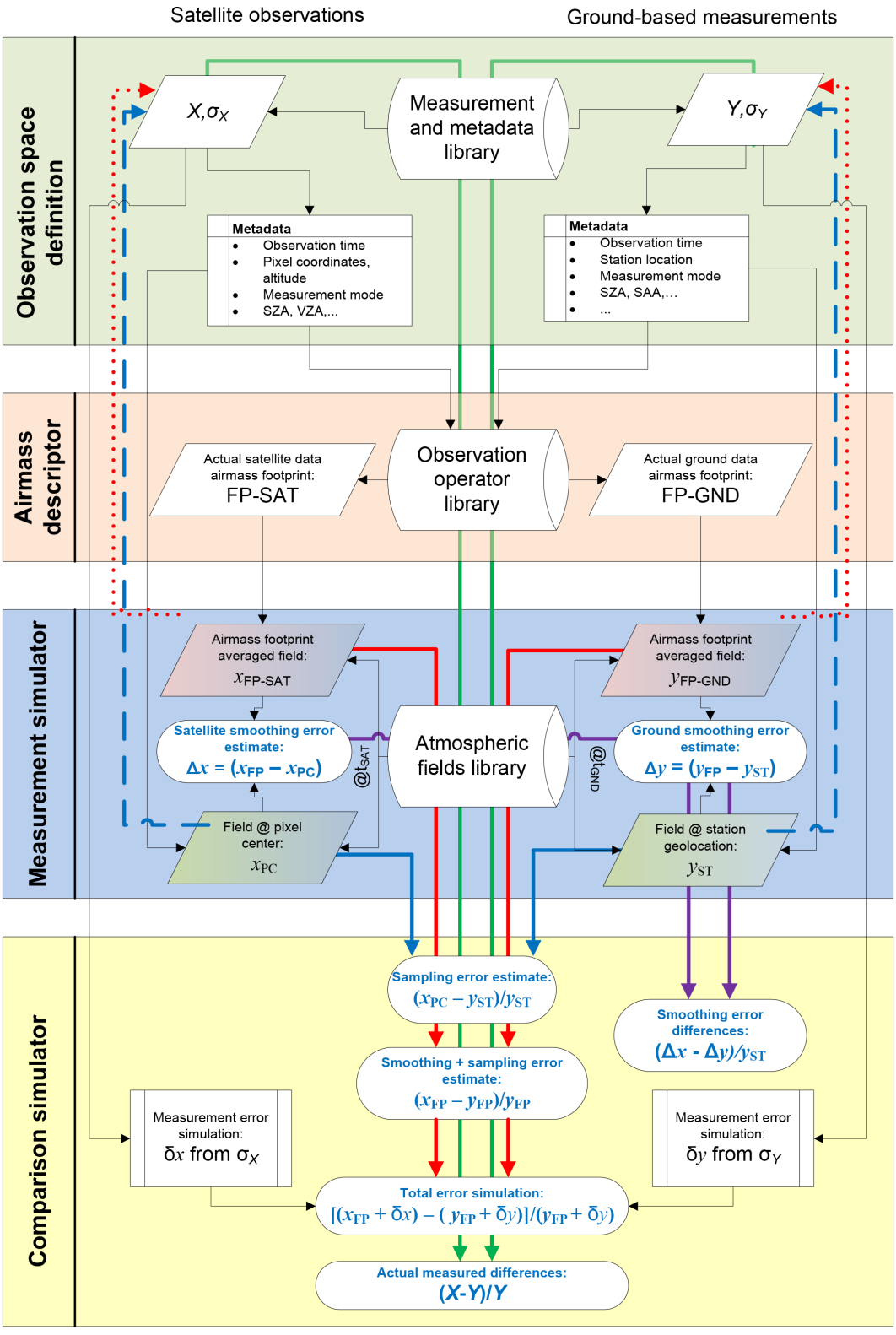


**Figure 2.** Architecture of the OSSSMOSE atmospheric metrology simulator as set up for the error budget closure of ground-based satellite validations. $X$ and $Y$ refer to the actual observations, e.g. hereafter total ozone data retrieved from GOME-2A and Brewer measurements, while $x$ and $y$ with varying subscripts refer to the simulated observations. The lateral feedback loops – highlighted in dashed blue and dotted red – show the possibility to compare the simulated observations to the real observations

ing the latest Network for the Detection of Atmospheric Composition Change (NDACC) guidelines for the ZSL-DOAS instruments (Hendrick et al., 2011). The simultaneous availability of reprocessed satellite and ground-based data with improved and document quality presents an ideal opportunity for the in-depth analysis of the ground-based TOC validation error budget reported here.

Section 2 contains the description of the different satellite and ground-based data sets used here, with due attention paid to the listed uncertainties and to the estimation of their areas of sensitivity (the observation operators). Section 3 contains the detailed description of the OSSE, including a description of the global modelled fields. In Sect. 4, three illustrative case studies, covering the different types of ground-based instruments, are analyzed in detail. Results for the comparisons between GOME-2/MetOp-A total ozone data and observations from a larger number of ground-based stations are discussed in Sect. 5. Finally, conclusions and prospects are summarized in Sect. 6.

## 2 Satellite and ground-based data: origin, uncertainties, and smoothing properties

This paper addresses the error budget of comparisons between satellite and ground-based TOC measurements. The TOC validation work performed within ESA's $O_3$ CCI and reported by Koukouli et al. (2015) represents a topical application of such comparisons. Consequently, the research presented here is based on the same co-located data sets, or subsets thereof. In this section, the specifics of these instruments and data sets are discussed, with emphasis on the known random and systematic uncertainties (characterizing the errors $\epsilon_N$ and $\epsilon_M$, respectively), and on the way they sample different air masses, information which is required to construct the corresponding observation operators.

### 2.1 Satellite data

The level-2 satellite data used here are part of a reprocessing of GOME/ERS-2, SCIAMACHY/Envisat, and GOME-2/MetOp-A observations, using the latest version of the GODFIT direct fitting retrieval algorithm, i.e. v3.0 (Lerot et al., 2014). In particular, this latest version of GODFIT deals with instrumental degradation through a soft-calibration scheme, effectively correcting level-1 radiance data by comparison with simulated spectra based on co-located Brewer total column measurements at selected sites. This and other improvements regarding a priori profiles, cloud and Ring-effect treatment, and polarization, help bring these records closer to the aforementioned GCOS requirements of 2 % total uncertainty and 1 % decade$^{-1}$ long-term stability.

Through a detailed sensitivity analysis, Lerot et al. (2014) estimate the total random uncertainty (instrument signal-to-noise ratio (SNR) plus cloud fraction and cloud top height uncertainty) to be better than 1.7 and 2.6 % for solar zenith angles (SZA) $< 80°$ and SZA $> 80°$, respectively. Systematic errors are derived to be lower than 3.6 and 5.3 %, again depending on these SZA regimes.

The area of sensitivity of such satellite nadir measurements contains the ground pixel footprint, an extension of that pixel in the direction of the sun, and, in case of a non-zero viewing angle, also an extension in the direction of the satellite. These extensions correspond to the projection on the ground of the air mass to which the measurement is sensitive, following the optical light path between sun, scatterer, and detector. A functional approximation of the horizontal spread of information (i.e. the observation operator describing the total air mass footprint) was derived from the horizontal projection of vertical averaging kernels which were computed for different solar zenith angles with the UVSPEC/DISORT (Mayer and Kylling, 2005) radiative transfer model. A full description can be found in Vandenbussche et al. (2009). The horizontal dilution in the direction of the sun ranges from a few 10s of kilometres at a SZA of 60° to almost 400 km at a SZA of 90°. For a viewing zenith angle of 31° (the maximum for normal GOME and SCIAMACHY operation modes) the horizontal dilution in the direction of the satellite is about 22 km, increasing up to 33 km for the 54° maximum viewing zenith angle (VZA) of GOME-2. An illustration of this observation operator can be found in Fig. 3.

### 2.2 Ground-based network data

Correlative ground-based total ozone column measurements used here were obtained using state-of-the-art instruments with documented quality assessment, and provided through the Network for the Detection of Atmospheric Composition Change (NDACC, http://ndacc.org). From the NDACC network, a non-exhaustive list of Brewer and Dobson direct-sun instruments is used, complemented by several Dobsons archiving data at the World Ozone and Ultraviolet Radiation Data Centre (http://woudc.org), to improve the latitude coverage, in particular in the Southern Hemisphere. The NDACC zenith–sky looking instruments which benefitted from a full data reprocessing by Hendrick et al. (2011), following the latest NDACC UV-Vis Working Group recommendations, are used as well.

All these data sources consitute the reference for the validation of satellite total ozone measurements (e.g. Lambert et al., 1999, 2000; Bramstedt et al., 2003; Balis et al., 2007a, b; McPeters et al., 2008; Koukouli et al., 2012, 2015). An "inverse" quality assessment, i.e. testing the ground-based Dobson and Brewer network by comparison with different satellite records, was performed by Fioletov et al. (2008) and revealed mean differences well below ±3 % for the better part of the stations. An overview of the stations and ground-based instruments used here is given in Table 1.

**Table 1.** Overview of the ground-based instruments used here as a source of reference data.

| Station | Lat. | Lon. | Alt. | Instrument | Institute |
|---|---|---|---|---|---|
| Direct sun instruments | | | | | |
| Sondre Stromfjord | 67.0° N | 50.7° W | 180 m | Brewer #053 (MkII) | DMI, Denmark |
| De Bilt | 52.1° N | 5.2° E | 4 m | Brewer #189 (MkIII) | KNMI, the Netherlands |
| Valentia | 51.9° N | 10.3° W | 14 m | Brewer #088 (MkIV) | ME, Ireland |
| Uccle | 50.8° N | 4.4° E | 100 m | Brewer #178 (MkII) | RMI, Belgium |
| Hohenpeißenberg | 47.8° N | 11.02° E | 980 m | Brewer #010 (MkII) | DWD, Germany |
| | | | | Dobson #104 | |
| Arosa | 46.8° N | 9.7° E | 1840 m | Dobson #101 | MeteoSwiss, Switserland |
| Obs. de Haute Provence | 43.9° N | 5.7° E | 650 m | Dobson #085 | GSMA, France + NOAA/ESRL, USA |
| Boulder | 40.0° N | 105.3° W | 1634 m | Dobson #061 | NOAA/ESRL, USA |
| Izaña | 28.3° N | 16.5° W | 2367 m | Brewer #157 (MkIII) | AEMET, Spain |
| Mauna Loa | 19.5° N | 155.6° W | 3397 m | Dobson #076 | NOAA/ESRL, USA |
| Paramaribo | 5.8° N | 55.2° W | 23 m | Brewer #159 (MkIII) | KNMI, the Netherlands |
| Darwin | 12.4° S | 130.9° E | 31 m | Dobson #078 | BoM, Australia |
| Bribane | 27.4° S | 153.1° E | 3 m | Dobson #012 | BoM, Australia |
| Lauder | 45.0° S | 169.7° E | 370 m | Dobson #072 | NIWA, New Zealand |
| Arrival Heights | 77.8° S | 166.7° E | 184 m | Dobson #017 | NIWA, New Zealand |
| UV-Vis instruments | | | | | |
| Scoresbysund | 70.5° N | 22.0° W | 68 m | SAOZ #4 | LATMOS-CNRS, France |
| Sodankylä | 67.4° N | 26.7° W | 100 m | SAOZ #17 | LATMOS-CNRS + FMI, Finland |
| Zhigansk | 66.8° N | 123.4° E | 50 m | SAOZ #12 | LATMOS-CNRS + CAO,Russia |
| Salekhard | 66.5° N | 66.7° E | 137 m | SAOZ #5 | LATMOS-CNRS + CAO,Russia |
| Harestua | 60.2° N | 10.8° E | 596 m | BISA-DOAS | BIRA-IASB, Belgium |
| Aberystwyth | 52.4° N | 4.1° W | 50 m | SAOZ #9 | Univ. of Manchester, UK |
| Jungfraujoch | 46.6° N | 8.0° E | 3580 m | SAOZ #11 | BIRA-IASB, Belgium |
| Obs. de Haute Provence | 44.0° N | 5.7° E | 650 m | SAOZ #13 | LATMOS-CNRS, France |
| Bauru | 22.3° S | 49.0° W | 640 m | SAOZ #1 | LATMOS-CNRS + UNESP, Brazil |
| Kerguelen | 49.3° S | 70.3° E | 10 m | SAOZ #3 | LATMOS-CNRS, France |
| Rio Gallegos | 51.6° S | 69.3° W | 650 m | SAOZ #26 | LATMOS-CNRS, France |
| Dumont d'Urville | 66.7° S | 140.0° E | 20 m | SAOZ #16 | LATMOS-CNRS, France |
| Dome Concorde | 75.1° S | 123.3° E | 3233 m | SAOZ #27 | LATMOS-CNRS, France |

### 2.2.1 Direct-sun instruments

Dobson and Brewer instruments measure the absorption of solar UV-light along the line-of-sight (LOS) towards the sun in the Huggins band using either a double prism monochromator (Dobson, 1957) or a grating spectrometer (Brewers, Kerr et al., 1981). Vertical columns are derived from the slant columns and provided to the users either as individual measurements (up to several tens per day) or as daily means. At SZA $> 75°$, measurements are affected by internal stray light (significantly reduced in the Mark-III and IV Brewer design with double monochromator) and by atmospheric refraction which varies amongst others with the aerosol load. The latter effect may lead to an underestimation by a few percent of the actual column at SZA $> 75°$ (Josefsson, 1992).

