Response to reviews on Representativeness errors in comparing chemistry transport and chemistry climate models with satellite UV/Vis tropospheric column retrievals.

San Francisco, 18 December 2015

Dear editor,

Below please find a point-by-point response to the reviewers' comments on our manuscript. The changes made to the manuscript in response to these comments are evident in our response.

Below the list of responses, we added the revised manuscript with changes in blue. Removed parts are indicated by strikethrough font.

We hope that with this revision, our manuscript is accepted for publication in GMD.

Kind regards,

Folkert Boersma

Anonymous Referee #1

This manuscript presents a methodology for properly comparing models for atmospheric composition with Level-2 retrievals from satellites in the UV/VIS. It does so by identifying three potential issues, testing the potential introduced errors for these issues and recommend procedures to avoid the errors. The manuscript is generally well written (apart from Section 2; see detailed comments below) and the scientific methods are sound. However, the presented concepts and methods are not really new and the manuscript is in that sense a bit disappointing. Sampling issues and the use of averaging kernels have been discussed in the literature for some time now. Nevertheless, I support the publishing of this manuscript as a model assessment methods paper, subject to minor revision, because it provides a good summary of the issues and guidelines, which will hopefully encourage scientists in the field to make better use of these concepts.

The concepts and methods used in our study indeed exist for some 20 years now. But whereas the data assimilation community may generally well aware of the issues at hand, we feel that the modeling community using tropospheric column retrievals of NO2, HCHCO, and SO2 from UV/Vis measurements is still struggling with representativeness errors, and generally lacks guidelines on how to perform a proper model-column comparison. This is partly because the UV/Vis retrievals do not always provide averaging kernels in their data products. Another reason is that the use of UV/Vis retrievals is relatively recent compared with the use of temperature and O3 profiles from satellite instruments in formal data assimilation systems, and new users lack a body of literature to guide them how to actually do that. Our paper indeed has the potential to encourage proper use of the UV/Vis satellite retrievals in the community.

Detailed comments:

Section 1: I am missing some important earlier work, such as the studies by Rodgers and Connor, JGR, 2003 and Migliorini et al., Monthly Weather Review, 2008. These could maybe be added in lines 16/17 of page 7827?

We added these relevant references in section 1.

P7827, line 11: The authors go a bit fast here. While the statement is correct, I had to read it several times to understand the logic. I would advise to add a sentence that explains that the contribution of the prior profile to the final solution increases with decreasing sensitivity of the measurement.

Thanks for the suggestion. We now added a sentence along these lines.

P7827, line 22-25: Assuming xa=0 makes the problem non-Gaussian, because xa cannot be negative. How does this assumption affect the discussion?

The DOAS retrieval is not an optimal estimation technique based on Bayes theorem. So this is not a problem. We refer to Eskes an Boersma [2003] for an extensive discussion of the retrieval theory for DOAS columns.

P7828, line 23: 'whereas satellite measurements provide "snapshots" at a particular local time': this is not a description of spatial error but of temporal error. Please remove.

We removed "at a particular local time" because this may be confusing.

Section 2.3 is a slightly odd section. In section 1 the authors say that section 2 will introduce the issues, while section 2.3 is actually used to describe the proposed solution. There is therefore overlap with text later in the manuscript, which does not help the reader. I suggest to either remove section 2.3 or make it part of text later in the document (e.g., Section 3.2). Also, the end of section 2.2 describes conclusions, which should not be part of this introduction.

Thanks for this excellent suggestion. We have moved section 2.3 to 3.2 and streamlined the text as necessary.

P7830, line 3: Should you not do the opposite in this case, i.e., average over all the grid boxes that fall within the satellite foot print?

That is in principle possible, but at the cost of no longer evaluating the model on its own grid, or a regular grid at all. Such evaluations are sometimes done on a daily basis (e.g. Huijnen et al., ACP, 2010). Ultimately, they still re-grid the satellite footprints to a coarser grid for final seasonal mean comparisons (maps, regional averages). Re-gridding the model-retrieval pairs from the irregular, daily varying satellite grid to a coarser common grid, brings the total number of manipulations on three (1. model-to-satellite, 2. model-at-satellite back to model grid, and 3. satellite to model grid), whereas the recipe in Eq. (3) only requires one such manipulation.

Section 2.4 is slightly strange as well. Of course, one has to be aware of clouds and their impact on the retrieval. But wasn't the whole point of recommending the use of averaging kernels exactly to deal with these kind of issues. Maybe, the authors could mention the impact of clouds in a shortened paragraph earlier in the manuscript and then present the use of averaging kernels later in the manuscript both in the context of vertical sensitivity in general and in the context of clouds.
Thanks for the suggestion. We have now merged the section on the influence of clouds with section 2.1. In that section we also state that it is in principle possible to use the cloudy retrievals as long as the kernels are being applied.

P7832, last paragraph: Treating systematic errors as random is a strange assumption considering the fact that the authors acknowledge later on that the systematic model errors can be as large as 50 %. This needs some further clarification.
We admit that this was formulated inadequately. It is better to interpret Eq. (6) as providing an overall envelop of the model-retrieval comparison error, where the individual (model, retrieval, representativeness) error contributions may have substantial systematic components. While some of these systematic contributions may cancel out in an area-averaged, monthly mean model-retrieval comparison such as presented in Section 6, robust systematic error will persist. We therefore removed the sentence "Here and in Sect. 6 we treat them as random errors, an approach usually followed in retrieval and data assimilation studies (Rodgers, 2000)".

P7833, lines 10 – 12: I am not sure if I understand this argument. Averaging will remove the random component from the error budget, so systematic errors will actually become more dominant. Or do I miss something?
In general, a priori data that is always and everywhere biased in the same way will then indeed become dominant. But a priori data can also suffer from systematic errors that are variable in space and time. Our argument has its origin in the use of a priori data in the retrieval (e.g. surface albedo) that may be systematically biased high for pixel 1, but biased low for pixel 2. Averaging retrievals over a large area, will then tend to cancel the effect of systematic errors in the a priori data, as long as the mean systematic error (in the albedo) is zero.

Section 3.2: Maybe the text of Section 2.3 could be used here.
Done.

Section 4.1: It is unclear what the relationship is between the model profile used for the AMF and the a priori profile (set to zero under the weak absorbing assumption) used in the retrieval. Both prior assumptions play a role in the definition of the end product, but it is not clear how this is accounted for in the averaging kernel.
The a priori profile $\mathbf{x_a}$ used in the air mass factor calculation is typically taken from a CTM. The a priori profile is used in the calculation of the AMF, and also in the averaging kernel, as discussed already in section 2.1, right below Eq. (1). The model that provides $\mathbf{x_a}$ may or may not be the same as the model under evaluation. In our study, the a priori profiles $\mathbf{x_a}$ in the OMI NO2 retrieval are from the TM4 model, and the models that are being evaluated are TM5 and GEOS-Chem. We decided to not go into much detail here, since this has been extensively described in many retrieval papers including our own (Eskes and Boersma [2004], Boersma et al., [2003; 2007; 2011]).

Page 7837, eq. 8: This was already described in Section 2.3

Good point. We removed Eq. (8) and instead refer to Eq. (3) now in the text.

Page 7838, lines 2 – 4: Does this mean that the DOAS error estimates are not correct? If they would be, they should be taken into account. I am not convinced by the argument that the method drives the statistical interpretation of the results.
The DOAS error, calculated from formal error propagation, is our best estimate of the uncertainty in a single retrieval. We stress that here we calculate an area average of individual retrievals that is the best spatial representation of the retrieved column over a large, grid-cell area. We take the individual errors (and their spatial correlation) into account when computing a superobservation error as in Eq. (9) in the manuscript.

Page 7841, line 13: Are the TM5 cloud fractions indeed simulated by the model or do they come from ERA-Interim. If the former is the case, it would be worth mentioning how these cloud fractions are simulated. If it is the latter, this should be mentioned as well.
The TM5 cloud information comes from ERA-Interim, as stated already in Section 4.2 and in 5.2. The details of the cloud fraction simulation are described in Dee et al. [2011].

Section 5.3: Although the presented comparisons do indicate a better agreement between model and observations, it is not rock-solid proof. There remains the possibility that the use of averaging kernels is masking/compensating other errors in the model. This is probably worth mentioning for completeness.
It depends on what is understood as 'proof'. Our point is not that the model is doing better if the kernel is applied, but rather that if the model is sampled as if the satellite would observe the modeled state, the model and retrieval can be properly compared according to the retrieval's capabilities. Such a comparison allows for a better-constrained evaluation of the model. Of course some model errors may still go undetected following such an evaluation, but the evaluation is no longer biased because of the difference in vertical sensitivity of the retrievals vs. the model.

Appendix D is not referenced in the text. However, it contains interesting content that could maybe be used in Section 4.1? See also my earlier comment about Section 4.1.
Thanks for this nice suggestion. We don't think it would be useful to include it in Section 4.1 that is about the retrieval settings. Still, we think referencing Appendix D is a good idea, and we do that now in Section 5.3, because this is where the actual application of the kernel is discussed.

Anonymous Referee #2
The authors have described an approach to quantity three types of errors that can arise when trying to compare modelled NO2 fields with UV/Vis retrievals of column NO2. They have focused on horizontal and vertical representative errors, and errors associate with cloud cover. These are all issues that the community is aware of and is struggling to address using various approaches. In that context, the manuscript is not innovative, but I believe that it will be useful to the community. By providing a coherent approach for dealing with these errors, the manuscript will help reduce misuse of the NO2 column data, and it may even spur new approaches for mitigating the errors. The manuscript is well written and I recommend it for publication after minor revisions to address my comments below.
We thank the reviewer for the constructive comments.

Comments
1) Page 7828, line 25, and Page 7840, line 8 (title of Section 5.2): It is not clear to me what is the temporal component here. It seems to me that this error, which is the focus of Section 5.2, is really an issue with the representativeness of the cloud cover. Referring to this as temporal (or meteorological) representativeness error is vague and confusing. Why not just call it representativeness errors in cloud cover?
On page 7828, we removed the adjective 'meteorological' to clarify that this category of errors is addressing temporal aspects that not just relate to cloud cover. Temporal representativeness errors arise from models not properly representing the temporal evolution associated with

changing cloud cover or with changing emissions such as the weekend effect, so just labelling them as 'representativeness errors in cloud cover' would not be a complete description. The title of section 5.2 is in our opinion appropriate as it is.

2) Page 7832, lines 21-23: Here the authors state that they treat all of the errors as random errors. However, in Section 5.1 and Appendix B they account for a correlation in the errors, so clearly they are accounting for some systematic errors. This confusing and needs to be better explained.
We removed the sentence that we treat the retrieval, modelling, and representativeness errors as random errors in response to Rev#1. In section 5.1, we discuss how to calculate representativeness errors, taking into account error correlation, so that combined observation and (spatial) representativeness may be obtained. In many data assimilation studies, representativeness errors are included in the observation errors, as already stated in our original manuscript, and now also stated right after introducing Eq. (3).