While estimates of the random uncertainty are generally provided with the data, and can be as good as 0.15 % uncertainty when looking at repeatability within 10 min for a Brewer at a well-established site (Scarnato et al., 2010), Van Roozendael et al. (1998) found that in order to achieve a mutual agreement between Dobson, Brewer, and UV-Vis data at the percent level across the ground-based network, several systematic effects must be taken into account: for the Dobson instruments, the temperature dependence of the ozone absorption coefficients used in the retrievals leads to a moderate seasonality in the differences (up to 1.7 % at Sodankylä), and to a systematic error up to 4 % (Bernhard et al., 2005). In winter polar vortex conditions, the effect can increase dramatically. For Brewer instruments this is less of a concern since the ratio of the cross sections at the wavelength pairs used in these instruments is less temperature dependent. In principle, it is possible to correct for this temperature dependence in the Dobson data (Komhyr et al., 1993), but this is not done for the present work. Both types of instruments are also affected by large contributions of diffuse light when observing at solar elevations below 15°. This problem is largely addressed by Brewer instruments with double monochromators (the MkIII and MkIV).

Assuming an optically thin atmosphere, a first-order approximation of the sensitivity along the LOS is the projec-

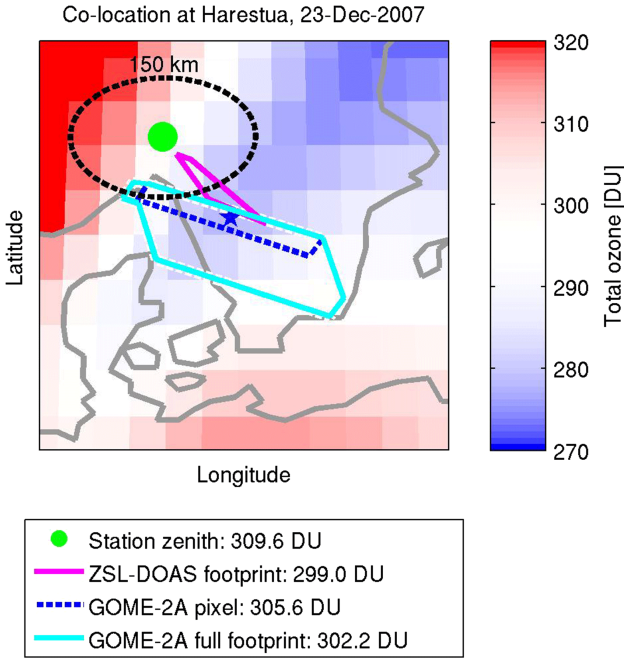


**Figure 3.** An illustration of the observation operators for a GOME-2A measurement co-located with a ZSL-DOAS observation at Harestua. The background represents the IFS-MOZART modelled TOC field at the time of the ZSL-DOAS measurement. The blue star represents the centre of the satellite pixel footprint, the blue dashed line denotes the edge of the satellite pixel footprint, and the solid cyan line represents the entire air mass of sensitivity of the satellite measurement. The latter has an extension towards the sun, in the south-east, and towards the satellite, in the west. Similarly, the green dot represents the station geo-location, while the magenta line represents the air mass of sensitivity of a morning ZSL-DOAS observation at that station. For reference, the dashed black circle describes a radius of 150 km around the station.

tion of the vertical ozone profile onto the LOS, followed by a normalization. Further projection of this sensitivity on the horizontal dimension provides a pragmatic estimate of the (1-D) air mass footprint, including relative sensitivity along the footprint. When multiple measurements are averaged into daily means, the associated range of solar azimuth angles (SAAs) leads to a 2-D footprint. In practice, the projection is limited to the middle part of the profile making up 90 % of the total column. The profile itself is taken from the Fortuin and Kelder (1998) climatology. At 75° SZA, the operational limit for Dobsons and early Brewers, the furthest point taken into account corresponds to a distance of roughly 100 km from the instrument location, with the bulk of the sensitivity around 50 km from the station. Further details can be found in Lambert and Vandenbussche (2011).

### 2.2.2 ZSL-DOAS instruments

Ground-based zenith-scattered light differential optical absorption spectrometers (ZSL-DOAS) play a key role in the long-term monitoring of stratospheric ozone and related trace gases since the late 1980s (e.g. Pommereau and Goutail, 1988; Solomon et al., 1987; McKenzie et al., 1991). Based on the differential optical absorption spectroscopy (DOAS, Platt and Stutz, 2008) technique applied to the visible Chappuis absorption band of ozone, they allow accurate observations at low sun and with limited cloud sensitivity. As such, they constitute a fundamental part of the ground-based reference instrument network used for satellite total ozone column validations, which is complementary to the direct-sun measurements obtained with Dobsons and Brewers. More than 35 of such instruments, located from pole to pole, contribute regularly to the NDACC and WOUDC archives.

While formal DOAS fitting uncertainties are generally provided with the data, these are significantly smaller than the random and systematic errors observed when comparing DOAS total columns with those obtained with direct-sun and satellite instruments (e.g. Van Roozendael et al., 1998). In particular, Van Roozendael et al. (1998) report systematic biases up to 5–6 % due to seasonal changes of the actual profile, biases up to 5 % for high altitude stations, and an average meridian dependence from −3 % at 67° N to +2.8 % at the tropics. These differences are generally attributed to uncertainties in cross sections and air mass factors (AMFs) used in the retrievals. Recently, Hendrick et al. (2011) report on a reprocessing of Système d'Analyse par Observation Zénithale (SAOZ) data (which constitute an automated subset of the ZSL-DOAS instrument network, operated by LATMOS), following homogenization recommendations by the NDACC UV-Vis working group and including a detailed error budget analysis, based on sensitivity studies w.r.t. profile climatology (for the AMF computation), clouds, aerosols, cross section, etc. The total random uncertainty of the SAOZ instruments is estimated to be about 4.7 %, and the total systematic uncertainty is conservatively put at 5.9 %.

Measurements following the typical NDACC procedure cover the range 86–91° SZA at either sunrise or sunset. Although the measurement is made by observing scattered light at zenith, the absorption signal effectively stems from the LOS between scattering agent and the sun. Using a ray-tracing code, the horizontal projection of the measurement sensitivity was derived, and taking into account the change in solar azimuth angle (SAA) during the measurement, a polygon (observation operator) can be constructed representing the air mass footprint of the measurement. Because of the very high SZA involved, the furthest points of these polygons can be located more than 500 km from the instrument. More details are available in Lambert and Vandenbussche (2011).

## 3 Metrology simulator

The core of OSSSMOSE is its metrology simulator, which consists of the following: (1) the design of an observation operator constrained by observational properties and describ-

ing the multi-dimensional sensitivity of the measurement to the atmosphere; followed by (2) the application of this observation operator onto a realistic representation of the atmospheric composition field; and (3) the calculation of metrological errors arising from the multi-dimensional nature of both the sensitivity of the observation and the atmospheric composition when point-to-area or volume-to-area assumptions are made. This suite of metrological elements is followed by an application processor enabling the calculation of, e.g., the smoothing errors associated with a single observation and with the comparison of two different observations. The modular design of OSSSMOSE is visualized in Fig. 2, and described hereafter.

### 3.1 Module 1: data and metadata

The starting point (upper green box in Fig. 2) is a library of co-located atmospheric measurements and their associated uncertainties ($X$, $\sigma_x$) and ($Y$, $\sigma_y$), built up either from existing databases (e.g., GOME-2A and NDACC total ozone data archives) or from virtual observing systems (e.g., new concept of satellite or modified network configuration). Each observation has associated with it the set of metadata and ancillary parameters needed to characterize the measurement and its 3-D sensitivity: date and time of the measurement, coordinates and elevation of the station or satellite footprint, measurement mode (e.g., ground-based direct-sun or zenith–sky, satellite nadir or limb), solar zenith and azimuth angles, viewing angle(s) and ground albedo. In particular, the basic properties of the data described in Sect. 2 are useful.

For the illustrations proposed in the following sections, the total ozone co-location libraries were built upon the following co-location criteria, reflecting community practices published in the total ozone validation literature in general and the recommendations of the international CEOS ACC ozone harmonization initiative in particular: (1) a maximum space/time distance of $150\,\mathrm{km}\,3\,\mathrm{h}^{-1}$ between the centre of the satellite field-of-view (FOV) footprint and the geolocation of the direct-sun instrument, or (2) a non-zero intersection between the centre of the satellite FOV footprint and the twilight zenith–sky air mass footprint with at most 10 h between the satellite and zenith–sky measurements, unless stated otherwise.

### 3.2 Module 2: air mass descriptor

The second module associates with each measurement a multi-dimensional description of the air mass contributing to the retrieved information: the so-called observation operator. OSSSMOSE contains a library of generic observation operators for a list of observation techniques and target molecules, including the satellite nadir UV (Vandenbussche et al., 2009), ground-based direct-sun UV and ground-based zenith–sky visible (Lambert et al., 1996; Lambert and Vandenbussche, 2011) total ozone measurement techniques considered as illustrations in the present paper. Resulting from direct and inverse simulations of the remote sensing measurement using ad hoc radiative transfer codes and retrieval tools, a generic observation operator usually consists of a parametrization of the multi-dimensional volume of the air mass contributing to the retrieved atmospheric information, including in some cases a further parametrization of the measurement sensitivity within this volume (e.g., MIPAS 2-D averaging kernels in von Clarmann et al., 2006). For total column data the air mass description can be given as the horizontal projection of this multi-dimensional object (e.g. Lambert et al., 1996; Balis et al., 2007b).

In Module 2 (orange box in Fig. 2) the metadata and parameters delivered by Module 1 (date and time, geolocation, and SZA) are used to constrain the appropriate generic observation operators of the library, yielding specific observation operators describing the actual air mass contributing to the considered observation. For the total ozone column illustrations hereafter, the actual air masses FP-SAT and FP-GND are described by either 2-D polygons (satellite and ZSL-DOAS observations) or 1-D intervals including the sensitivity curve within this interval. The actual air mass contributing to a measurement can differ significantly from either the Field-Of-View (FOV) footprint of a satellite observation or the geolocation of a ground-based instrument. Details of the computation of the specific observation operators are presented for each instrument type in Sect. 2. For nadir satellite TOC measurements the most important information concerns the pixel size and pixel location, and the solar and viewing zenith angles at the time of observation. For a ground-based measurement, required metadata are the location of the station (latitude, longitude and elevation above sea level), the instrument type (Brewer, Dobson, ZSL-DOAS), the observing mode (e.g. direct-sun or zenith–sky, a single exposure or a daily mean), and the SZA.

### 3.3 Module 3: observation simulator

The 3rd module of the system simulates each observation by applying the specific air mass descriptor generated by Module 2 into atmospheric fields. Therefore Module 3 includes a library of measured and modelled atmospheric fields at sufficiently high spatial resolution to enable accurate use of the observation operators (centre of the blue box in Fig. 2): global gridded data generated by chemical-transport models and data assimilation systems, high resolution measurements over an area taken during an airborne campaign etc. For the intended total ozone illustrations, which target among others seasonal cycles and global statistics, the fixed set-up of high-resolution reanalyses by data assimilation systems make these an appropriate source of global fields.

Ideally the atmospheric fields should have quantitative uncertainties associated with them, like systematic and random uncertainty estimates, in order to enable OSSSMOSE to calculate error propagation along its suite of operations. Un-

fortunately, uncertainties on modelled atmospheric fields are difficult to assess and the quality information documented in the literature is usually not of direct use for quantitative error propagation: it consists mainly in comparison results with reference measurements and in other quality diagnostics peculiar to the data assimilation technique. To evaluate the validity of the modelled fields for the intended use, the least that can be done is to test the robustness of the metrology simulations by feeding OSSSMOSE with different (and as independent as possible) modelled fields. Hereafter results are reported for two substantially different atmospheric representations: (1) the MACC-IFS-MOZART reanalysis performed at ECMWF, and (2) the MERRA reanalysis performed by NASA's GMAO. Their general set-up and characteristics are described below. Table 2 summarizes the relevant characteristics of each reanalysis.

#### 3.3.1 MACC (IFS-MOZART)

In the context of the EU FP7 Monitoring Atmospheric Composition and Climate Interim Implementation (MACC-II, Inness et al., 2013), the Integrated Forecast System (IFS) at European Centre for Medium-Range Weather Forecast (ECMWF) was coupled with the Model for OZone And Related chemical Tracers (MOZART-3) transport model to include chemically reactive gases (Stein et al., 2012). IFS is run at T255 spectral truncation, corresponding to roughly 80 km horizontal resolution, but MOZART-3 resolution is slightly lower at $1.125° \times 1.125°$. The vertical grid consists of 60 hybrid sigma-pressure levels, with of top of atmosphere (TOA) at 0.1 hPa. Data assimilation follows an incremental formulation of the 4D-VAR approach. The list of ozone observations that are assimilated by IFS are listed in Table 2. Global model ozone fields are available on a 6 hourly basis at the MOZART-3 horizontal resolution. Lefever et al. (2015) compared IFS-MOZART (near real time) total ozone data with ground-based reference measurements acquired by NDACC certified instrumentation (Dobson, Brewer, ozonesondes), and they find good agreement (biases below 5 % at both polar and tropical latitudes), including a reliable performance in ozone-hole conditions (reported biases below 2 %).