3) Page 7833, lines 8-12: I do not agree with the statement that the vertical transport errors will be smaller, in an average sense, when aggregated over a month. That is, for example, unlikely to be the case in the tropics or over eastern North America in summer, when convective transport is strong. Indeed, the monthly mean differences in the vertical distribution of NO2 in the lower troposphere between GEOS-Chem andTM5 shown in Figure 7a are not that different from those shown in Figure 6a for Feb 18th.
They will be smaller in a monthly average sense, because not all days in the month will have strong convective activity, especially in a region like North America. Furthermore, Figures 6 and 7 do not necessarily indicate that GEOS-Chem suffers from vertical transport errors but rather that GEOS-Chem and TM4 simulate vertical transport of NO2 differently.

4) Page 7836, line 17: Are there additional references besides Lin (2012) that should be included here? What about Lamsal et al. (2010, J. Geophys. Res., D05302, doi:10.1029/2009JD013351)?
We have now included the Lamsal et al. [2010] reference, also in a number of other occasions where we thought it appropriate.

5) Page 7845, lines 2-6: Not all of the NO2 products provide averaging kernels. It would be helpful if the authors could explain what one should do to mitigate the vertical representativeness errors in the case when averaging kernels are not provided.
There is not much one can do except but go back to the NO2 product developers and make a request to make a available the averaging kernels or the scattering weights for each pixel in the product. Using the averaging kernel of another product would not help, unless all the a priori data used in calculating the averaging kernels (albedo, cloud parameters, atmospheric profile) is exactly the same, which is highly unlikely to ever happen.

6) Figure 7 caption: Explain that the numbers given in each panel, e.g. 5.75 for GEOS-Chem, are the integrated column abundances.
Done, also for Figure 6.

7) Figure 10 caption: Should "100% x (B/C – 1)" be "100% x (B/A – 1)"? Why not make these differences relative to experiment A, which is believed to be better, so that for C vs A is it 100% x (C/A – 1) and for B vs A it is 100% x (B/A – 1)?
Thanks for spotting this. The experiments were indicated relative to experiment A. We adapted the caption accordingly.

Representativeness errors in comparing chemistry transport and chemistry climate models with satellite UV/Vis tropospheric column retrievals

K. F. Boersma[1,2], G. C. M. Vinken[3], and H. J. Eskes[1].

[1]{Royal Netherlands Meteorological Institute, De Bilt, The Netherlands}

[2]{Wageningen University, Meteorology and Air Quality department, Wageningen, The Netherlands}

[3]{Eindhoven University of Technology, Eindhoven, The Netherlands}

Correspondence to: K. F. Boersma (boersma@knmi.nl)

Abstract

UV/Vis satellite retrievals of trace gas columns of nitrogen dioxide (NO_2), sulphur dioxide (SO_2), and formaldehyde (HCHO) are useful to test and improve models of atmospheric composition, for data assimilation, air quality hindcasting and forecasting, and to provide top-down constraints on emissions. However, because models and satellite measurements do not represent the exact same geophysical quantities, the process of confronting model fields with satellite measurements is complicated by representativeness errors, which degrade the quality of the comparison beyond contributions from modelling and measurement errors alone. Here we discuss three types of representativeness errors that arise from the act of carrying out a model-satellite comparison: (1) horizontal representativeness errors due to imperfect collocation of the model grid cell and an ensemble of satellite pixels called superobservation, (2) temporal representativeness errors originating mostly from differences in cloud cover between the modelled and observed state, and (3) vertical representativeness errors because of reduced satellite sensitivity towards the surface accompanied with necessary retrieval assumptions on the state of the atmosphere. To minimize the impact of these representativeness errors, we recommend that models and satellite measurements be sampled as consistently as possible, and our paper provides a number of recipes to do so. A practical confrontation of tropospheric NO_2 columns simulated by the TM5 chemistry transport model

(CTM) with Ozone Monitoring Instrument (OMI) tropospheric NO_2 retrievals suggests that horizontal representativeness errors, while unavoidable, are limited to within 5-10% in most cases and of random nature. These errors should be included along with the individual retrieval errors in the overall superobservation error. Temporal sampling errors from mismatches in cloud cover, and, consequently, in photolysis rates, are on the order of 10% for NO_2 and HCHO, and systematic, but partly avoidable. In the case of air pollution applications where sensitivity down to the ground is required, we recommend that models should be sampled on the same mostly cloud-free days as the satellite retrievals. The most relevant representativeness error is associated with the vertical sensitivity of Ultraviolet-visible (UV/Vis) satellite retrievals. Simple vertical integration of modelled profiles leads to systematically different model columns compared to application of the appropriate averaging kernel. In comparing OMI NO_2 to GEOS-Chem NO_2 simulations, these systematic differences are as large as 15-20% in Summer, but, again, avoidable.

1 Introduction

Chemistry transport models (CTMs) are increasingly being evaluated with satellite column retrievals from UV/Vis solar backscatter satellite instruments. Satellite retrievals of trace gas concentrations constitute a rich source of information on key tropospheric species such as nitrogen dioxide (NO_2), sulphur dioxide (SO_2), and formaldehyde (HCHO) that is beginning to be exploited on an ever-larger scale. Ultraviolet-visible (UV/Vis) satellite observations are being used to:

- evaluate the capability of models to simulate atmospheric concentrations of various species (e.g. Uno et al. [2007], Herron-Thorpe et al. [2010], Huijnen et al. [2010[a]]),
- drive data assimilation experiments aimed at improving estimates of the atmospheric state (e.g. Wang et al. [2011], Inness et al. [2013], Miyazaki et al. [2014]),
- provide constraints on uncertain model inputs such as emission inventories through inverse modelling (e.g, Wang et al. [2007], Müller et al. [2008], Mijling and van der A [2012], Barkley et al. [2013]), and to identify new emissions sources (for instance newly built power plants [Zhang et al., 2009]),
- test processes influencing the lifetime of crucial chemical species (e.g. Schaub et al., 2007; Lamsal et al., 2010; Beirle et al., 2011; Stavrakou et al., 2013), or, more

broadly, the chemical regime of the atmosphere (e.g. Martin et al., 2004; Duncan et al., 2010)

When comparing model simulations to satellite measurements, both modelling errors and measurement errors are usually taken into account. Measurement errors are often reasonably well characterized (e.g. Boersma et al., 2004; De Smedt et al., 2008; Lee et al., 2009), but modelling errors are more difficult to establish, because of the large number of uncertain model processes, uncertain boundary (e.g. emissions) and initial conditions, and unresolved or misrepresented aspects of atmospheric physics and chemistry. Modelling errors are best characterized by comparing model simulations to observations. Unfortunately, the observations available for such comparisons are mostly limited in vertical range and regional coverage such as in the case of ground-based networks, or they are merely sporadic in space and time, such as for aircraft campaigns. Satellite data records are based on robust retrieval methods, provide global coverage, and cover decadal time spans. Satellite data has recently been successfully used for dedicated modelling error studies (e.g. Lin et al., 2012; Stavrakou et al., 2013).

When using satellite data, modellers need to be aware that most UV/Vis-retrievals generally contain little information on the vertical distribution of a species (the exception is stratospheric ozone profile retrieval in the far UV of the spectrum, but this species will not be considered in this study). Here we focus on the application of tropospheric UV/Vis retrievals, and we limit ourselves to retrievals of tropospheric species NO_2, SO_2, and HCHO for comparison with models. These species are all relatively short-lived and their retrievals are generally based on differential optical absorption spectroscopy (DOAS, Platt and Stutz [2008]). DOAS retrievals in the UV/Vis match relevant absorption cross section spectra to the solar backscatter spectrum measured by the satellite instrument in order to infer the column integral (slant column density, expressed in molecules cm^{-2}) of a species along the effective atmospheric photon path. The subsequent retrieval step requires the conversion of the slant column density into a vertical column density, and this conversion depends on knowledge (assumptions) of the state of the atmosphere, e.g. on the presence of clouds and aerosols, the vertical distribution of the species, and surface properties. When these assumptions are very different from the atmospheric state modelled by a chemistry transport model (CTM), this will lead to inflated differences between modelled (by, say, CTM 1) and retrieved columns (aided by CTM 2). Such differences, however, can be avoided or in any case minimized, if the

user of satellite data accounts for the representativeness and averaging kernels of the satellite data while interpreting model simulations. Representativeness here is defined as the context in which the satellite measurement holds, i.e. the horizontal coverage, the temporal representativeness, and the vertical information content of the retrieval. It is the goal of this study to provide guidelines on how users can take the representativeness of the UV/Vis column retrievals into account when comparing CTM simulations to satellite retrievals, and by how much the model-retrieval differences would inflate if aspects of representativeness are neglected.

In Sect. 2, we introduce the definitions and terminology for sources of error in the comparison of models and observations, and relate these to what is common practice in the data assimilation community. In doing so, we follow the notation proposed by Ide et al. [1997], also used in relevant work by Rodgers and Connor [2003] and Migliorini et al. [2008]. Section 3 will give an overview of the common features shared by various UV/Vis retrievals with an emphasis on the assumptions made in the retrieval approach that are relevant to modellers and other data users, and it provides a recipe for constructing an appropriate observation operator. Section 4 introduces the TM5 and GEOS-Chem models that we will evaluate to demonstrate the nature and magnitude of representativeness errors. In Sect. 5, we discuss the error budgets associated with a confrontation of CTM simulations with satellite measurements, and, in particular, how the representativeness errors contribute to that budget. Section 6 presents the result of a practical assessment of representativeness errors made when comparing global CTM simulations of tropospheric NO_2 to satellite measurements from the Ozone Monitoring Instrument, and provides recommendations on how to minimize these.

2 Comparing models and UV/Vis satellite measurements

2.1 UV-VIS satellite retrievals

Over the last two decades, tropospheric NO_2, SO_2, and HCHO columns have been retrieved from measurements by the GOME, SCIAMACHY, OMI and GOME-2 (on Metop-A and Metop-B) satellite sensors. The retrievals generally use a two-step approach, based on the DOAS-technique. In step 1, the reflectance spectra measured by the satellite instruments are modelled with a fitting routine that accounts for the spectral signatures from trace gas absorption, inelastic scattering, and (broadband) Rayleigh, Mie, and surface scattering. For each of the above species, spectral regions are selected where the absorption structures are

most distinct, and spectral interference from other species is minimal. The species' slant column density is then calculated from the inferred absorption in combination with knowledge of the species' absorption cross-section. Before converting the slant column densities into tropospheric vertical columns, background corrections may be required to account for the fact that a portion of the slant column has originated from the species' absorption of light in the stratosphere.

In step 2 of the retrieval, the tropospheric slant column densities are converted into vertical column estimates, using a radiative transfer (forward) model and forward model parameters, that influence the retrieval. For DOAS UV/Vis retrievals, forward model parameters typically include the sensor viewing geometry, and best estimates of the surface albedo, terrain height, cloud and aerosol properties (or an effective representation thereof), as well as the a priori vertical distribution of the species ($\mathbf{x_a}$) of interest. The radiative transfer calculations are expressed as so-called air mass factors, defined as the (forward) modelled ratio of slant (N_S) and vertical columns (N_V), given the set of forward model parameters: $M = N_S/N_V$. Tropospheric air mass factors have been shown to be very sensitive to choices for surface albedo, for cloud correction, and for a priori vertical distribution, and, consequently, air mass factor uncertainties are large, and dominate the retrieval error budget for tropospheric columns (e.g. Boersma et al. [2004], Millet et al. [2006], Lee et al. [2009]).

Data users need to be aware of the important role played by clouds in UV/Vis-retrievals. With the exception of elevated plumes resulting from volcanoes, lightning, and aircraft, most tropospheric NO_2, SO_2, and HCHO generally resides in the lower atmosphere, close to their surface sources. Clouds thus typically obscure the absorbing species from (satellite) view, leading retrieval groups to advise against the use of their satellite data when taken under cloudy conditions. Trace gas retrievals under cloudy situations suffer from larger errors (e.g. Schaub et al. [2006]), because the detectable fraction corresponds to the column above the cloud, leaving a so-called 'ghost column' below the cloud to be added somehow. Because ghost columns are generally taken from climatology or a CTM, they do not contribute to the measured information in any way, so that inclusion of columns under cloudy situations compromises a model – satellite comparison, unless the averaging kernels are taken into account [Schaub et al., 2006]. ~~This does not mean that satellite measurements taken under cloudy conditions should not be used at all.~~ In data assimilation systems, cloudy measurements still provide valuable information on the abundance and vertical information of

trace gases above the cloud, for instance for constraints on e.g. lightning-produced NO_2 [Boersma et al., 2005] and in recent cloud-slicing techniques [Choi et al., 2014; Belmonte-Rivas et al., 2015]. ~~In cloud-covered situations, it is essential to take the averaging kernels into account.~~

DOAS UV/Vis nadir retrievals are characterized by a vertical sensitivity that generally reduces with increasing atmospheric pressure, and require an a priori vertical profile of the species $\mathbf{x_a}$ to interpret the slant column (e.g. Palmer et al. [2001]; Richter et al. [2006]). Because Rayleigh scattering of sunlight is more effective in the UV, fewer photons reach the lower atmosphere in the spectral range where SO_2 has distinct absorption spectral features (300-330 nm), compared to the spectral windows for HCHO (340-360 nm) or NO_2 (400-500 nm). This implies that the measurement sensitivity to species in the lower atmosphere is lowest for SO_2, followed by HCHO, and highest for NO_2. The contribution of the a priori profile to the retrieved column increases with decreasing sensitivity of the measurement. Uncertainty in the species a priori vertical profile thus propagates stronger for SO_2 (up to 22% error [Lee et al., 2009]), and somewhat less for NO_2 (10-15% error, e.g. Hains et al. [2010]; Vinken et al. [2014]).

This a-priori profile error contribution to model-satellite comparisons can be eliminated by application of the averaging kernel to the model output (Eskes and Boersma [2003]; Boersma et al. [2004]; Rodgers and Connor [2003]). The averaging kernel for UV/Vis retrievals describes the relationship between the true column and the estimated, or retrieved column $\hat{y}_o$ where the hat denotes that the retrieval represents an estimated value of the true column:

$$\hat{y}_o = \mathbf{A} \cdot \mathbf{x_{true}} \quad (1)$$

with $\mathbf{A}$ the averaging kernel whose discretized elements can be described as $A_l = \frac{m_l}{M(x_a)}$, with m_l the scattering weights [Palmer et al., 2001], or box air mass factors for layer l (see Eskes and Boersma [2003], and Boersma et al. [2004] for more detail). Note that the retrieval problem has been linearised around $\mathbf{x_a} = 0$, related to the weak absorber character of the species, which implies that the a-priori state does not explicitly appear in Eq. (1).

2.2 Model evaluation with UV/Vis satellite retrievals

A comparison between satellite measurements $\hat{y}_o$ (e.g. the retrieved tropospheric NO_2 columns within a model grid cell), and the model state $\mathbf{x_m}$ (e.g. the modelled vertical NO_2

distribution in the troposphere), in the form of measurement-minus-model departures (d) is expressed as:

$$d = \hat{y}_o - \mathbf{Hx_m} \quad (2)$$

with **H** the observation operator that describes the relation between the observed data and the modelled state. Apart from the observation errors (σ_o in the following) and the modelling errors (σ_m), we also need to take into account representativeness errors (σ_r) associated with the fact that model simulations and satellite measurements provide different representations of a geophysical quantity. We generalise the representativeness errors as the errors introduced in a satellite-to-model evaluation by an incorrect description of the relation between the grid-cell mean concentrations and the satellite retrieval(s), i.e. we can think of them as errors in the observation operator **H**. In data assimilation, representativeness errors are normally included in the observation errors (e.g. Jones et al., [2003]; Miyazaki et al., [2012]).

Substantial representativeness error may arise when the observation operator **H** is simplified and the model is not sampled in a manner fully consistent with the satellite observation. We can identify three types of representativeness errors associated with model-satellite comparisons:

- Spatial representativeness errors. Such errors will arise because models provide a spatially smoothed representation of the atmospheric state, whereas satellite measurements provide 'snapshots' ~~at a particular local time~~, and often resolve variability at scales (pixels) smaller than the model grid cell.
- Temporal ~~and meteorological~~ representativeness errors. In applications focusing on clear-sky situations such as emission estimates, failure to sample the model for the same clear-sky conditions and overpass time as the satellite measurements, will lead to systematic sampling errors.
- Vertical representativeness errors. Because the sensitivity of the UV/Vis satellite retrievals is altitude-dependent (Palmer et al., [2001]), UV/Vis retrievals should be regarded as estimates of the state weighted by the averaging kernel (Eskes and Boersma [2003]). Neglecting the averaging kernel or vertical sensitivity of the retrieval in the comparison will inevitably introduce additional representativeness errors to the comparison in Eq. (2).

To minimize these representativeness errors in comparing CTMs and satellite measurements, we recommend to follow the recipe given in Section 3.2 ~~the box inset on this page~~.

This recipe on how to compare a CTM with satellite observations is a set of mathematical operations on satellite and model data. This is particularly relevant for short-lived species that have a high spatial and diurnal variability such as NO_2, SO_2, and HCHO (e.g. Boersma et al. [2008], Vrekoussis et al. [2009], Barkley et al. [2013]). Details of the approach may differ (e.g. spatial interpolation of the model state to the location of the pixel, averaging over different model times close to the satellite measurement time, replacing the a priori profile with the model profile in the retrieval), as long as the general principle of consistent sampling is observed. We advise against a comparison of the original satellite column (retrieved with a

~~**Recipe for minimizing representativeness errors**~~

~~1. The first step in comparing satellite observations to model simulations is to ensure that the satellite measurements are spatially representative for the area of the model grid cell. This is achieved by calculating the weighted average of all individual retrievals $\hat{y}_i^o$ within the superobservation model grid cell over the entire area covered by all (valid) retrievals, where the weight is given by the pixel area w_i (in km^2):~~

$$\hat{y}_o = \frac{\sum_i w_i \hat{y}_i^o}{\sum_i w_i} \quad \text{(B1)}$$

~~If the model grid cell happens to be smaller than the satellite pixel, Eq. (B1) will reduce to $\hat{y}_o = \hat{y}_1^o$ for grid cells that are completely overlapped by a single satellite pixel (w_i=1).~~

~~2. The second step is to sample the CTM field sequence $\mathbf{x_m}[t]$, here expressed as a discrete series of periodic fields with t an integer, when model time t is closest to the satellite overpass time t_o.~~

$$\mathbf{x_m} = \mathbf{x_m}[t]\delta[t], \text{ with } \delta[t] = \begin{cases} 0, & t \neq t_o \\ 1, & t = t_o \end{cases} \quad \text{(B2)}$$

~~The model sequence is sometimes also sampled with somewhat looser criteria, by requiring that the absolute model-satellite time difference stays within 1-2 hours (e.g. Martin et al. [2003]).~~

~~3. The third step is to apply the averaging kernel on the model vertical distribution $\mathbf{x_m}$ to obtain the model estimate $\hat{y}_m$ that can be directly compared to the observed state $\hat{y}_o$:~~

$$\hat{y}_m = \boldsymbol{A}\,\mathbf{x_m} = \sum_{i=1}^{L} A_i S_i x_{m,i} \quad \text{(B3)}$$

~~where S_i are the components at the i-th vertical layer of an operator that executes a mass-conserving vertical interpolation or integration followed by a conversion to sub-columns (molec.cm^{-2}) in case the model vertical distribution $x_{m,i}$ is not yet given in those units. The product of the mathematical expressions (B2) and (B3) forms the observation operator $\mathbf{H}$ in Eq. (2), which describes the relation between the superobservation and the modelled state.~~

priori profile $\mathbf{x_a}$) to the model column $\hat{x}_m$ because in that case differences between the a priori and modelled vertical profiles would inflate the overall error d, see Sect. 6 and recommendations in Sect. 2.3 of Boersma et al. [2004], and Duncan et al. [2014].

~~2.3 The role of clouds in UV/Vis retrievals~~

~~Data users need to be aware of the important role played by clouds in UV/Vis-retrievals. With the exception of elevated plumes resulting from volcanoes, lightning, and aircraft, most tropospheric NO_2, SO_2, and HCHO generally resides in the lower atmosphere, close to their surface sources. Clouds thus typically obscure the absorbing species from (satellite) view, leading retrieval groups to advise against the use of their satellite data when these are taken under cloudy conditions. Trace gas retrievals under cloudy situations suffer from larger errors (e.g. Schaub et al. [2006]), because the detectable column then corresponds to the column above the cloud, leaving a so-called 'ghost column' below the cloud to be added. Because ghost columns are generally taken from climatology or a CTM, they do not contribute to the measured information in any way, so that inclusion of columns under cloudy situations compromises a model – satellite comparison, unless the averaging kernels are taken into account [Schaub et al., 2006].~~

~~This does not mean that satellite measurements taken under cloudy conditions should not be used at all. In data assimilation systems, for above-cloud constraints on e.g. lightning-produced NO_2 [Boersma et al., 2005], but also in recent cloud-slicing techniques [Choi et al., 2014; Belmonte-Rivas et al., 2015], cloudy measurements provide valuable information on the abundance and vertical information of trace gases above the cloud. In cloud-covered situations, it is essential to take the averaging kernels into account.~~

3 Theoretical model evaluation error budget

3.1 Sources of errors in evaluating CTMs with UV/Vis retrievals

A comparison between model simulations and satellite retrievals begins with a comparison of their theoretical capabilities. A model-satellite comparison will be influenced by:

1. modelling errors σ_m, related to an incomplete knowledge and description of the atmospheric state $\mathbf{x_m}$,
2. retrieval errors σ_o, because of instrument noise and uncertainty in the (external) forward model parameters, and

3. representativeness errors σ_r, arising from fundamental differences between the atmospheric sampling by models and satellites, i.e. errors in the observation operator **H**.