#### 3.3.2 MERRA

The Modern-Era Retrospective Analysis for Research and Applications (MERRA) is a reanalysis undertaken by the National Aeronautics and Space Administration (NASA)'s Global Modelling and Assimilation Office (GMAO) with the aim to place observations from NASA's Earth Observation (EO) satellites into a climate context (Rienecker et al., 2011), with a particular emphasis on an accurate representation of the hydrological cycle. MERRA was generated with version 5.2.0 of the Goddard Earth Observing System (GEOS) atmospheric model and data assimilation system (DAS). The circulation model is based on finite-volume dynamics and the data ingestion is done with a 3-D variational data assimilation (3DVAR) algorithm, based on the Gridpoint Statistical Interpolation scheme (GSI), using a 6 h update cycle. MERRA makes extensive use of satellite radiance data, using the Community Radiative Transfer Model (CRTM, Han et al., 2006) to calculate model-equivalent radiances. An extensive overview of the observations used in the production of MERRA, is given in Appendix B of Rienecker et al. (2011). Assimilated ozone data are Version 8 retrievals of SBUV2, available from October 1978 to present and provided by NASA's Goddard Earth Sciences Data and Information Services Center (GES DISC). The MERRA native grid measures $1/2°$ latitude $\times$ $2/3°$ longitude with 72 fixed-pressure vertical levels from the surface to 0.01 hPa, but assimilated chemical fields (e.g. ozone) are provided as 3-hourly instantaneous fields on a "reduced" grid of $1.25° \times 1.25°$, with 42 vertical levels. MERRA's time span was chosen to cover most of the satellite era, with an effective starting date (after a 3-year spin-up period) of 01 January 1979, and extending up to the present. While MERRA ozone data are being used for scientific purposes (e.g. Smith and Polvani, 2014), no validation or quality-assessment study of these data appears to have been published hitherto.

### 3.4 Measurement simulation

From these fields, simulated observations are calculated either as an interpolation on the nominal location of the measurement ($x_{\mathrm{PC}}$ with PC referring to the pixel centre and $y_{\mathrm{ST}}$ with ST referring to the station location), or as an averaging over the footprint derived in the previous step ($x_{\mathrm{FP}}$ and $y_{\mathrm{FP}}$). The difference between both approaches, $\Delta x$ for the simulated satellite measurements and $\Delta y$ for the simulated ground-based measurements, yields an estimate of the horizontal smoothing for both measurements. This completes the 3rd, blue, box in Fig. 2.

These simulated measurement, whether as a point-like interpolation or through averaging over the FOV footprint or over the actual air mass, can be compared to the actual measurements to gauge both the fitness-for-purpose of the modelled fields and the benefit of taking into account the smoothing properties. This is represented by the blue dashed and red dotted lines in Fig. 2. Moreover, this feedback loop can be used to further optimise the co-location criteria and the observation operators, e.g. in adjusting the somewhat ad hoc choice of vertical sensitivity limits for the ZLS-DOAS observation operator, as detailed in Sect. 4.3.3.

An illustration of these measurement simulations based on an averaging of the reanalysis field over the appropriate air mass using the associated observation operator is presented in Fig. 3.

**Table 2.** Characteristics of the two reanalyses from which atmospheric ozone fields were used as input to metrology simulations.

| Name | office | time step | lat–lon grid | vertical grid | assimilated ozone observations |
|---|---|---|---|---|---|
| IFS-MOZART-3 | ECMWF | 6 hourly | 1.125° × 1.125° | 60 levels | GOME, MIPAS, SCIAMACHY, SBUV/2, OMI, MLS |
| MERRA | NASA GMAO | 3 hourly | 1.25° × 1.25° | 42 levels | SBUV/2 |

### 3.5 Module 4: comparison simulator

Finally, the different metrological components of the error budget can be estimated and confronted with the actual difference between the retrieved total ozone values (bottom yellow box in Fig. 2) as follows.

- Using the simulated smoothing errors $\Delta x = x_{FP} - x_{PC}$ and $\Delta y = y_{FP} - y_{ST}$, for the satellite and ground-based observations, respectively, we can estimate the smoothing error differences, $\epsilon_{SH} = (\Delta x - \Delta y)$,
- Using the point-like simulated measurements at the pixel centre ($x_{PC}$) and at the station location ($y_{ST}$), each at the time of the respective observations, we can estimate the spatiotemporal sampling error, $\epsilon_{dO_3/dH} + \epsilon_{dO_3/dt} = (x_{PC} - y_{ST})$,
- Using the simulated smoothed measurements ($x_{FP}$ and $y_{FP}$, respectively), we can estimate the combined smoothing and sampling error, $\epsilon_{SH} + \epsilon_{dO_3/dH} + \epsilon_{dO_3/dt} = (x_{FP} - y_{FP})$,
- And finally, by adding simulated measurement errors, $\delta x$ and $\delta y$ to each simulated measurement, we can reconstruct the total expected distribution of differences and derive both the median error and the spread, which can be compared to the median measured difference and the measured spread on the differences.

Note that through this approach, the total error budget is not computed as the sum of individual terms, which would be incorrect since several of the terms may be correlated. For instance, the horizontal sampling and smoothing errors can be highly correlated as they are sensitive to the same gradient in the atmospheric field.

In the following section, the details and results of this OSSE are presented for three representative satellite–to-NDACC comparisons.

## 4 Case studies

In this section, the error budget OSSE is applied to three representative cases: SCIAMACHY/ENVISAT measurements vs. the Dobson at the Regional Dobson Calibration Center of Hohenpeissenberg (Germany, 47.8° N), GOME-2/MetOp-A vs. the Brewer at the Regional Brewer Calibration Center of Izana (Canary Islands, 28.3° N) and finally GOME/ERS2 vs. the SAOZ instrument at Dumont d'Urville (Antarctica, 66.7° S). These examples cover the different types of satellite and ground-based reference measurements used in O3 CCI, and they represent different atmospheric regimes: on the one hand, the comparisons at Hohenpeissenberg and Izana represent cases of relatively small comparison spread due to well-calibrated reference instruments, small satellite ground pixels, a well-behaved atmosphere, and tight co-location criteria (within O3 CCI). On the other hand, the comparisons at Dumont d'Urville are affected by the strong TOC gradients around the polar vortex, combined with large areas of measurement sensitivity. Total error budget closure requires that one can fully account for the comparison spread and median, including their temporal behaviour, with known, quantified, sources of random and systematic differences.

### 4.1 Co-located measurements and measurement footprints

An illustration of the comparison pairs at these three stations is shown in Fig. 4, one pair per season. In the context of O3 CCI, only coincidences within a 150 km radius from the station are used for direct-sun observations, such as those obtained with the Dobson at Hohenpeißenberg or the Brewer at Izaña, with at most a 3 h time difference. For the zenith–sky observations such as those at Dumont d'Urville, an intersection between the satellite pixel footprint and the ground-based air mass footprint is already enforced to minimize sampling difference errors. For these comparisons with ZSL-DOAS instruments, a larger 12 h time difference is allowed so that both sunrise and sunset ground-based measurements can be co-located with satellite observations. An evaluation of the consequences of using different (more relaxed) co-location criteria is performed in Sect. 4.5.

Also visualized in Fig. 4 are the air mass footprints of the different measurements, represented by the observation operators introduced in Sect. 2. Since a direct-sun measurement is sensitive to the absorption along the line-of-sight towards the sun, the daily means of DS measurements cover an area which depends on the SZA and SAA evolution throughout the day. The zenith–sky observations during twilight conditions cover a smaller range in SAA, but the high SZA leads to sensitivity very far from the station. Pixel sizes differ among satellite instruments (and observing modes), and further dilution of measurement sensitivity (and hence of the observa-

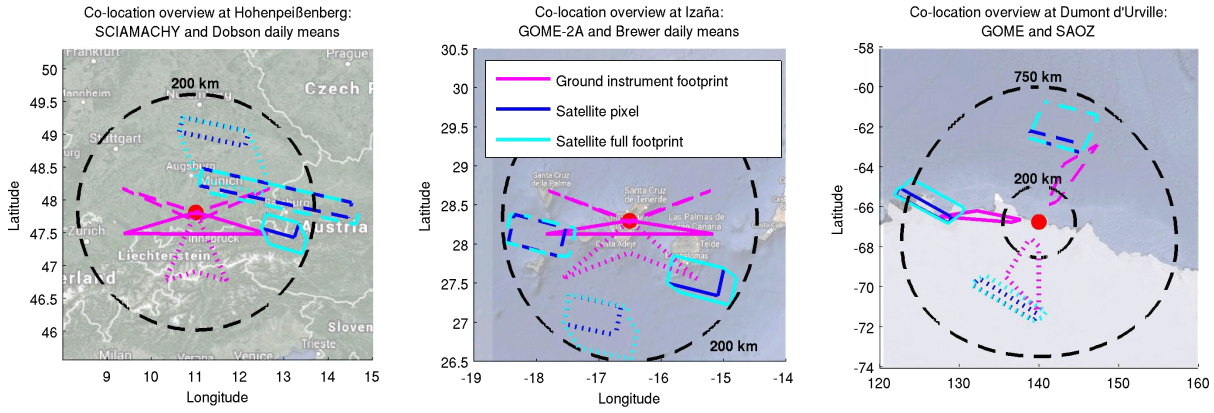


**Figure 4.** Co-located ground-satellite measurement pairs near summer and winter solstice (dashed and dotted lines, respectively) and near the autumn and spring equinox (solid line). The station is indicated by a red dot, the ground observation operators in magenta, the satellite pixel in dark blue and the full satellite observation operator in cyan.

tion operator) towards the sun or satellite depends on SZA and VZA.

### 4.2 Observed and modelled TOC time series

The corresponding observed TOC time series for both satellite $(X, \sigma_X)$ and ground-based $(Y, \sigma_Y)$ measurements are presented in Fig. 5. These illustrate the different atmospheric regimes probed by the three case studies. Also shown in these graphs are the modelled TOC time series for the satellite instrument ($x_{\mathrm{FP}}$), as derived by averaging the IFS-MOZART reanalysis fields over the observation operator shown in cyan in Fig. 4. While minor differences between observations and models are evident, the correlation coefficients ($r_{X,x_{\mathrm{FP}}} > 0.96$) and root mean square error (RMSE, $\sim$ 2–3 %) indicate a very good agreement, almost to within measurement uncertainty for stable atmospheric conditions such as those near Hohenpeißenberg and Izaña. Note that the correlation coefficient at Izaña is somewhat lower due to the intrinsic low variability of the ozone field at (sub-)tropical latitudes. A similar level of agreement is obtained using the MERRA reanalysis fields (not shown here, but further elaborated in Sect. 4.6). The use of the full observation operators (rather than pixel centres or station coordinates) for the averaging of the reanalysis field yields only minimal improvement in observation-model agreement, except for the twilight UV-Vis measurements, where the RMSE can be significantly reduced by using the observation operator (from 5.3 % down to 4.2 % in the case of Dumont d'Urville). Use of the satellite observation operator even degrades somewhat the correspondence between GOME and the IFS-MOZART reanalysis fields, but this is not surprising since the GOME data were assimilated in the IFS-MOZART reanalysis without taking into account the dilution of sensitivity towards the sun and satellite. A more detailed analysis of the use of these observation operators in the context of model-observation comparisons is beyond the scope of the current paper, but such prospects are expanded in Sect. 6.

### 4.3 Comparison error budget: observed and simulated

The satellite-ground differences, both observed ($(X-Y)/Y$, marked in black) and simulated ($[(x_{\mathrm{FP}}+\delta x)-(y_{\mathrm{FP}}+\delta y)]/(y_{\mathrm{FP}}+\delta y)$, marked in green) are visualized as 3-month running medians in Fig. 6. Some derived quantities, including model-quality indicators, are summarized in Table 3. Moreover, the simulated differences are decomposed into the different components resulting from the metrology aspects of the comparison: smoothing difference errors ($\Delta x - \Delta y / y_{\mathrm{ST}}$) in blue and sampling difference errors ($(x_{\mathrm{PC}} - y_{\mathrm{ST}})/y_{\mathrm{ST}}$) in red. The magenta line represents the combined random measurement uncertainty $\sqrt{\sigma_X^2 + \sigma_Y^2}$. Depending on the instruments involved, $\sigma_X$ and $\sigma_Y$ are taken from the data files, from the literature, or estimated here. Because the differences between satellite and ground-based measurements contain these metrological components, which depend on atmospheric structures and are thus not necessarily of a random nature, the total error budget is quantified using medians and interquantiles instead of means and variances.

#### 4.3.1 SCIAMACHY/ENVISAT vs. Dobson DS at Hohenpeißenberg

The left panel of Fig. 6 contains the 3-month running median and spread of the SCIAMACHY vs. Dobson comparisons at Hohenpeißenberg, both observed and modelled. The median difference (top panel) contains a clear seasonal component with an amplitude of roughly 2–3 %, which is not at all reproduced by the simulation. This can partly be explained by the well-known cross-section issue of the Dobson measurements already touched upon in Sect. 2.2.1. However, the amplitude of that effect is assumed to be somewhat smaller (1 %

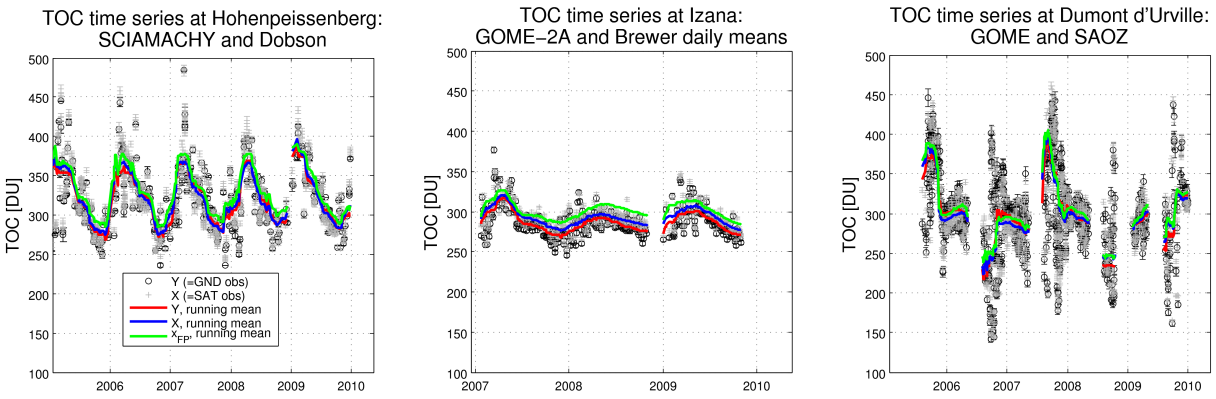


**Figure 5.** Total ozone column time series measured at the three sites with the different instruments that are being compared, including a running median of both the observed and simulated time series.