Assuming that these error terms are independent, the error analysis for a satellite-model column difference $(\hat{y}_o - \mathbf{Hx_m})$ can be written as:

$$\sigma = (\sigma_o^2 + \sigma_m^2 + \sigma_r^2)^{1/2} \quad (3)$$

with σ_o^2 the best estimate for the (relative) column retrieval errors, σ_m^2 for the (relative) modelling error, and σ_r^2 the contribution to the error arising from the act of carrying out the comparison itself (i.e. from errors in the observation operator). Some studies (e.g. Jones et al. [2003]) include representativeness errors in the observation errors. Below we will show that representativeness errors may contribute substantially to the overall error in satellite-model confrontations.

The retrieval, modelling, and representativeness errors will all have systematic and random components. ~~Here and in Sect. 6 we treat them as random errors, an approach usually followed in retrieval and data assimilation studies [Rodgers, 2000].~~ In principle, one would like to distinguish between the random and systematic contributions, but in practice this is very complicated, because many systematic contributions to retrieval and model errors are only weakly correlated in space and time. Examples of subtle systematic retrieval effects are errors in individual albedo values with a small spatial correlation length but with 100% correlation in time (for instance because residual cloud effects in the albedo climatology are strongly variable from one location to the other [Kleipool et al., 2008]). When averaged over a larger region such as the spatial extent of a coarse model grid cell, the impact of such errors tends to reduce. Likewise, models will suffer from systematic errors in for instance the description of vertical transport. In particular circumstances, such as strong, small-scale convective activity, such errors tend to be acute, but in an average sense, such as comparisons aggregated over a month and a region, we may expect these errors to be smaller.

3.2 Recipe for minimizing representativeness errors

1. The first step in comparing satellite observations to model simulations is to ensure that the satellite measurements are spatially representative for the area of the model grid cell. This is achieved by calculating the weighted average of all individual retrievals $\hat{y}_i^o$ within the

superobservation model grid cell over the entire area covered by all (valid) retrievals, where the weight is given by the pixel area w_i (in km^2):

$$\hat{y}_o = \frac{\sum_i w_i \hat{y}_i^o}{\sum_i w_i} \quad (4).$$

If the model grid cell happens to be smaller than the satellite pixel, Eq. (4) will reduce to $\hat{y}_o = \hat{y}_1^o$ for grid cells that are completely overlapped by a single satellite pixel (w_1=1).

2. The second step is to sample the CTM field sequence $\mathbf{x_m}[t]$, here expressed as a discrete series of periodic fields with t an integer, when model time t is closest to the satellite overpass time t_o.

$$\mathbf{x_m} = \mathbf{x_m}[t]\delta[t], \text{ with } \delta[t] = \begin{cases} 0, & t \neq t_o \\ 1, & t = t_o \end{cases} \quad (5)$$

The model sequence is sometimes also sampled with somewhat looser criteria, by requiring that the absolute model-satellite time difference stays within 1-2 hours (e.g. Martin et al. [2003]).

3. The third step is to apply the averaging kernel on the model vertical distribution $\mathbf{x_m}$ to obtain the model estimate $\hat{y}_m$ that can be directly compared to the observed state $\hat{y}_o$:

$$\hat{y}_m = \boldsymbol{A}\,\mathbf{x_m} = \sum_{l=1}^{L} A_l S_l x_{m,l} \quad (6)$$

where S_l are the components at the l-th vertical layer of an operator that executes a mass-conserving vertical interpolation or integration followed by a conversion to sub-columns (molec. cm^{-2}) in case the model vertical distribution $x_{m,l}$ is not yet given in those units. The product of the mathematical expressions (5) and (6) forms the observation operator $\mathbf{H}$ in Eq. (2), which describes the relation between the superobservation and the modelled state.

3.3 Representativeness errors in evaluating CTMs with UV/Vis retrievals

The total representativeness error σ_r is composed of horizontal representation errors, (temporal model) sampling errors, and vertical smoothing errors, and these three contributions may be assumed to be largely uncorrelated:

$$\sigma_r = (\sigma_h^2 + \sigma_t^2 + \sigma_v^2)^{1/2} \quad (7)$$

For an appropriate comparison between model simulations and satellite retrievals, it is important to sample the CTM as closely as possible to the satellite's sampling of the

atmosphere (see Sect. 3.2). These may seem like trivial conditions for comparison, yet one or more of these conditions are often violated.

4 Data used in this study

4.1 Satellite data

In this study, we use tropospheric NO_2 retrievals from the Dutch OMI NO_2 (DOMINO) algorithm v2.0 [Boersma et al., 2011]. These retrievals proceed along the lines discussed above, with spectral fitting of NO_2 in the 405-465 nm window [van Geffen et al., 2014], data assimilation of the NO_2 slant columns in the TM4 chemistry transport model [Williams et al., 2009] to estimate the stratospheric background [Dirksen et al., 2011], and final conversion of the tropospheric slant columns with air mass factors based on radiative transfer calculations with the DAK model. In the DOMINO algorithm, altitude-dependent AMFs are interpolated from pre-calculated look-up tables using the best available information on the satellite viewing geometry, surface albedo [Kleipool et al., 2008], terrain height (3km-resolution elevation data provided with Aura data). Subsequently, the local altitude-dependent AMFs are combined with the predicted local vertical NO_2 distributions (from TM4), to produce the (tropospheric) air mass factors. The air mass factor step also includes a correction for the temperature-dependency of the NO_2 absorption cross-section [Boersma et al., 2004], because only the 220 K cross section is used in the spectral fit. The DOMINO v2.0 data have been evaluated in a number of validation exercises (e.g. Irie et al. [2012], Ma et al. [2013], Lin et al. [2014]), showing their quality and use, although a number of relevant improvements is planned and currently being implemented (Maasakkers [2013], van Geffen et al. [2014]). DOMINO v2.0 has been used in many applications and model studies (e.g. Stavrakou et al. [2013], Castellanos et al. [2014], McLinden et al. [2014], Verstraeten et al. [2015]), which makes the data product well-suited for evaluating satellite-to-model comparisons and the errors associated with such comparisons, which is the purpose of this study.

Chemistry-transport models (CTMs) are the central tools to simulate tropospheric concentrations of NO_2, SO_2, and HCHO, and to help interpret and use satellite measurements of these species. For the short-lived species studied here, previous studies indicate modelling biases of ±20-30% for NO_2 (e.g. van Noije et al. [2006]), and 20-50% for HCHO (e.g. Dufour et al. [2009]; Williams et al. [2012]) over regions with substantial pollution.

4.2 TM5

We use the TM5, the global 3-D CTM version 3.0 [Huijnen et al., 2010] with a grid of 3° longitude × 2° latitudes × 34 vertical layers, and a model top at 0.1 hPa [Krol et al., 2005]. The TM5 model is used in many studies for atmospheric chemistry (e.g. Williams et al. [2014]), aerosol haze (e.g. von Hardenberg et al. [2012]), data assimilation, and inversion applications (e.g. Hooghiemstra et al. [2012]; Krol et al. [2013]). The model is driven by ERA-Interim meteorological reanalysis data from the European Centre for Medium Range Weather Forecats (ECMWF) [Dee et al., 2011] and the base time step is 1 hour. In the version used here, TM5 operates with Carbon Bond Mechanism 4 chemistry [Gery et al., 1989] to describe the production of ozone, hydrogen oxide radicals (HO_x=OH+HO_2) and oxidation of nitrogen oxides (NO_x=NO+NO_2), SO_2, and volatile organic compounds (VOCs), with 40 species, 64 gas-phase, and 16 photolysis reactions. In TM5, SO_2 is oxidized in clouds and on aerosols, and nighttime hydrolysis of N_2O_5 into nitric acid (HNO_3) is parameterized with a global mean uptake coefficient of 0.02 following recommendations by Evans and Jacob [2005]. NO_x emissions are from the RETRO inventory for the anthropogenic sectors (Regional Emission inventory in ASia – REAS for Asia) with a total of 33 Tg N/yr, 9 Tg N/yr from soil, 5 TgN/yr from biomass burning (from the Global Fire Emissions Database v2 (GFED2) van der Werf et al. [2006]), and 6 TgN/yr for lightning. Global anthropogenic SO_2 emissions are taken from the AeroCom project at 108 Tg SO_2/yr [Dentener et al., 2006]. Biogenic VOC emissions, including the important HCHO and its precursor isoprene, are from the ORCHIDEE database [Lathière et al., 2006], and are 10 TgC/yr for HCHO and 565 Tg C_5H_8/yr for isoprene. We simulated the year 2006 with a one-year spin-up.

TM5 simulations of NO_2 and HCHO have been evaluated by Huijnen et al. [2010] and Williams et al. [2012]. These studies indicate that tropospheric NO_2 columns in TM5 are 20-30% low compared to DOMINO v2.0 columns, but the model captures the seasonality, and shows realistic vertical distributions of NO_2 relative to INTEX-B aircraft measurements. TM5 captures the seasonality of HCHO tropospheric columns but also overestimates these columns by 0-50%, partly because of inadequate photolysis rates in the model [Williams et al., 2012].

4.3 GEOS-Chem

We also use the GEOS-Chem model, v9-02i, with a grid of 2.5° longitude × 2° latitude × 47 vertical layers, and the model top at 80 km. The GEOS-Chem model is a CTM in use by a

large community of scientists for a wide range of applications including, shipping NO_x plume-in-grid chemistry [Vinken et al., 2011], and estimating isoprene and ammonia emissions (e.g. Millet et al. [2008]; Paulot et al., [2014]). GEOS-Chem is driven by GEOS-5 meteorological fields from NASA GMAO, with a time step of 30 minutes. As TM5, GEOS-Chem uses a condensed O_3-NO_x-HO_x-VOC-aerosol chemistry scheme (described in Mao et al. [2010] and references therein). The standard chemistry scheme has 66 species, and 236 chemical reactions. GEOS-Chem takes into account heterogeneous chemistry on aerosol and cloud particles [Mao et al., 2010], including the uptake of N_2O_5 on aerosols leading to nighttime HNO_3 formation following the parameterization by Evans and Jacob [2005]. Anthropogenic NOx emissions are from the global EDGAR 3.2FT2000 inventory [Olivier and Berdowski, 2000], but these are replaced by regional inventories over various continents. Other NO_x emissions in GEOS-Chem include soil, lightning, biomass burning, biofuel, aircraft and ship, resulting in a global total source of 51.5 TgN/yr for 2006 (similar to TM5 with 53 TgN/yr for the same year). A two-year spin-up was performed (2004-2005), and GEOS-Chem output was stored for the year 2006. For more details on the GEOS-Chem simulation, see Vinken et al. [2014].