**Table 3.** OSSE quality indicators and related information for the 3 case studies discussed in Sect. 4. The first row lists the correlation between actual observations and simulated measurements and the second row lists the corresponding RMSE. The third row lists the random measurement uncertainties, either as provided with the data sets, or proposed here. The last row contains the correlation coefficient between observed and simulated satellite-ground differences.

| | Hohenpeißenberg (47.8° N) | | Izaña (28.3° N) | | Dumont d'Urville (66.7° S) | |
|---|---|---|---|---|---|---|
| | SCIAMACHY | Dobson DM | GOME-2A | Brewer DM | GOME | SAOZ |
| $r_{X,x_{FP}}$ or $r_{Y,y_{FP}}$ | 0.99 | 0.99 | 0.96 | 0.97 | 1.00 | 0.97 |
| $X - x_{FP}$ or $Y - y_{FP}$ RMSE [%] | 1.8 | 2.1 | 1.8 | 1.7 | 1.7 | 4.2 |
| $\sigma_x$ or $\sigma_y$ [%] | 1.0 | 0.8 | 0.7 | 1.0 | 1.0 | 2.5 |
| $r_{X-Y,x_{FP}-y_{FP}}$ | 0.43 | | 0.63 | | 0.77 | |

at mid latitudes, see Van Roozendael et al., 1998), and also the Brewer comparisons at Hohenpeißenberg show some seasonality (see Fig. 15), though these should not be affected by a temperature dependence. It can therefore not be ruled out that an unaccounted for effect is introducing additional seasonality in the comparison median. Additionally, some smaller features can be observed which do also appear in the simulations and can as such be attributed to either smoothing or sampling difference errors. The observed comparison spread (bottom panel) exceeds the combined measurement uncertainty (magenta line) almost continuously, including several particularly large features. The simulated errors, and in particular those due to the sampling differences, can account for the average comparison spread and for most of these features (except for winter 2007–2008). Smoothing difference errors remain below the combined measurement uncertainty and are thus only a minor component of the total error budget for this particular case.

The derivation of the combined measurement uncertainty used here warrants some discussion. Uncertainties provided with NDACC archive data files for the daily mean Dobson measurements represent the uncertainty on the mean of the individual measurements. These are used directly as $\sigma_Y$. As discussed in Sect. 2.1, the uncertainties provided with the GODFITv3 satellite data contain only the formal fit uncertainty and are known not to represent the full random uncertainty. The random uncertainty derived from sensitivity studies by Lerot et al. (2014), i.e. 1.7–2.6 %, on the other hand, is found here to be too conservative: using a 1.7 % measurement uncertainty in the simulation leads to a clear overestimation of the comparison spread. In fact, best agreement between observed and simulated comparison spread is achieved using a 1 % uncertainty on the satellite measurements. In Sect. 5, it is shown that this value holds for the comparisons at all mid- and high-latitude NDACC stations, regardless of ground instrument type. At tropical latitudes, the precision appears to be even better, as demonstrated in the following section.

### 4.3.2 GOME2/MetOp-A vs. Brewer DM at Izaña

The middle panel of Fig. 6 contains the results for GOME-2 vs. Brewer (daily mean) comparisons at Izaña. The comparison median (upper panel) contains both a clear non-zero median and a seasonal component. The seasonal component is very well reproduced by the simulation and thus is not an indication of cross-section or SZA-dependence issues. The large positive median difference of about 3 % is typical for high-altitude stations within a low-altitude region: the ground-based measurements miss the column below the station altitude, while the larger satellite pixel en-

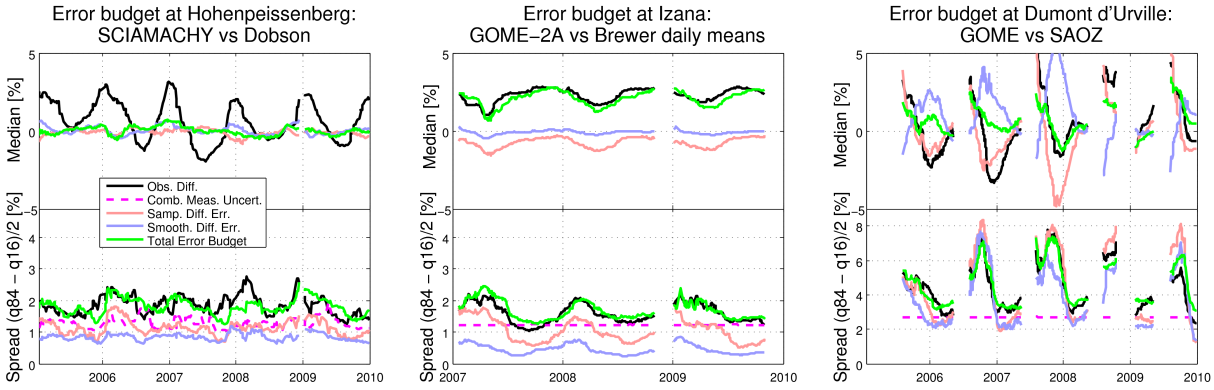


**Figure 6.** Running 3-month comparison median and spread (as derived from 16 and 84 % quantiles), both observed (black) and simulated (green), and the decomposition in the different metrological components of the simulations. Note the larger range of the bottom right-hand panel.

compasses the entire column. The 4-D reanalysis fields used here contain the required vertical information to estimate this effect and although an extensive analysis is beyond the scope of this paper, a simulation for Izaña (2367 m a.s.l.) with the IFS-MOZART fields suggests a missing column in the ground-based measurements of $3.0 \pm 0.5$ %, which is in excellent agreement with the observations: the green curve takes this vertical metrology component into account as a time-invariant 3 % shift. The comparison spread also contains a strong seasonal component with a minimum corresponding to the combined measurement uncertainty (assuming 0.7 % uncertainty on the satellite data) during local summer–autumn and almost double that spread in local winter-spring. This seasonal increase in comparison spread is fully reproduced by the simulations and mostly due to spatiotemporal sampling differences. Smoothing difference errors are estimated to reach up to 0.8 %, but this is still below the combined measurement uncertainty. Both comparison median and spread are therefore fully understood for this comparison.

### 4.3.3 GOME/ERS2 vs. SAOZ at Dumont d'Urville

At the Antarctic ground station of Dumont d'Urville, the atmospheric dynamics are much more complex, with the ozone-depleted polar vortex either encompassing the station or not, and this on variable time scales. Moreover, the zenith–sky ground instrument operated there has a large horizontal area of sensitivity, which can mean that while the station is on one side of the vortex edge, the actual sounded air mass is on the other side. From the right panel in Fig. 6, it is clear that both the comparison spread and bias are much larger and more structured than for the other two cases. Interestingly, the OSSE manages to qualitatively reproduce this behaviour, both in comparison median and spread, for the better part of the time series. This did require the use of an assumed SAOZ measurement uncertainty of 2 %, which is considerably larger than the DOAS fitting uncertainties provided with the NDACC data files (well below 1 %) but far smaller than the 4.7 % precision derived by Hendrick et al. (2011).

An interesting exception to this overall good performance is the comparison median in 2006 and 2007, which has a more pronounced observed seasonality than seen in the simulations. Examination of the curves representing smoothing and sampling errors reveals that the sampling errors appear to match the observed differences, but that they are negated by smoothing difference errors of opposite sign. This raises the question whether our smoothing difference errors, which depend on the pragmatic observation operators, are not overestimated, e.g. by too large an assumed footprint (up to 600 km, see Sect. 2.2.2.

Indeed, from Fig. 7, it appears the best agreement between ZSL-DOAS observations and simulated measurements is obtained with a somewhat smaller assumed measurement footprint: for all ZSL-DOAS stations studied in this paper, the lowest spread in observation vs. model comparisons is obtained when using an observation operator scaled down by about 50 % compared to the default one. Unfortunately, this adjustment does not suffice to really improve the agreement between observed and simulated comparison median at Dumont d'Urville in 2006 and 2007. The extreme TOC gradient at the edge of the polar vortex may amplify the impact of both short-comings in the modelled fields and in the observation operators.

Comparisons with ZLS-DOAS measurements are particularly useful to gauge the quality of the satellite measurements at high SZA, and as such they are complementary to comparisons with direct-sun instruments which are often limited to a 75° SZA. ZLS-DOAS instruments therefore extend the validation potential of the ground-based networks considerably in the polar regions, where the SZA is high for extended periods in time. Figure 8 illustrates an analysis of the SZA dependence of the comparisons between GOME-2A and the SAOZ at Dumont d'Urville, for different seasons. Some clear

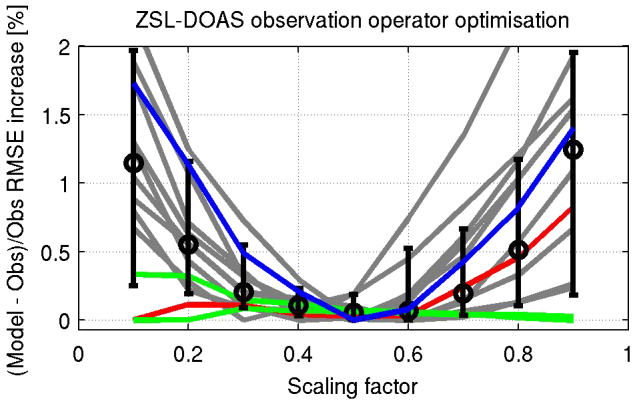


**Figure 7.** Increase in comparison spread (w.r.t. the optimum) between simulated and observed SAOZ measurements as a function of observation operator scaling factor for all NDACC stations (grey and coloured lines). The median curve with 0.16–0.84 interpercentile error bars is shown with black markers. The optimal observation operator size appears to be about half the currently assumed size. The red curve corresponds to the results at Hohenpeißenberg, and the green curves, showing no clear minimum, correspond to tropical stations (Bauru and St. Denis). At tropical latitudes, the TOC variability is low at the scale of a few hundred km and hence the exact shape of the observation operator is not of great importance. The blue curve represents the optimization at Dumont d'Urville, i.e. the current case study.

signals are detected, in particular at SZA > 70°. For instance, in local winter, the median difference increases with increasing SZA, up to almost 10 %. Also, in local spring, a particular feature is observed near 80° SZA. Interestingly, these features are at least qualitatively reproduced by the simulations, which suggests that this behaviour is mostly related to the comparison metrology, and not to instrumental or retrieval issues.

### 4.4 Error distributions

The analyses conducted in the previous sections have relied on robust statistical tools based on quantiles to determine central tendency and variability. However, in the context of meteorology and climate change, extreme values are believed to be of great importance (e.g., Katz and Brown, 1992). While it is not necessarily so that extreme values of an ECV will lead to extreme values in the differences between two instruments measuring the same event, the large gradients that occur during such events can indeed lead to large smoothing and sampling difference errors. The case study at Dumont d'Urville during ozone hole conditions is in fact an illustration of this situation (Sect. 4.3.3). To assess the quality of the simulations for differences larger than those captured by the quantiles used hitherto, entire error histograms are shown in Fig. 9 for two representative cases, corresponding to the comparisons already analyzed in Sects. 4.3.2 and 4.3.3. The comparisons between histograms of observed differences with those of modelled differences illustrate that also the tails of the distributions, beyond the 16 and 84 % quantiles, are well reproduced by the simulations, even at Dumont d'Urville where the yearly ozone hole leads to extremely low TOC values.

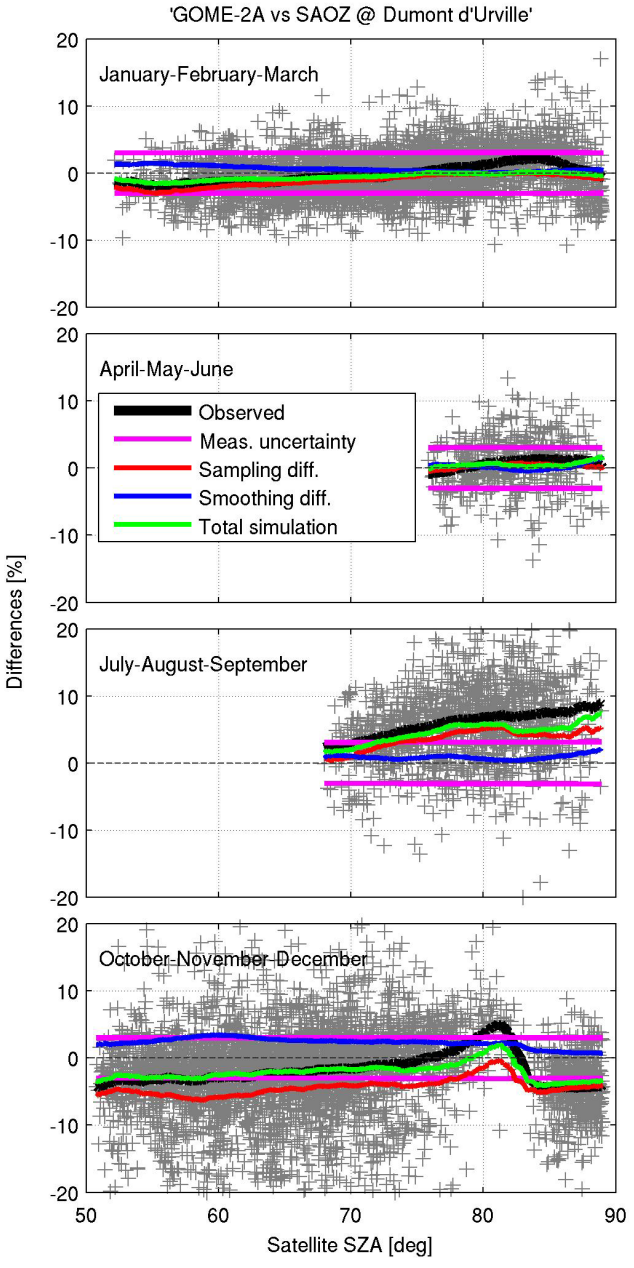


**Figure 8.** SZA dependence of the differences between GOME-2A and the SAOZ at Dumont d'Urville, grouped per season and covering 2007 to 2009. Grey crosses denote individual observed differences and solid lines represent the running 5° SZA median. While not perfect, the simulations qualitatively reproduce the observed SZA dependence, e.g. the increasing median in local winter, and the feature at 80° SZA in local spring.

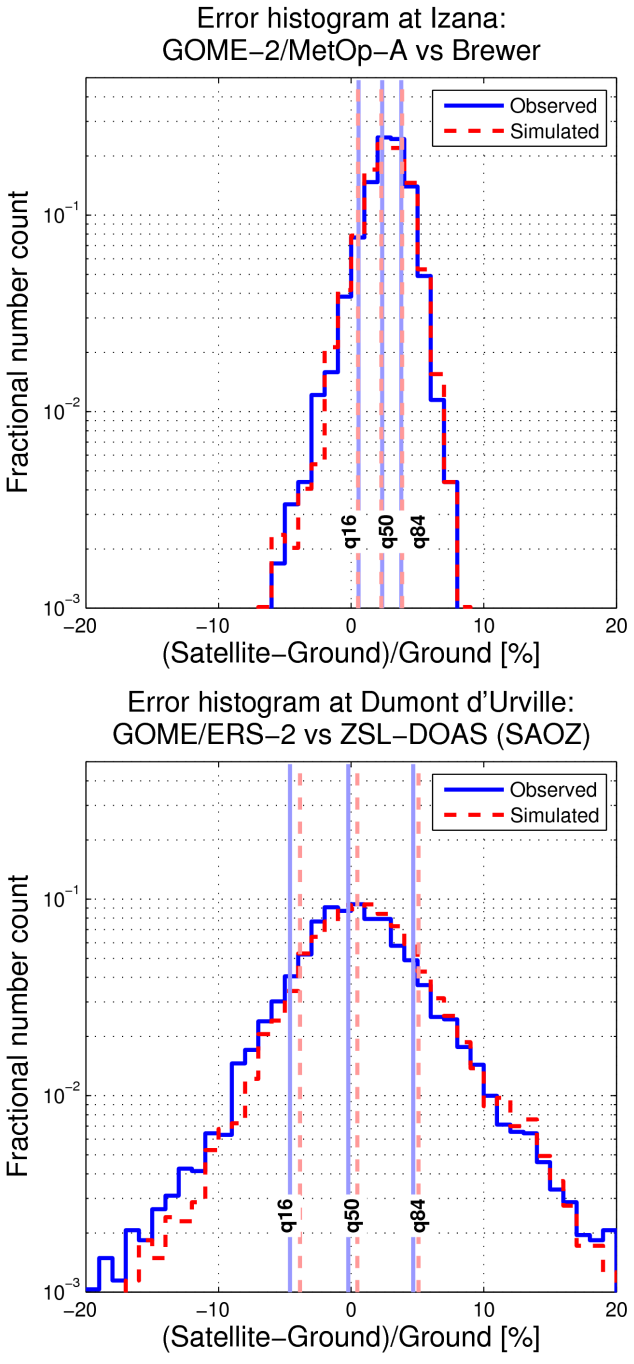


**Figure 9.** Histograms of observed and modelled differences between satellite and co-located ground-based measurements, for the comparisons at Izaña analyzed in Sect. 4.3.2 in the upper panel and for those at Dumont d'Urville from Sect. 4.3.3 in the lower panel.

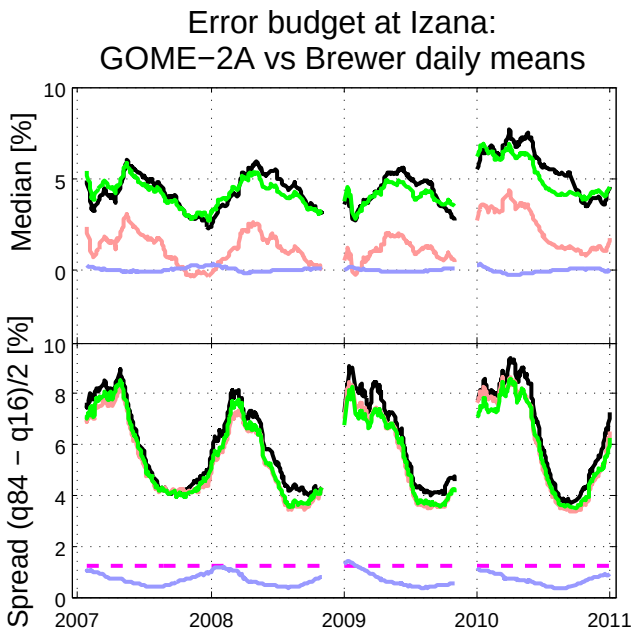


**Figure 10.** Error budget of 4 years of GOME-2A vs. Brewer comparisons at Izaña using a very relaxed spatial co-location criterium of 1000 km maximum distance. Colours as in Fig. 6.

### 4.5 Different co-location criteria

The co-location criteria used hitherto for the direct-sun comparisons, i.e. 150 km maximum spatial separation, and at most 3 h time difference, are only those of the O3 CCI validation work. Other validation campaigns have used different criteria, most often determined by the need to have a statistically representative sample of comparison pairs. For example, in earlier work a maximum spatial separation of up to 300 km was typical. As an example of the impact of more relaxed co-location criteria on the comparison statistics, Fig. 10 shows the error budget of comparisons at Izaña with a 1000 km distance maximum, to be compared to the middle panel of Fig. 6. The spread has increased from 1.5–2 to 4–9 %, and is entirely dominated by sampling mismatch errors, as expected. The median shows a seasonal behaviour of similar magnitude as for the D150 comparisons, well matched by the simulations and therefore fully due to metrological differences. Note that also the small-scale temporal structure of the median curve can be directly traced back to sampling difference errors (the red curve).

Figure 11 shows the observed comparison spread and median as a function of the spatial co-location criterium (maximum distance) for these comparisons at Izaña. The values at 1000 km correspond to the temporal average of Fig. 10. The comparison spread increases almost linearly when relaxing the co-location criterium, both in the observations and in the simulation, and this up to at least 1000 km. This behaviour is expected to saturate at distances where the autocorrelation of the ozone field is reduced to zero, but no attempt was made here to estimate that scale as it is beyond any reasonable co-location criterium used in validation work. In this particular case, the comparison median also depends strongly on co-location criterium, suggesting the presence of persistent atmospheric gradients which are sampled in a non-

homogeneous way. The green curves demonstrate that the spread and median of the OSSSMOSE simulated differences accurately reproduce the observed statistics. The $\sim 3\,\%$ offset between observed and simulated median difference is again due to the station altitude, as discussed in Sect. 4.3.2.

In fact, the simulations are realistic not only in the statistical sense (total sample spread and median), but even at the level of each individual comparison pair. This is illustrated by the cyan curves which represents the observed comparison spread and median after substraction of the metrology differences predicted by the OSSE for each individual comparison pair. As the resulting spread and bias are almost independent on co-location criterium, it is clear that the simulated differences are an excellent qualitative proxy of the real sampling and smoothing difference errors.

The error bars in Fig. 11, obtained using a bootstrap approach, illustrate the impact of the sample size on the accuracy of the spread and bias determination: a strict co-location criterium, e.g. $< 100$ km leads to a small observed comparison spread, but because that number is based on very few pairs, it has a large uncertainty. On the other side of the graph, at very large numbers of comparison pairs, the precision on the derived spread and bias is very high, but because of the large contribution to the total error budget by the sampling (and smoothing) differences, these numbers are of little direct meaning for the validation campaign. Best practice in validation work usually argues against the contamination of the data with information derived from models and as such the use of metrology-corrected observed differences is not advised, but in particular cases, such as retrieval algorithm delta validations, a metrology-correction approach may allow the detection of small improvements in measurement bias and noise which do not show up when using very strict co-location criteria.

### 4.6 Choice of modelled fields

The metrology simulations presented above were all based on the reanalyses produced in the IFS-MOZART system. While it was found that the modelled observations agree with the actual measurements almost to within measurement uncertainty, indicating very low model uncertainty for IFS-MOZART total ozone columns, independent confirmation of the reliability of the simulations can be obtained by use of fully independent reanalysis fields, such as those produced by NASA's GMAO for MERRA (see also Sect. 3.3). In general, we find the agreement between MERRA and the observations to be somewhat noisier than for IFS-MOZART (see Fig. 13 in the next section), but the satellite-ground comparisons statistics are very similar, as is illustrated for the GOME-2/MetOp-A vs. Brewer daily mean comparisons at Izaña in Fig. 12, to be compared to the middle panel of Fig. 6.

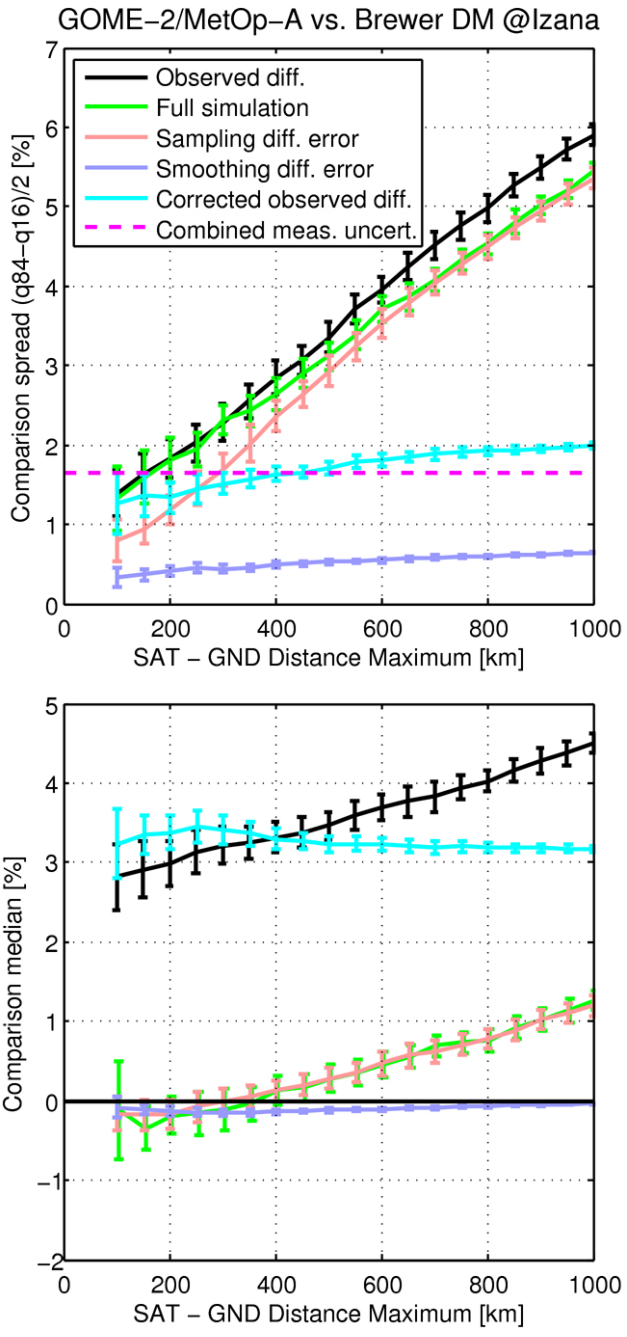


**Figure 11.** Upper panel: observed and simulated comparison spread between GOME-2/MetOp-A TOC measurements and correlative Brewer observations as a function of maximum co-location distance for the Izaña station over the period 2007–2010. Lower panel: comparison median for the same sets of comparisons. Colours as in the upper panel.