GEOS-Chem simulations of tropospheric NO_2 columns have been evaluated before by Lamsal et al. [2010] and Lin [2012], who found, similar to the TM5 evaluation discussed above, that the model underestimates tropospheric NO_2 by 20-35% (over China). Zhang et al. [2012], in a study targeting nitrogen deposition over the United States, found excellent agreement between the modelled and OMI-observed spatial distribution of tropospheric NO_2, but also underestimates of 10% in the northeastern US, and 40% locally in southern California, were also evident.

5 Representativeness errors

5.1 Horizontal representativeness errors

If the complete spatial extent of a model grid cell is covered with valid retrievals, a good comparison is straightforward because a spatially fully representative area average can be calculated. For partly covered cases, the difficulty lies in estimating the magnitude of the (horizontal representativeness) errors associated with limited coverage of a model grid cell. One way to calculate a representative grid cell average is by averaging all valid satellite observations that were taken within the boundaries of the grid cell within a given model time

step, as in Eq. (3), with w_i the fractional grid cell coverage defined as A_{pixel}/A_{cell} with A_{pixel} the area (in km^2) covered by the fraction of the satellite pixel that falls within the boundaries of the model grid cell with area A_{cell} (in km^2). In this manner, one obtains a 'superobservation' that may be considered as representative for the grid cell average (Dirksen et al. [2011]; Miyazaki et al. [2013]). In some model-satellite confrontations, the number of satellite retrievals is thinned out to 1 per grid cell, but we advise against such an approach in view of the strong sub-grid variations and the considerable errors in individual measurements. In many global applications, the spatial resolution of the model is coarser than the resolution of the satellite observations. ~~In that case, we recommend computing a set of superobservations defined as the grid cell average trace gas column $\overline{N}$:~~

$$\overline{N} = \frac{\sum_{i=1}^{n} w_i N_i}{\sum_{i=1}^{n} w_i} \quad (5)$$

~~with w_i the fractional grid cell coverage of retrieval N_i, defined as A_{pixel}/A_{cell} with A_{pixel} the area (in km^2) covered by the fraction of the satellite pixel that falls within the boundaries of the model grid cell with area A_{cell} (in km^2), and n the total number of valid retrievals within the cell.~~ We caution against applying additional weighting by the individual retrieval errors in Eq. (4). Because, by nature of the DOAS approach, retrieval errors are largest for large column values (see e.g. Boersma et al. [2004]), error weighting would skew the average to the lower values in the distribution. The measurement error for superobservations can be calculated from area-weighting the individual pixel errors $\sigma_{o,i}$ to provide an area-weighted average (statistical) retrieval error σ, and by accounting for a partial correlation in the errors between pixels as in Eskes et al. [2003] (see Appendix B for a derivation):

$$\sigma_o = \sigma \sqrt{\frac{1-c}{n} + c} \quad (8)$$

with the second term on the right side representing the error correlation (c) between the n retrievals. Miyazaki et al. [2012] propose c=0.15, based on the consideration that errors in clouds, albedo, a priori profile, and aerosol in retrievals are typically correlated in space, but they acknowledge that the exact number is difficult to estimate.

Some studies take a different approach than the superobservations proposed in Eq. (4) and interpolate the model simulations to the centre of a satellite pixel, but the difficulty with this approach is the questionable spatial representativeness of the interpolated model value, especially if the model grid cells cover a larger area than the satellite pixels.

Both individual pixel errors and representativeness errors contribute to the total error in the superobservation. Following Miyazaki et al. [2012], we calculate the horizontal representativeness error σ_r as a function of the total fractional coverage achieved by all valid pixels by random reduction of the number of retrievals used to calculate the mean grid cell value. For homogeneous scenes with little variability of NO_2, SO_2, or HCHO, such errors will obviously be small. But for grid cells covering strong inhomogeneous sources of air pollution, such as megacities or coal plants, we may expect the area average to depend strongly on the spatial sampling. Figure 1 illustrates the horizontal representativeness error as a function of total fractional coverage for one polluted model grid cell, here taken over the eastern United States (greater New York City), at two resolutions, i.e. 3° × 2° (typical for a global CTM) and 0.5° × 0.5° (regional CTM). To calculate the horizontal representativeness error, we randomly reduced the number of pixels *n* in Eq. (4) first by 1, then by 2, and so on, until there was only one pixel left, to obtain new estimates $\hat{y}_o'$. We repeated this 100 times and interpret the root mean squared difference with the original $\hat{y}_o$ as the horizontal representativeness error, which is zero in situations of full coverage. Complete coverage of the grid cell is typically achieved by more than 100 OMI pixels in the case of 3° × 2° resolution grid cells, and by ±5 pixels[1] for 0.5° × 0.5°. The horizontal representativeness errors appear higher for the 0.5° × 0.5° than for the 3° × 2° grid cell, due to the smaller sample (*n*=5) size and the strong spatial gradients over the central New York area for the higher resolution model. For models with higher spatial resolution (0.5° × 0.5°), there is less tolerance for reduced area coverage over strongly inhomogeneous areas such as central New York, as indicated by the steeper representativeness error increase with reduced cover (blue dashed line in Figure 1). This reflects the more heterogeneous distribution of polluted NO_2 column values for the high-resolution model with a small sample (5 pixels) than for the coarse resolution with a large sample (> 100 pixels). The 3° × 2° case with complete area coverage by OMI NO_2 pixels (on 17 July 2006) illustrates the potential for horizontal representativeness errors. For a fractional coverage of 0.5, the horizontal representativeness error increases to 10-15%, which is still considerably smaller than the 20-30% errors in the satellite measurements themselves. For fractional coverage of 0.1 however, the representativeness error increases to 35%, a level that exceeds the theoretical NO_2 retrieval error (Boersma et al., 2011) and NO_2 validation errors

[1] Because OMI pixel sizes vary with viewing zenith angle (largest pixels at the edge of the swath), the exact number of pixels covering a model grid cell depends on which part of the OMI swath covers the grid cell.

(e.g. Irie et al., 2012). However, by averaging over multiple days, the representativeness error can be reduced further, depending on the day-to-day variability of the columns. Table S1 (Supplementary Material) shows the statistics of a comparison between monthly mean observed and simulated columns over the greater eastern United States in July 2006, for different degrees of fractional coverage required.

In data assimilation systems, any fractional coverage may be used as long as the horizontal representativeness error is well described and accounted for along with the observation error. This can be achieved by adding in quadrature the measurement error and representativeness error $\sqrt{\sigma_{\vec{N},o}^2 + \sigma_{\vec{N},r}^2}$ to represent the overall superobservation error.

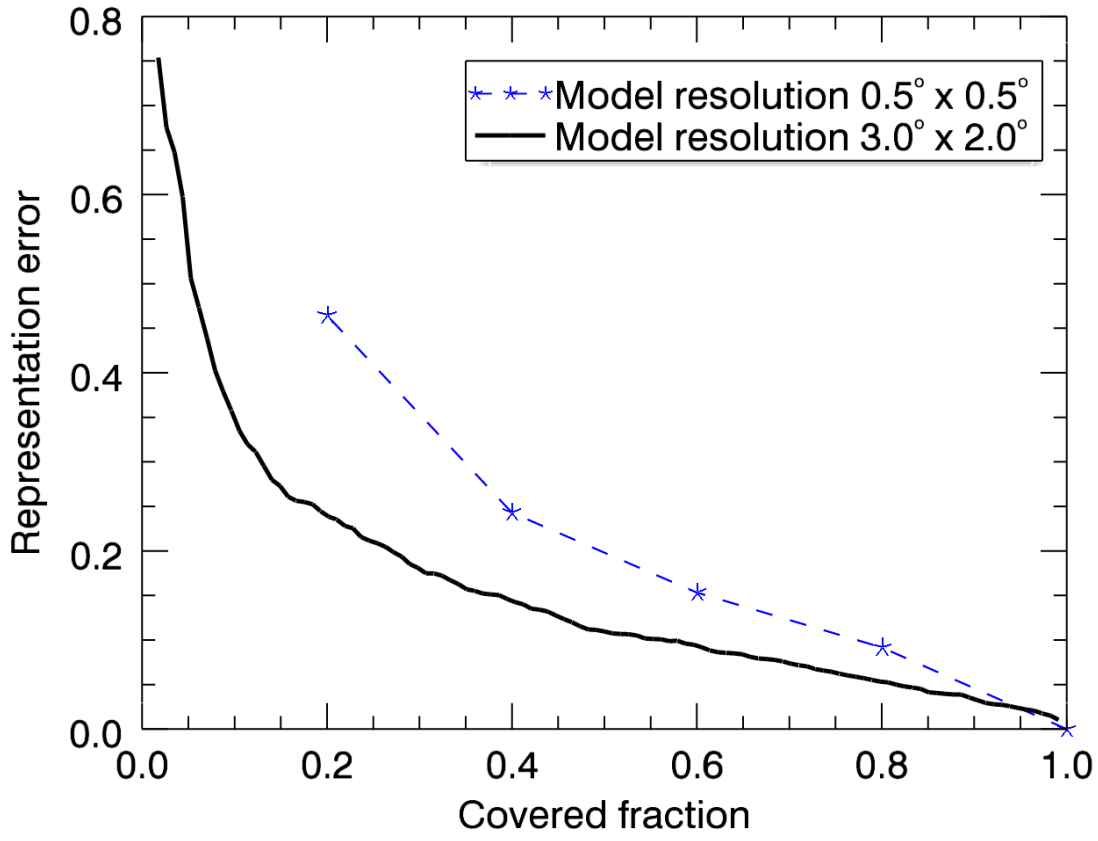


Figure 1. Relative horizontal representativeness errors as a function of the covered fraction of one model grid cell in the case of OMI tropospheric NO_2 columns for polluted area(s) (mean column 5×10^{15} molec.cm^{-2}). The black line indicates the error as a function of the fractional coverage for a 3° × 2° grid cell over the area of New York City on one day (17 July 2006, 114

OMI pixels). The blue asterisks indicate the mean error as a function of fractional coverage for various 0.5° × 0.5° grid cells on 17 July 2006.