## 5 GOME-2/MetOp-A vs. the NDACC network

In this section, the methodology developed in Sect. 3 and illustrated in detail in Sect. 4 is extended to the comparisons of GOME-2/MetOp-A total columns with the entire NDACC

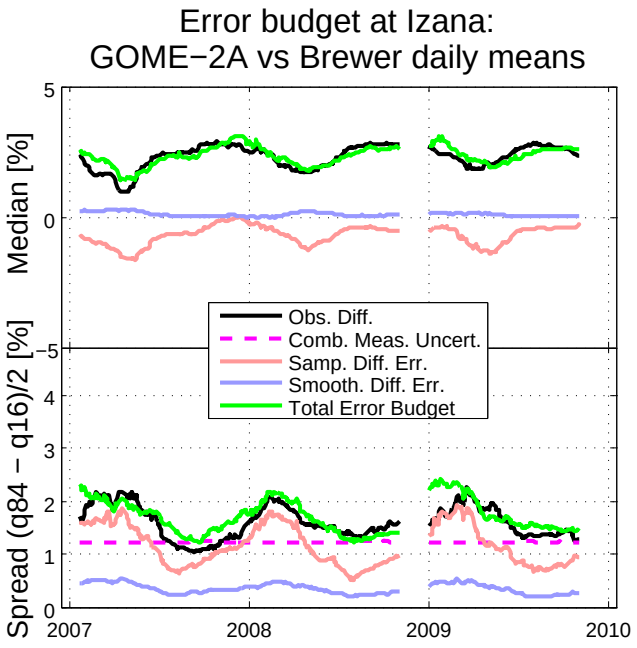


**Figure 12.** Error budget of the GOME-2/MetOp-A vs. Brewer daily mean comparisons at Izaña, derived from simulation based on MERRA fields rather than IFS-MOZART fields.

network of direct-sun and zenith–sky instruments over a 3-year period (2008–2010). This allows a more comprehensive study of the comparison error budget as a function of latitude and atmospheric regime. Further details about the NDACC network and the contributing instruments were already described in Sect. 2.

### 5.1 Models vs. GOME-2 and NDACC observations

Figure 13 illustrates the quality of the simulated TOC measurements, and hence of the underlying model fields, for both the IFS-MOZART and MERRA reanalyses. None of the observations used for this graph were assimilated in the modelled fields. The IFS-MOZART fields in general lead to the lowest comparison spread between model and observation. In particular at high southern latitudes, the difference in agreement is significant. For this reason, the analysis in this section is based only on IFS-MOZART fields. However, as illustrated in Sect. 4.6, the results do not critically depend on the choice of model fields.

### 5.2 Direct-sun instruments

Error budget simulations for comparisons between GOME-2 and NDACC Brewers and Dobsons are analyzed in Figs. 14 and 15. These comparisons follow the co-location criteria used for the validation work performed within ESA's $O_3$ CCI project, i.e. at most 150 km spatial separation between station location and satellite pixel center, and at most 3 h time difference.

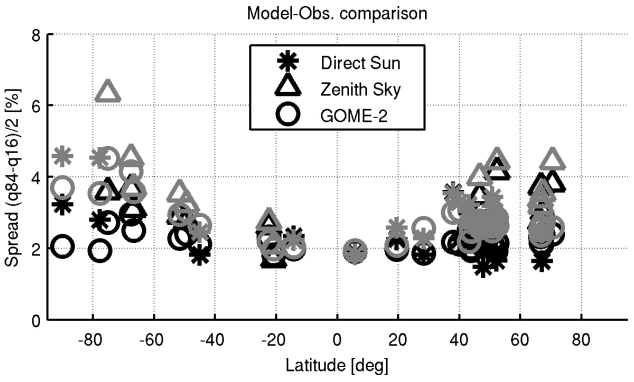


**Figure 13.** Spread of the differences between simulated TOC measurements, based on either the IFS-MOZART fields (black) or the MERRA fields (grey), and actual observations.

#### 5.2.1 Spread of the differences

The spread of the differences (Fig. 14) is remarkably well reproduced across the network, in both stable and highly variable atmospheric conditions, see, e.g. the Izaña vs. the Lauder comparisons. While smoothing difference errors (blue lines) remain below combined random measurement uncertainties (magenta lines) in all cases, sampling difference errors (red lines) often dominate the comparison spread, in particular at mid and high latitudes. At the tropical station of Paramaribo, this is not the case: both smoothing and sampling errors are well below the combined measurement uncertainties.

For two stations, Uccle and Arosa, no measurement uncertainty estimate is present in the files provided through the NDACC archive, which implies that some guestimate had to be made here. Good agreement between simulated and observed comparison spread was obtained assuming 1.5 % uncertainty for the Brewer at Uccle, 2.5 % for the Dobson there, and 1.5 % for the Dobson at Arosa. These numbers appear realistic.

As discussed in Sect. 4, the uncertainty estimate provided with the satellite data takes into account only the formal fitting uncertainty and as such is known to be too optimistic. However, the uncertainty estimate published by Lerot et al. (2014), which includes all known sources of random and systematic uncertainty, is confirmed here to be too conservative across the entire NDACC network, as already expected from the case studies in Sect. 4. Indeed, a 1 % satellite random uncertainty suffices at all stations, with the data at the tropical stations requiring only 0.7 % random uncertainty to account for the comparison spread. These numbers also hold in comparisons with zenith–sky instruments (Sect. 5.3).

It is noteworthy that for most stations, the minimum observed comparison spread roughly corresponds to the combined measurement uncertainty, i.e. there are periods during

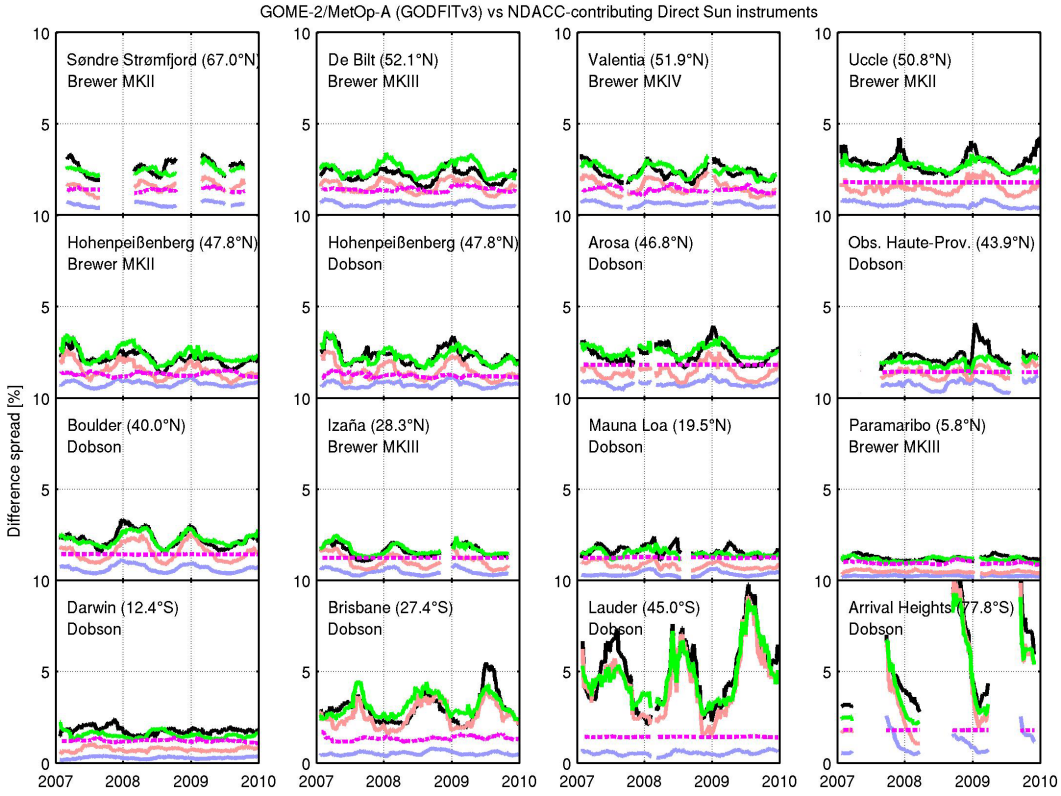


**Figure 14.** Spread of the differences (3-month running 16–84 % interquantiles) between GOME-2/MetOp-A observations and correlative direct-sun measurements (Brewers and Dobsons) from all NDACC network stations with sufficient co-locations during this period. The legend and the definition of comparison spread are the same as in Fig. 6. Note that the magenta line, representing the combined measurement uncertainty, is based on the revised estimates of the random satellite measurement uncertainty (Sect. 5.2.1).

which metrological errors are still well below measurement uncertainties, for the 150 km 3 $h^{-1}$ co-location criterium.

When relaxing the co-location criteria, as done for Hohenpeißenberg and Izaña in Sect. 4.5, the results are qualitatively the same for all stations: the errors due to sampling differences determine the comparison spread more and more, totally dominating the other error terms (smoothing and measurement errors), which do not depend on co-location distance.

### 5.2.2 Median of the differences

For the 3-month median of the differences (Fig. 15), the results are in general less satisfactory, as the observed comparison median often deviates far from zero, with strong temporal features, which can not be traced back to the comparison metrology. Still, good agreement between observed and simulated comparison median is found for the Brewers at De Bilt and Izaña (with the offset in the latter known to be due to the station altitude), and to a lesser extent also for the Brewer at Hohenpeißenberg and for the Dobson at Boulder. For the latter two stations, the simulations predict fairly significant smoothing and sampling errors, with an amplitude and structure similar to the observed comparison mean, but some discrepancies remain. Dobsons are known to have a seasonal systematic error (see Sect. 2), which could play a role here, as it appears to do in many of the other comparisons with Dobsons (Uccle, Observatoire de Haute-Provence, Lauder). For Arosa, Izaña, and Mauna Loa, the large offset can be traced back to the station altitude (w.r.t. its immediate surroundings), as was already discussed for Izaña in Sect. 4.3.2.

## 5.3 Zenith–sky instruments

Error budget simulations for comparisons between GOME-2 and NDACC UV-Vis zenith–sky instruments (SAOZ and DOAS) are analyzed in Figs. 16 and 17. Here also the comparisons follow the co-location criteria used for the validation work performed within ESA's $O_3$ CCI project, i.e. the satellite pixel footprint is required to intersect the ground-measurement air mass as quantified by the observation operator described in Sect. 2.2.2 and illustrated in the right-hand panel of Fig. 4. The observation operator used to calculate the smoothing difference errors is however the scaled-down version derived in Sect. 4.3.3. The maximum time difference is 12 h, implying that a GOME-2 measurement can be co-

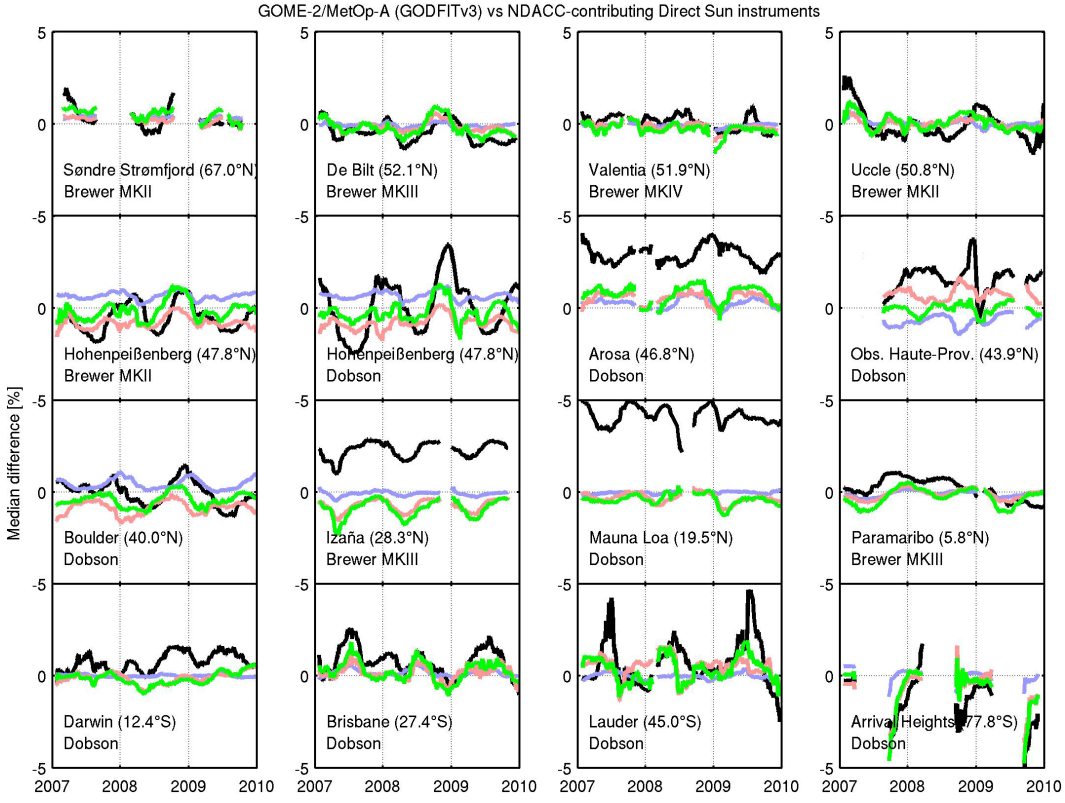


**Figure 15.** Similar to Fig. 14 but now for the median of the differences. The large median differences for Arosa, Mauna Loa, and Izaña are due to the high-altitude location of these stations, for which no correction was implemented here.

located with both sunrise and sunset zenith–sky ground measurements.

### 5.3.1 Spread of the differences

As already discussed in Sect. 2.2.2, the measurement uncertainties provided with the ground-based data are not representative for the total measurement uncertainty as they only include formal DOAS fitting uncertainties. On the other hand, the 4.7 % precision estimated by Hendrick et al. (2011) based on a detailed investigation of all sources of random and systematic uncertainty is confirmed here to be too pessimistic for all NDACC stations, as was already found for Dumont d'Urville in Sect. 4.3.3. Aiming for error budget closure, a random uncertainty of 2 to 2.5 % suffices at mid and high latitudes, and only 1 to 1.5 % is required at tropical latitudes.

As for the comparisons with direct-sun instruments, the simulations agree very well with the observed comparison spread, except for a few isolated events such as spring 2009 at Aberystwyth and winter 2009–2010 at Rio Gallegos. The comparisons at Bauru show an increase in spread towards 2010 which is not reproduced by the simulations.