5.2 Temporal representativeness errors related to clouds

In the case where UV/Vis satellite retrievals of the tropospheric column are used for air pollution applications (taken under cloud-free situations, see e.g. Schaub et al. [2006], Millet al. [2006], Geddes et al. [2012]), both measurements and models should be sampled under similar clear-sky situations. As long as the model appropriately simulates the effects of clouds on photolysis rates, this ensures that measurement and model represent the trace gas concentrations under similar photochemical regimes. Failure to sample the model on clear-sky days only, will introduce a bias in the modelled average. Short-lived trace gases may have a longer lifetime against photochemical loss in situations with overhead clouds (assuming they are represented well in models), when actinic fluxes and temperatures are lower and chemistry slower than in clear-sky situations. For trace gases whose emissions reflect distinct anthropogenic patterns, it is also necessary to sample the model according to the observations, in order to properly weigh well-documented weekend (e.g. Beirle et al., 2003; Boersma et al., 2009) and national holiday reductions (Lin et al., 2011) when calculating the model average.

We first evaluate the TM5 model's ability to simulate the effective cloud cover as observed by OMI at 13:30 hrs local time. Cloud cover (and cloud optical thickness) data in TM5 are hourly interpolated from 3-hourly pre-processed ECMWF fields [Huijnen et al., 2010]. Since the OMI cloud retrieval reports effective cloud fractions, based on the assumption that clouds are optically thick (optical thickness of 40, with a corresponding cloud albedo of 0.8) [Acarreta et al., 2004; Stammes et al., 2008], we converted the TM5 geometrical cloud cover into an effective fraction comparable to the OMI observations. To do so, we used the maximum-random overlap assumption [Morcrette and Jakob, 2000] to compute the total geometrical cloud cover and total cloud optical thickness from the vertically resolved cloud cover and optical thickness in TM5. We used the modelled relationship between the total cloud optical thickness for a liquid water cloud and its spherical cloud albedo in Buriez et al. [2005] to calculate the effective cloud albedo associated with each grid cell's cloud cover. Finally, we weighted the total geometric cloud cover with the ratio of the effective cloud albedo to 0.8, the value assumed for all clouds in the OMI retrieval (Acarreta et al. 2004]; Stammes et al. [2008]). For more details we refer to the Appendix C.

Figure 2 shows monthly mean effective cloud fractions as retrieved from OMI and simulated with TM5 for February and August 2006. The model was sampled within 30 minutes of the OMI overpass time of 13:30 hrs, and model and satellite were matched in space and time for further analysis. We see that TM5 captures the spatial patterns observed by OMI, with low cloud fractions over the subtropics, and high cloud fractions over the tropical ITCZ and the middle-to-high latitudes (> 40°). Largest differences occur at the edges of areas flagged as snow-covered in the OMI retrieval (February 2006), and over areas where TM5 predicts cloud optical thickness to exceed 40, such as over the tropics, where ice clouds often occur (and the relationship for water clouds from Buriez et al. [2005] is less valid).

To evaluate the simulated effective cloud fractions, we report the correlation coefficient, mean bias, and root mean square error relative to the OMI-observed cloud fractions over Europe for February and August 2006. Figure 2 shows significant positive correlation between TM5 and OMI effective cloud fractions over Europe both in February (r=0.70, n=3379) and August (r=0.75, n=4665). The mean bias between TM5 and OMI is -0.08 in February and +0.02 in August, and the root mean square error is 0.23 in February and 0.20 in August. The agreement between TM5 and OMI, while far from perfect, suggests that TM5 has some success in simulating the contrast between 'cloud-free' (f_{OMI}<0.2) and 'cloudy sky' (f_{OMI}>0.2) situations, i.e. the likelihood that OMI reports a clear-sky scene, while TM5 simulates a cloudy sky, and vice versa is <20% and <14%, respectively.

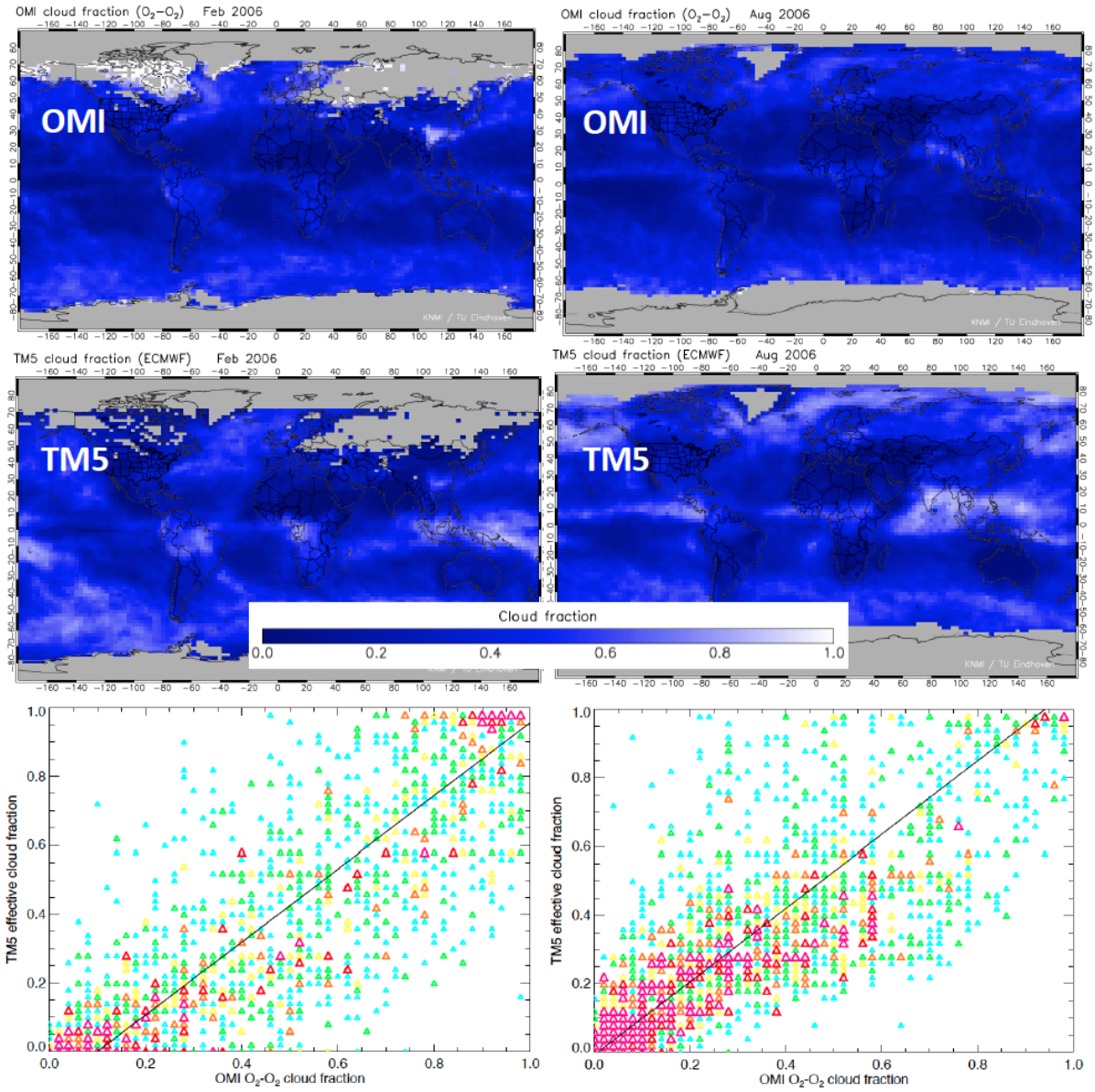


Figure 2. Monthly average effective cloud fraction observed from OMI (upper panels) and simulated by TM5 based on ECMWF meteorological fields (middle panels) in February (left column) and August 2006 (right). Cloud fractions have been selected only for those days and locations that had a successful OMI O_2-O_2 retrieval. Grey areas indicate less than 3 successful coincidences. Bottom panels: scatterplot of daily pairs of OMI (x-axis) and TM5 cloud fractions (y-axis) in February 2006 (left) and August 2006 (right) over Europe (10° W-30° E; 35°-60° N). The colours indicate the number of times a particular grid cell has been filled, where light blue corresponds to 2×, green 3×, yellow 4×, orange 5×, red 6×, and magenta to 7× or more. TM5 effective cloud fractions can be expressed as -0.10 + 1.06 f_{OMI} (February) and -0.01 + 1.07 f_{OMI} (August).

Figure 3 shows a box and whisker plot for OMI and TM5 effective cloud fractions over Europe in February and August 2006. The figure indicates that for OMI measurements of effective cloud fractions smaller than 0.2, TM5 reproduces similar small effective cloud fractions (February median OMI: 0.09, TM5: 0.06; August median OMI: 0.05, TM5: 0.04). For days and locations when OMI observes effective cloud fractions larger than 0.2 (February: 0.59, August: 0.47), TM5 simulates comparable high effective cloud fractions (January: 0.49, July: 0.45), providing some confidence in the TM5 model, driven by ECMWF meteorological fields, to capture the observed effective cloud fractions.

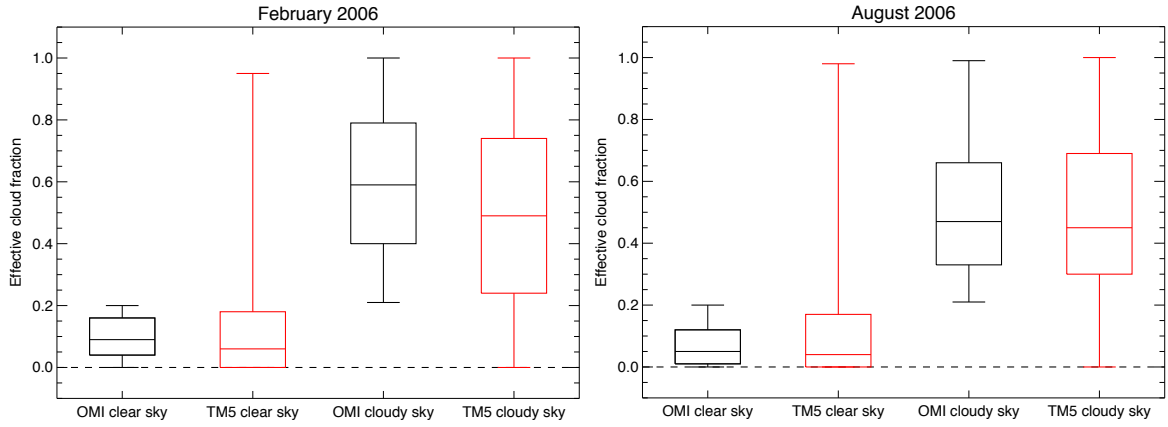


Figure 3. Box and whisker plots for OMI (black) and TM5 (red) effective cloud fractions over Europe in February 2006 (left panel) and August 2006 (right panel). The two left boxes of each panel indicate the clear sky situations when the OMI cloud fraction < 0.2. The centre line of the boxes indicate the median cloud fraction, the upper and lower edges indicate the 25^{th} and 75^{th} percentiles and the lower and upper whiskers represent the minimum and maximum value in the sample. For February 2006, the sample consisted of 3379 pairs (737 clear sky, 2642 cloudy), and for August, the sample size was 4665 (1991 clear sky, 2674 cloudy).