### 5.3.2 Median of the differences

The median difference for the GOME-2 vs. zenith–sky UV-Vis instrument comparisons shows strong deviations from zero, with both seasonal and irregular components. While the simulations predict some non-zero medians, they do not match the observed statistics, except for a few particular features at selected stations, e.g. at Scoresbysund and at the Observatoire de Haute Provence. Surprisingly, the best agreement is in fact observed at high southern latitudes (Dumont d'Urville and Rothera). In general though, most stations show some level of pathology, be it strong seasonality (e.g. Zhigansk), a drift (e.g. Aberystwyth), or any other erratic behaviour (e.g. Bauru). The SAOZ data obtained at Sodankylä were analyzed in detail by Hendrick et al. (2011), who find a similar disagreement with the Brewer located at the same station.

## 6 Conclusions and prospects

The ever increasing accuracy of satellite total ozone column data records, required for both stratospheric and tropospheric ozone research and monitoring, and obtained through improved instrumentation and optimized retrieval methods, places correspondingly stringent requirements on the

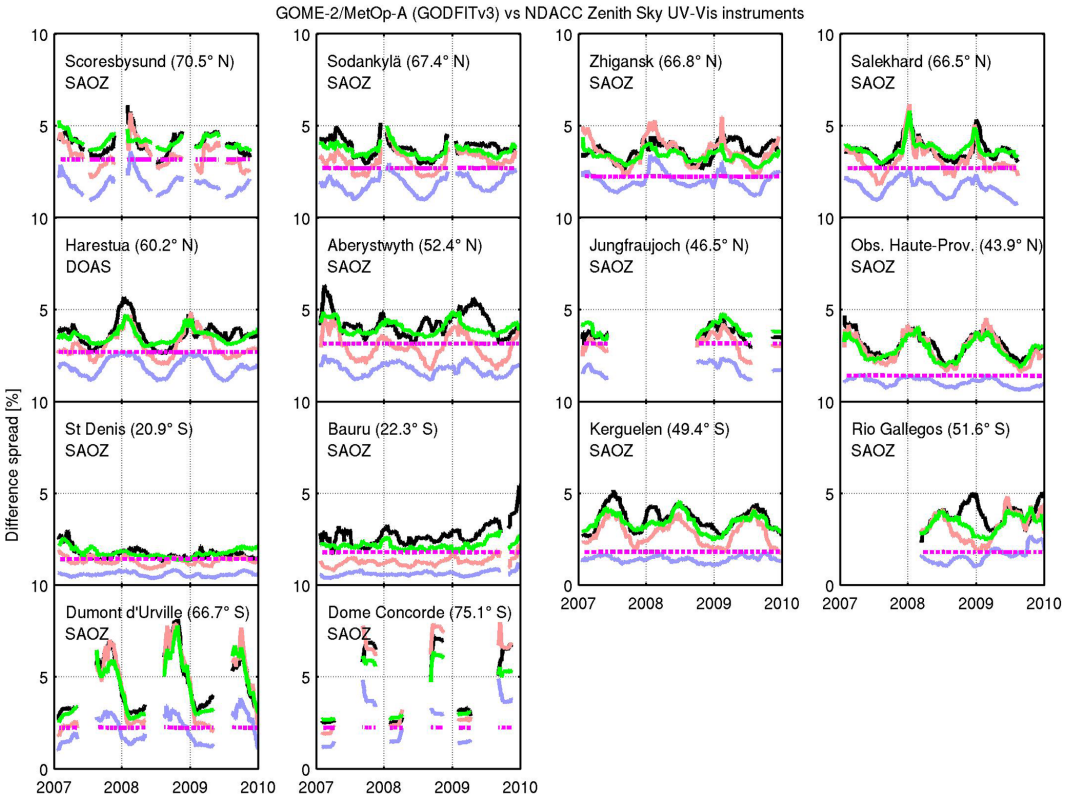


**Figure 16.** Similar to Fig. 14 but now for all NDACC UV-Vis (ZSL-DOAS) instruments with sufficient co-locations.

ground-based validation of these records. Besides the need for accurate and representative reference measurements, also the validation methodology has to be fine-tuned to current and future requirements. A key hurdle in ground-based satellite TOC validation is the introduction of additional errors in the comparisons by natural variability through non-perfect spatial and temporal co-location, including differences in smoothing of the TOC field.

In this paper, the error budget of total ozone column ground-based validation work was analyzed in detail, including for the first time the errors due to the interplay of both sampling and smoothing differences between the satellite and ground-based measurements, and an inhomogeneous and variable ozone field. These error terms were estimated using a versatile system for Observing System Simulation Experiments (OSSEs), named OSSSMOSE. The simulations are based on the real observation metadata, pragmatic observation operators, and 4-D high-resolution global ozone fields. Several station-based case studies were analyzed in detail, and extended to comparisons between GOME-2/MetOp-A and NDACC-affiliated direct-sun and zenith–sky instruments, complemented with some further stations to improve the pole-to-pole coverage.

From this work, the following conclusions could be drawn.

1. Both the modelled fields (IFS-MOZART and MERRA reanalyses) and the pragmatic observation operators are accurate enough to closely reproduce the actual satellite and ground-based observations, almost to within measurement uncertainty.

2. Comparison statistics (spread and median of the differences) derived from the simulated measurements accurately reproduce the observed comparison statistics for most satellite vs. ground-based measurement combinations, at most NDACC stations. Discrepancies, in particular in the comparison median which is indicative of systematic errors, could mostly be traced back to known instrumental issues, e.g. the Dobson's temperature-dependent (and therefore seasonal) bias.

3. Sampling difference errors range from less than 1 % to well above 10 %, depending on parameters such as co-location criterium, station latitude, and season. They are found to be a significant contributor to the error budget in almost all cases, except at tropical stations, even when using the tight co-location criteria adopted in the Committee on Earth Observation Satellites (CEOS) Atmospheric Composition Constellation (ACC) and in ESA's O3 Climate Change Initiative. Their contribution increases further if the co-location criteria are relaxed.

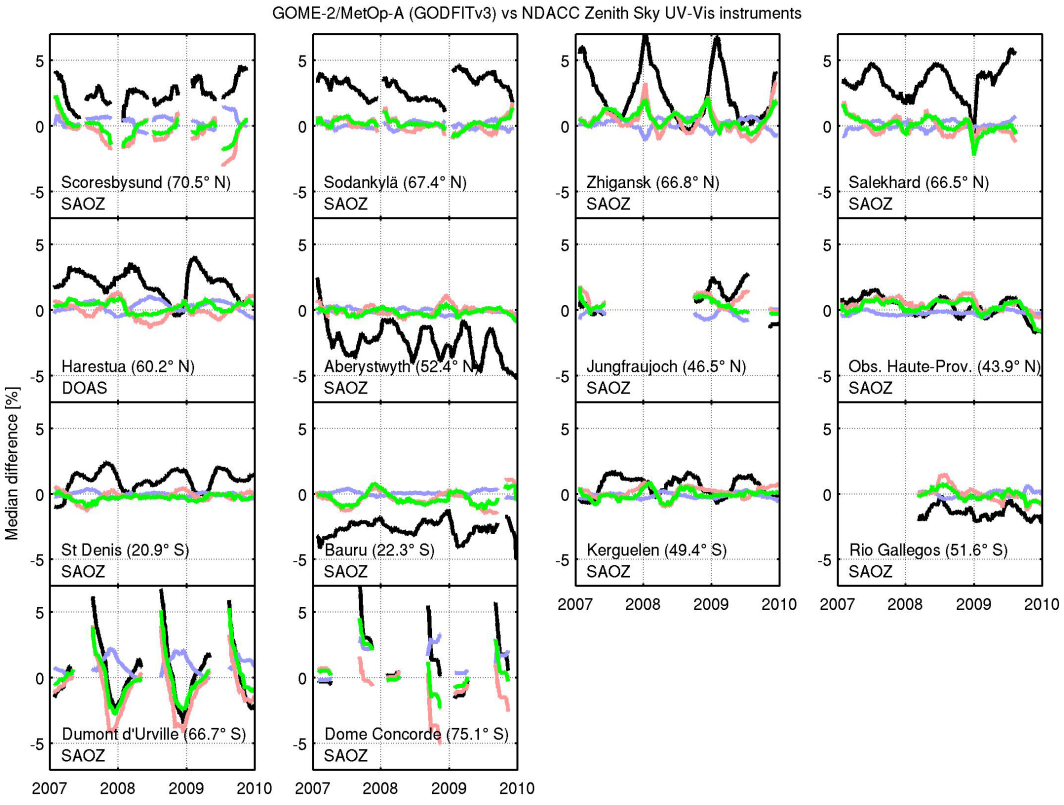


**Figure 17.** Similar to Fig. 16 but for the median of the differences.

4. Smoothing difference errors contribute only occasionally to the error budget, with amplitudes typically below 1 % for comparisons with direct-sun instruments, and below 2 % for comparisons with ZLS-DOAS measurements. They become comparable to the measurement noise only for the comparisons with zenith–sky measurements in atmospheric conditions with particularly large gradients (e.g. near the polar vortex border).

5. By correcting the observed differences with the simulated metrology errors, the comparison spread and median become almost independent of co-location criterium, illustrating that the OSSSMOSE simulations are not only meaningful in a statistical sense, but also at the level of individual comparison pairs.

6. Uncertainties provided with the satellite data records contain only the formal spectral-fit uncertainties and as such underestimate the full (random) measurement uncertainty. The random uncertainties estimated by Lerot et al. (2014), however, are found to be too conservative. For the GODFITv3 GOME-2/MetOp-A product, a random uncertainty between 0.7 % (tropics) and 1 % (mid and high latitudes) is shown here to suffice for comparison error budget closure.

7. Random uncertainties for the ground-based measurements appear reliable for most Brewers and Dobsons, except for the few stations that do not provide uncertainties. For the zenith–sky measurements, only DOAS fit uncertainties are provided with the data, and these clearly make up only a small part of the random uncertainty. The detailed uncertainty estimate by Hendrick et al. (2011) however, is found to be too conservative, as 1 % (tropics) to 2.5 % (mid and high latitudes) random uncertainty suffices for comparison error budget closure.

8. The median of the differences, used to gauge systematic errors in the data sets over periods of the order of months and longer, often deviates much further from zero than can be accounted for by the OSSSMOSE simulations. Strong biases due to sampling and smoothing issues occur only in the presence of persistent atmospheric gradients, such as near the polar vortex. Comparisons with Brewers in general show very little systematic errors (well below 1 %), while comparisons with Dobson and zenith–sky instruments on the other hand show significant (often seasonal) deviations from zero (up to 3 % for the former and up to 5 % for the latter), at least part of which can be understood from known instrumental effects in the reference measurements. The

amplitude of these features is in general found to be within the estimates of the systematic errors of these instruments published in the literature (4 % for the Dobsons and 6 % for the ZSL-DOAS instruments), but the very strong seasonality and drift at a few stations require further study.

In this paper, the OSSSMOSE system was presented and applied to a first case study: total ozone column validation work. The versatile nature of the system facilitates several further avenues of research, not yet covered in this paper. First, co-location criteria for satellite validation studies can be studied and optimized in greater detail in order to minimize the introduction of metrological errors, e.g. using wind or potential vorticity information. Also the representativeness of the ground network can be assessed and recommendations for future observing sites formulated. Similar work can be done for other reactive and greenhouse gases, meteorological variables and other ECVs (provided that reliable global gridded data, either from models or observations, are available), and for satellite intercomparison studies. Finally, the use of observation operators may improve model-observation comparisons as performed for instance in chemical data assimilation.

*Acknowledgements.* Part of this work was funded by the Belgian Science Policy Office (BELSPO) and ESA via the ProDEx projects A3C and ACROSAT, and by the EU H2020 project GAIA-CLIM (Ares(2014)3708963/Project 640276). GOME, SCIAMACHY and GOME-2A data reprocessing at IASB-BIRA was funded by ESA's CCI Ozone project. ECMWF and NASA/GMAO are thanked for providing IFS-MOZART and MERRA reanalysis, respectively. The Brewer, Dobson, and ZLS-DOAS data used in this publication were obtained as part of WMO's Global Atmosphere Watch (GAW) and the Network for the Detection of Atmospheric Composition Change (NDACC). They are publicly available via the NDACC Data Host Facility (http://ndacc.org) and the World Ozone and Ultraviolet Data Centre (http://woudc.org). The authors acknowledge the dedication of the PIs and staff at the stations to acquire and maintain long-term ozone data records of high quality, as well as supporting projects like ESA's CEOS Intercalibration of Ground-Based Spectrometers and Lidars. The authors are grateful to M. Koukouli and D. Balis for fruitful discussions on total ozone column validation. Finally, the authors acknowledge the pioneering research carried out in 2008–2012 by C. De Clercq and S. Vandenbussche and funded by the EU FP6 project GEOmon (FP6-2005-Global-4-036677) and ProDEx project SECPEA.

Edited by: P. Stammes

## References

Arnold Jr., C. P. and Dey, C. H.: Observing-systems simulation experiments: past, present, and future, B. Am. Meteorol. Soc., 67, 687–695, doi:10.1175/1520-0477(1986)067<0687:OSSEPP>2.0.CO;2, 1986.

Balis, D., Kroon, M., Koukouli, M. E., Brinksma, E. J., Labow, G., Veefkind, J. P., and McPeters, R. D.: Validation of Ozone Monitoring Instrument total ozone column measurements using Brewer and Dobson spectrophotometer ground-based observations, J. Geophys. Res., 112, D24S46, doi:10.1029/2007JD008796, 2007.