Figure 4(a) shows a comparison of average TM5 tropospheric NO_2 columns simulated under clear-sky and cloudy situations over Europe in February and August 2006. TM5 was sampled for polluted situations (cells with monthly mean NO_2 columns in excess of 1.0×10^{15} molec.cm^{-2}) between 12:00-15:00 hrs local time, on days with clear skies and on days with cloud-cover. Under clear-sky situations, TM5 simulates tropospheric NO_2 columns that are on average 15-20% lower than under cloudy circumstances, in line with in situ observations reported by Boersma et al. [2009] and Geddes et al. [2012] over Israeli and Canadian cities, respectively. Both in February and August, the clear-sky mean NO_2 column is 12% below the

28-day monthly mean in February and 31-day monthly mean in August. Although we cannot rule out that other effects than enhanced photochemical loss may have contributed to lower NO_2 columns over the polluted grid cells (e.g. increased ventilation or deposition) on clear-sky days, a comparison of NO_2 columns for all European grid cells showed that the geometrical mean of the local clear-sky to cloudy column ratios was 0.74 in February and 0.89 in August, suggesting that reduced clear-sky NO_2 columns presented in Fig. 4 show a robust effect.

The results for August 2006 indicate that clear-sky sampling of the model is also relevant for HCHO in the growing season (Fig. 4(b)). Average HCHO columns are 12% higher under clear-sky situations than on cloudy days and the clear-sky mean HCHO column is 8% higher than the all-sky monthly mean (August 2006). In winter, HCHO concentrations are generally low over Europe and differences between clear and cloudy sky are well below the detection limit of UV-Vis satellite sensors.

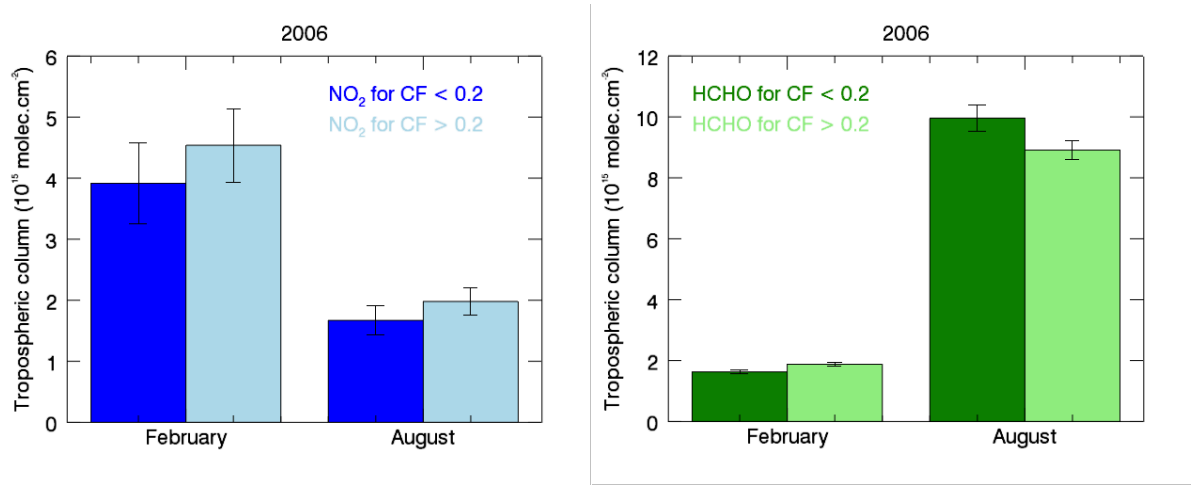


Figure 4.(a) monthly mean tropospheric NO_2 columns simulated by TM5 for polluted grid cells (with all-sky monthly means > 1.0×10^{15} molec.cm^{-2}, n=18 in February, n=17 in August). The blue bars represent the average of the tropospheric NO_2 column sampled on days when the OMI cloud fraction was smaller than 0.2. Light blue: average for columns sampled when OMI cloud fraction > 0.2. (b): monthly mean TM5 HCHO columns for clear-sky and cloudy situations (n=18 in February, for August: all-sky monthly mean > 7.5×10^{15} molec.cm^{-2}, n=12).

Exclusive sampling of the model on clear-sky days is important, because photolysis rates J[NO_2] in the lower troposphere are significantly higher on those days and can be simulated well by TM5 [Williams et al., 2012], so that NO_2 columns will be systematically lower. The

differences between HCHO columns sampled on clear-sky and cloudy days are somewhat smaller than for NO_2 columns because both the formation and destruction of HCHO are driven by photochemistry. Nevertheless, the stronger summertime production of HCHO from the (OH-driven) oxidation of methane and especially isoprene outpaces the increased loss of HCHO through photolysis and oxidation [Fried et al., 1997] on clear-sky days compared to cloudy days, in line with observations (e.g. Munger et al. [1995], Cerquiera et al. [2003]).

To estimate the magnitude of the temporal representativeness errors arising from the particular choice of model sampling, we evaluated the satellite-model comparison results for different sampling strategies. Again, we use the averaged ratio of satellite measurements to model simulations ($\hat{y}_o/\hat{x}_m$), and the spatio-temporal correlation coefficient, as appropriate indicators of representativeness errors. Since the model – measurement bias may well be due to unrelated systematic errors in either the CTM (emissions, chemistry) or the satellite retrievals, we are not concerned with the absolute value of the measurement-to-model ratio, but we are interested in the sensitivity of the ratio to various sampling strategies. We tested four strategies for comparing tropospheric NO_2 over large polluted regions: (A) both OMI (for OMI effective cloud-fraction) and TM5 (TM5 effective cloud fraction) collocated and sampled for mostly clear-sky scenes only at the OMI overpass time of 13:30 hrs, (B) OMI and TM5 collocated and co-sampled for situations with OMI effective cloud radiance fractions < 0.5[2], (C) OMI sampled for situations with OMI effective cloud radiance fractions < 0.5, but TM5 more loosely sampled for OMI effective cloud fractions < 0.6, and (D) OMI sampled for situations with OMI effective cloud radiance fractions < 0.5, but TM5 sampled for all days in the month (i.e. no temporal collocation except for appropriate overpass time). Strategy (A) is considered to be optimal, but to our knowledge has not been applied in studies to date. Strategy (B) has been followed in numerous studies, and relies on the assumption that CTMs capture the observed cloud cover well. In spite of its erroneous co-sampling with the satellite measurements, strategy (D) has also been used frequently, and therefore we tested its impact on the temporal representativeness errors. Finally, strategy (C) holds middle ground between (B) and (D). Figure 5 shows that the model-to-measurement ratio shows substantial dependence on the comparison strategy, especially in Winter. The differences between strategies (A) and (B) are negligible, but with strategy (D) the OMI/TM5 ratio drops more

[2] The cloud radiance fraction is defined as the relative contribution of top-of-atmosphere radiance received by the cloud part of the pixel. A cloud radiance fraction of 0.5 corresponds to a geometric cloud fraction of ±0.2.

than 25% below the values obtained by strategies (A) and (B). These strategies also demonstrate that strategy (D) leads to a reduced capacity of the model to explain the observed variability in the NO_2 spatial patterns, with R^2 dropping almost 10% (from 0.64 to 0.55 in Winter and from 0.66 to 0.59 in Summer).

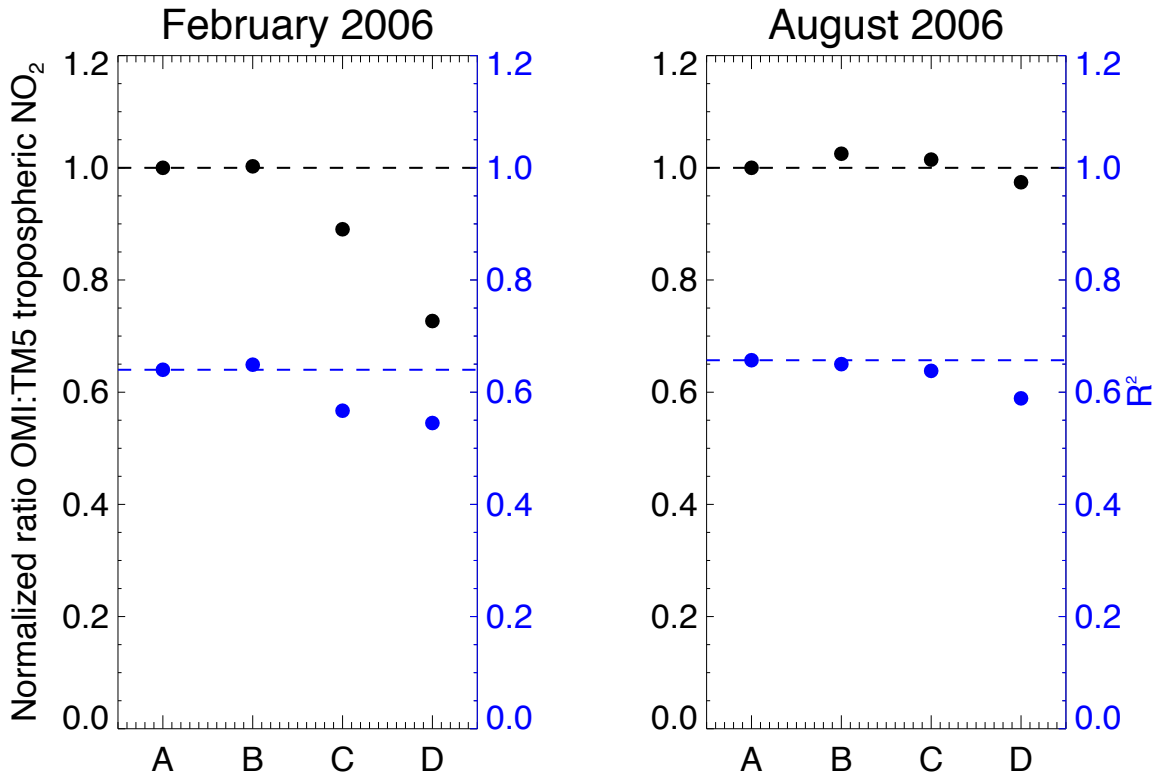


Figure 5. Impact of sampling strategy on monthly averaged OMI:TM5 ratio of tropospheric NO2 columns (black dots) and on spatial correlation coefficient (R^2, blue dots) over the eastern United States (30°-44° N, 90°-72° W). Left panel: ratio and R^2 for February 2006 (*n*=28). Right panel: August 2006 (*n*=32). Grid-cells were selected in the comparison when the covered fraction exceeded 0.5. The dashed black line shows the normalized OMI:TM5 ratio for strategy (A), and the dashed grey line shows the R^2 for strategy (A) as a guide to the eye.