Balis, D., Lambert, J.-C., Van Roozendael, M., Loyola, D., Spurr, R., Livschitz, Y., Valks, P., Amiridis, V., Gerard, P., and Granville, J.: Ten years of GOME/ERS-2 total ozone data – the new GOME Data Processor (GDP) Version 4: II Ground-based validation and comparisons with TOMS V7/V8, J. Geophys. Res.-Atmos., 112, D07307, doi:10.1029/2005JD006376, 2007b.

Bernhard, G., Evans, R. D., Labow, G. J., and Oltmans, S. J.: Bias in Dobson total ozone measurements at high latitudes due to approximations in calculations of ozone absorption coefficients and air mass, J. Geophys. Res.-Atmos., 110, D10305, doi:10.1029/2004JD005559, 2005.

Bramstedt, K., Gleason, J., Loyola, D., Thomas, W., Bracher, A., Weber, M., and Burrows, J. P.: Comparison of total ozone from the satellite instruments GOME and TOMS with measurements from the Dobson network 1996–2000, Atmos. Chem. Phys., 3, 1409–1419, doi:10.5194/acp-3-1409-2003, 2003.

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.

Dobson, G. M. B.: Observer's Handbook for the Ozone Spectrophotometer, Annales International Geophysical Year, V, Part I: Ozone, Pergamon Press Ed., New York, 1957.

Errico, R. M., Yang, R., Privé, N. C., Tai, K.-S., Todling, R., Sienkiewicz, M. E., and Guo, J.: Development and validation of observing-system simulation experiments at NASA's Global Modeling and Assimilation Office, Q. J. Roy. Meteor. Soc., 139, 1162–1178, doi:10.1002/qj.2027, 2013.

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.

Fioletov, V. E., Labow, G., Evans, R., Hare, E. W., Köhler, U., McElroy, C. T., Miyagawa, K., Redondas, A., Savastiouk, V., Shalamyansky, A. M., Staehelin, J., Vanicek, K., and Weber, M.: Performance of the ground-based total ozone network assessed using satellite data, J. Geophys. Res., 113, D14313, doi:10.1029/2008JD009809, 2008.

Fortuin, J. and Kelder, H.: An ozone climatology based on ozonesonde and satellite measurements, J. Geophys. Res., 103, 709–734, 1998.

GCOS: Systematic Observation Requirements for Satellite-based products for Climate – 2011 Update, GCOS-154, available at: http://www.wmo.int/pages/prog/gcos/Publications/gcos-154.pdf, 2011.

Han, Y., van Delst, P., Liu, Q., Weng, F., Yan, B., Treadon, R., and Derber, J.: JCSDA Community Radiative Transfer Model (CRTM) – Version 1, NESDIS 122, Tech. rep., NOAA, 2006.

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.

Ignaccolo, R., Franco-Villoria, M., and Fassó, A.: Modelling collocation uncertainty of 3D atmospheric profiles, Stoch. Env. Res. Risk A., 29, 417–429, doi:10.1007/s00477-014-0890-7, 2015.

Inness, A., Baier, F., Benedetti, A., Bouarar, I., Chabrillat, S., Clark, H., Clerbaux, C., Coheur, P., Engelen, R. J., Errera, Q., Flemming, J., George, M., Granier, C., Hadji-Lazaro, J., Huijnen, V., Hurtmans, D., Jones, L., Kaiser, J. W., Kapsomenakis, J., Lefever, K., Leitão, J., Razinger, M., Richter, A., Schultz, M. G., Simmons, A. J., Suttie, M., Stein, O., Thépaut, J.-N., Thouret, V., Vrekoussis, M., Zerefos, C., and the MACC team: The MACC reanalysis: an 8 yr data set of atmospheric composition, Atmos. Chem. Phys., 13, 4073–4109, doi:10.5194/acp-13-4073-2013, 2013.

Joint Committee for Guides in Metrology: International Vocabulary of Metrology - Basic and General Concepts and Associated Terms, 3, available at http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf, 2012.

Josefsson, W. A. P.: Focused sun observations using a Brewer ozone spectrophotometer, J. Geophys. Res.-Atmos., 97, 15813–15817, doi:10.1029/92JD01030, 1992.

Katz, R. and Brown, B.: Extreme events in a changing climate: Variability is more important than averages, Climate Change, 21, 289–302, doi:10.1007/BF00139728, 1992

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.

Kerr, J., McElroy, C., and Olafson, R.: Measurements of total ozone with the Brewer spectrophotometer, in: Proc. Quad. Ozone Symp., 1980, edited by: London, J., Natl. Cent. for Atmos. Res., Boulder CO, 74–79, 1981.

Komhyr, W. D., Mateer, C. L., and Hudson, R. D.: Effective Bass-Paur 1985 ozone absorption coefficients for use with Dobson ozone spectrophotometers, J. Geophys. Res.-Atmos., 98, 20451–20465, doi:10.1029/93JD00602, 1993.

Koukouli, M. E., Balis, D. S., Loyola, D., Valks, P., Zimmer, W., Hao, N., Lambert, J.-C., Van Roozendael, M., Lerot, C., and Spurr, R. J. D.: Geophysical validation and long-term consistency between GOME-2/MetOp-A total ozone column and measurements from the sensors GOME/ERS-2, SCIAMACHY/ENVISAT and OMI/Aura, Atmos. Meas. Tech., 5, 2169–2181, doi:10.5194/amt-5-2169-2012, 2012.

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.

Labow, G. J., McPeters, R. D., Bhartia, P. K., and Kramarova, N.: A comparison of 40 years of SBUV measurements of column ozone with data from the Dobson/Brewer network, J. Geophys. Res.-Atmos., 118, 7370–7378, doi:10.1002/jgrd.50503, 2013.

Lambert, J.-C. and Vandenbussche, S.: 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, 2011.

Lambert, J.-C., Van Roozendael, M., Granville, J., Gérard, P., Simon, P., Claude, H., and Staehelin, J.: Comparison of the GOME ozone and NO2 total amounts at mid-latitude with ground-based zenith-sky measurements, in: Atmospheric Ozone, Proceedings of the XVIII Quadrennial Ozone Symposium, L'Aquila, Italy, 12–21 September 1996, 1998.

Lambert, J.-C., Van Roozendael, M., De Mazière, M., Simon, P., Pommereau, J.-P., Goutail, F., Sarkissian, A., and Gleason, J.: Investigation of pole-to-pole performances of spaceborne atmospheric chemistry sensors with the NDSC, J. Atmos. Sci., 56, 176–193, 1999.

Lambert, J.-C., Van Roozendael, M., Simon, P., Pommereau, J.-P., Goutail, F., Gleason, J., Andersen, S., Arlander, D., Van Bui, N., Claude, H., de La Noe, J., De Mazière, M., Dorokhov, V., Eriksen, P., Green, A., Karlsen Tornkvist, K., Kastad Hoiskar, B., Kyro, E., Leveau, J., Merienne, M.-F., Milinevsky, G., Roscoe, H., Sarkissian, A., Shanklin, J., Staehelin, J., Wahlstrom Tellefsen, C., and Vaughan, G.: Combined characterisation of GOME and TOMS total ozone measurements from space using ground-based observations from the NDSC, Adv. Space Res., 26, 1931–1940, doi:10.1016/S0273-1177(00)00178-2, 2000.

Lambert, J.-C., De Clercq, C., and von Clarmann, T.: Chapter 9: comparing and merging water vapour observations: a multi-dimensional perspective on smoothing and sampling issues, in: Ground-Based Remote Sensing and In-Situ Methods for Monitoring Atmospheric Water Vapour, ISSI, 177–199, 2012.

Lefever, K., van der A, R., Baier, F., Christophe, Y., Errera, Q., Eskes, H., Flemming, J., Inness, A., Jones, L., Lambert, J.-C., Langerock, B., Schultz, M. G., Stein, O., Wagner, A., and Chabrillat, S.: Copernicus stratospheric ozone service, 2009–2012: validation, system intercomparison and roles of input data sets, Atmos. Chem. Phys., 15, 2269–2293, doi:10.5194/acp-15-2269-2015, 2015.

Lerot, C., Van Roozendael, M., Spurr, R., Loyola, D., Coldewey-Egbers, M., Kochenova, S., van Gent, J., Koukouli, M., Balis, D., Lambert, J.-C., Granville, J., and Zehner, C.: Homogenized total ozone data records from the European sensors GOME/ERS-2, SCIAMACHY/Envisat, and GOME-2/MetOp-A, J. Geophys. Res.-Atmos., 119, 1639–1662, doi:10.1002/2013JD020831, 2014.

Loyola, D. G., Koukouli, M.-E., Valks, P., Balis, D.-S., Hao, N., Van Roozendael, M., Spurr, R., Zimmer, W., Kiemle, S., Lerot, C., and Lambert, J.-C.: The GOME-2 total column ozone product: retrieval algorithm and ground-based validation, J. Geophys. Res., 116, D07302, doi:10.1029/2010JD014675, 2011.

Mayer, B. and Kylling, A.: Technical note: The libRadtran software package for radiative transfer calculations – description

and examples of use, Atmos. Chem. Phys., 5, 1855–1877, doi:10.5194/acp-5-1855-2005, 2005.

McKenzie, R. L., Johnston, P. V., McElroy, C. T., Kerr, J. B., and Solomon, S.: Altitude distributions of stratospheric constituents from ground-based measurements at twilight, J. Geophys. Res.-Atmos., 96, 15499–15511, doi:10.1029/91JD01361, 1991.

McPeters, R., Kroon, M., Labow, G., Brinksma, E., Balis, D., Petropavlovskikh, I., Veefkind, J. P., Bhartia, P. K., and Levelt, P. F.: Validation of the Aura Ozone Monitoring Instrument total column ozone product, J. Geophys. Res.-Atmos., 113, D15S14, doi:10.1029/2007JD008802, 2008.

Platt, U. and Stutz, J.: Differential Optical Absorption Spectroscopy: Principles and Applications, Springer, available at: http://www.springerlink.com/content/978-3-540-75776-4#section=214158&page=1, 2008.

Pommereau, J. and Goutail, F.: $O_3$ and $NO_2$ ground-based measurements by visible spectrometry during Arctic winter and spring 1988, Geophys. Res. Lett., 15, 891–894, doi:10.1029/GL015i008p00891, 1988.

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.

Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., Bosilovich, M. G., Schubert, S. D., Takacs, L., Kim, G.-K., Bloom, S., Chen, J., Collins, D., Conaty, A., da Silva, A., Gu, W., Joiner, J., Koster, R. D., Lucchesi, R., Molod, A., Owens, T., Pawson, S., Pegion, P., Redder, C. R., Reichle, R., Robertson, F. R., Ruddick, A. G., Sienkiewicz, M., and Woollen, J.: MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications, J. Climate, 24, 3624–3648, doi:10.1175/JCLI-D-11-00015.1, 2011.

Rodgers, C.: 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.

Scarnato, B., Staehelin, J., Stübi, R., and Schill, H.: Long-term total ozone observations at Arosa (Switzerland) with Dobson and Brewer instruments (1988–2007), J. Geophys. Res.-Atmos., 115, D13306, doi:10.1029/2009JD011908, 2010.

Smith, K. L. and Polvani, L. M.: The surface impacts of Arctic stratospheric ozone anomalies, Environ. Res. Lett., 9, 074015, doi:10.1088/1748-9326/9/7/074015, 2014.

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.

Solomon, S., Schmeltekopf, A. L., and Sanders, R. W.: On the interpretation of zenith sky absorption measurements, J. Geophys. Res., 92, 8311–8319, doi:10.1029/JD092iD07p08311, 1987.

Sparling, L. C., Wei, J. C., and Avallone, L. M.: Estimating the impact of small-scale variability in satellite measurement validation, J. Geophys. Res.-Atmos., 111, D20310, doi:10.1029/2005JD006943, 2006.

Stein, O., Flemming, J., Inness, A., Kaiser, J. W., and Schultz, M. G.: Global reactive gases forecasts and reanalysis in the MACC project, Journal of Integrative Environmental Sciences, 9, 57–70, doi:10.1080/1943815X.2012.696545, 2012.

Valks, P., Hao, N., Gimeno Garcia, S., Loyola, D., Dameris, M., Jöckel, P., and Delcloo, A.: Tropical tropospheric ozone column retrieval for GOME-2, Atmos. Meas. Tech., 7, 2513–2530, doi:10.5194/amt-7-2513-2014, 2014.

Van Roozendael, M., Peeters, P., Roscoe, H., De Backer, H., Jones, A., Bartlett, L., Vaughan, G., Goutail, F., Pommereau, J.-P., Kyro, E., Wahlstrom, C., Braathen, G., and Simon, P.: Validation of ground-based visible measurements of total ozone by comparison with Dobson and Brewer Spectrophotometers, J. Atmos. Chem., 29, 55–83, doi:10.1023/A:1005815902581, 1998.

Vandenbussche, S., De Clercq, C., and Lambert, J.-C.: EC FP6 GEOmon Technical note D4.2.1 – Multi-dimensional characterisation of remotely sensed data – Chapter 3: Satellite measurements of nadir-scattered ultraviolet-visible light, GEOmon TN-IASB-OBSOP/Chapter 3, BIRA-IASB, 2009.

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.

Weber, M., Dikty, S., Burrows, J. P., Garny, H., Dameris, M., Kubin, A., Abalichin, J., and Langematz, U.: The Brewer-Dobson circulation and total ozone from seasonal to decadal time scales, Atmos. Chem. Phys., 11, 11221–11235, doi:10.5194/acp-11-11221-2011, 2011.