Analyses for other regions showed similar results as in Fig. 5. These results imply that for applications of satellite data such as emission estimates or model evaluations, substantial systematic errors may occur in the final estimate, if sampling strategies such as (D) are used. We therefore strongly discourage the use of such comparison strategies, as they lead to considerable temporal representativeness errors, and, thus, systematic underestimations in measurement:model ratios.

5.3 Vertical representativeness errors

Here we evaluate the representativeness errors introduced in a satellite-model comparison if the averaging kernel is not accounted for. To illustrate the way the kernels work, Figure 6 shows GEOS-Chem NO_2 vertical profiles with and without the averaging kernel applied over the Beijing grid cell on clear-sky days with excellent spatial coverage (18 February and 23 August 2006). On both days, application of the kernel leads to a higher value for the model column, reflecting the relatively larger amounts of NO_2 aloft in GEOS-Chem simulations compared to the a priori TM4 NO_2 profiles. The lower panels show that on two other clear-sky days (17 February and 31 August 2006) the kernel has only little effect on the GEOS-Chem tropospheric NO_2 column. On these days, the TM4 a priori and GEOS-Chem NO_2 profiles show similar, less pronounced vertical distributions. Nevertheless, in Figure 7 we see that, on average, for February and August 2006, the OMI averaging kernels result in increases in GEOS-Chem NO_2 columns over Beijing of 15% (February) and 8% (August), and a closer agreement with OMI NO_2 retrievals. This result can be understood from the stronger vertical mixing in the GEOS-Chem model compared to TM4, rather than from differences in NO_x emissions or chemistry between models (NO_2 amounts are quite similar between TM4 and GEOS-Chem over Beijing in 2006).

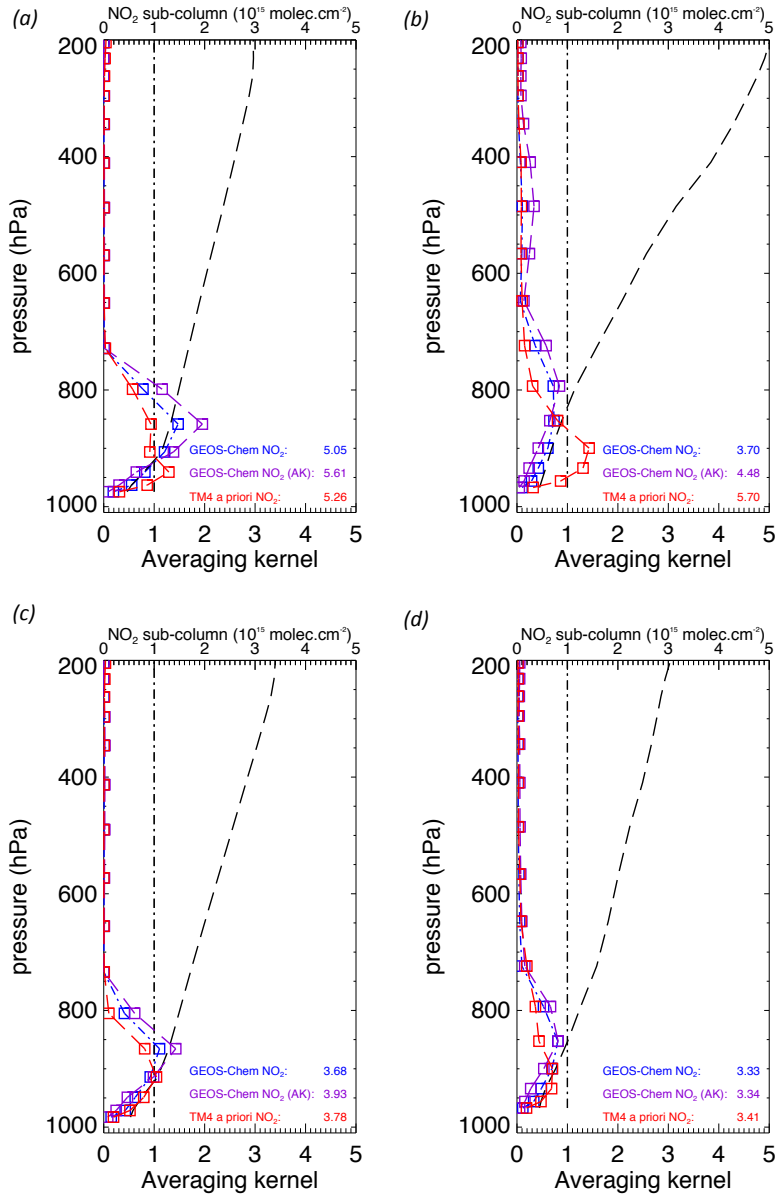


Figure 6. Vertical averaging kernel (black dashed line) and NO_2 profiles simulated by GEOS-Chem (blue), TM4 (red, a priori profiles in OMI NO_2 retrieval), and GEOS-Chem convolved with the averaging kernel (purple) following Eq. (6); (a) 18 February 2006, (b) 23 August 2006 over the Beijing grid cell (centered on 40°N, 116.25°E), (c) 17 February 2006, and (d) 31 August 2006. The numbers given in blue, purple, and red indicate the tropospheric vertical NO_2 columns in GEOS-Chem and TM4.

The above finding does not have general validity in the sense that applying the kernel on any other model will also result in a tropospheric column increase. Applying the kernels to NO_2 profiles from a model with weaker vertical mixing than TM4 (rather than generally stronger vertical mixing as in the case of GEOS-Chem) is likely to reduce those columns. Figure S1 in

the Supplementary Information shows as much for the North Sea grid cell in February 2006, when GEOS-Chem exceeds TM4 NO_2 concentrations below 900 hPa, and for Siberia in August 2006, when GEOS-Chem simulates a substantially enhanced tropospheric NO_2 column compared to TM4.

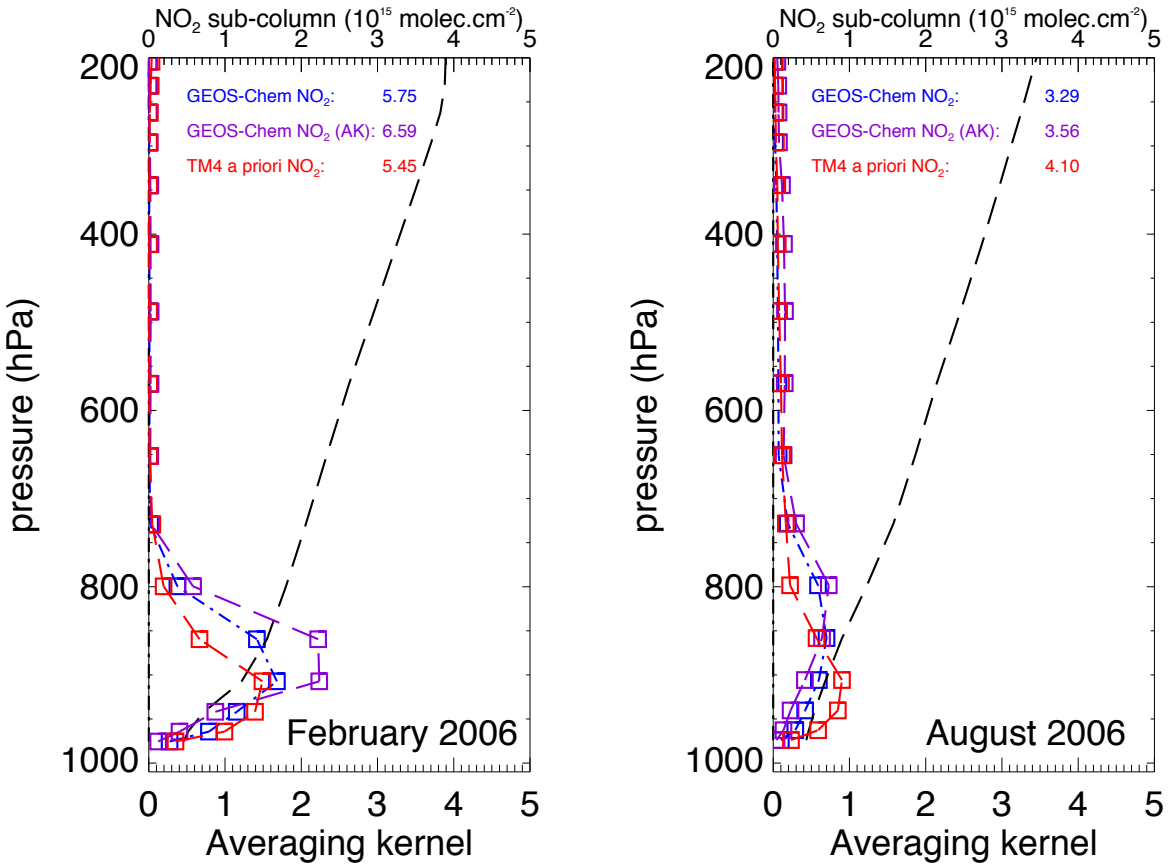


Figure 7. Monthly mean averaging kernel (black dashed line) and NO_2 profiles simulated by GEOS-Chem (blue), TM4 (red, a priori profiles in OMI NO_2 retrieval), and GEOS-Chem convolved with the averaging kernel (purple) following Eq. (6); left panel: February 2006, right panel: August 2006 over the Beijing grid cell (centered on 40°N, 116.25°E). The numbers given in blue, purple, and red indicate the tropospheric vertical NO_2 columns in GEOS-Chem and TM4.

We next compare the monthly averaged GEOS-Chem tropospheric NO_2 column fields for February and August 2006 with and without the kernels applied. Figure 8 shows that applying the kernel leads to substantial increases of up to 2×10^{15} molec. cm^{-2} in the columns for the polluted source regions in the northern hemisphere (eastern USA, Europe, and China). At the periphery of these regions in wintertime, and over regions with possible biomass burning in summer, we see that the smoothed columns can be lower than the original columns, indicating

that the GEOS-Chem vertical NO_2 profile is more skewed towards the surface than the TM4 a priori in those situations, as confirmed by the profiles shown in Figure S1.

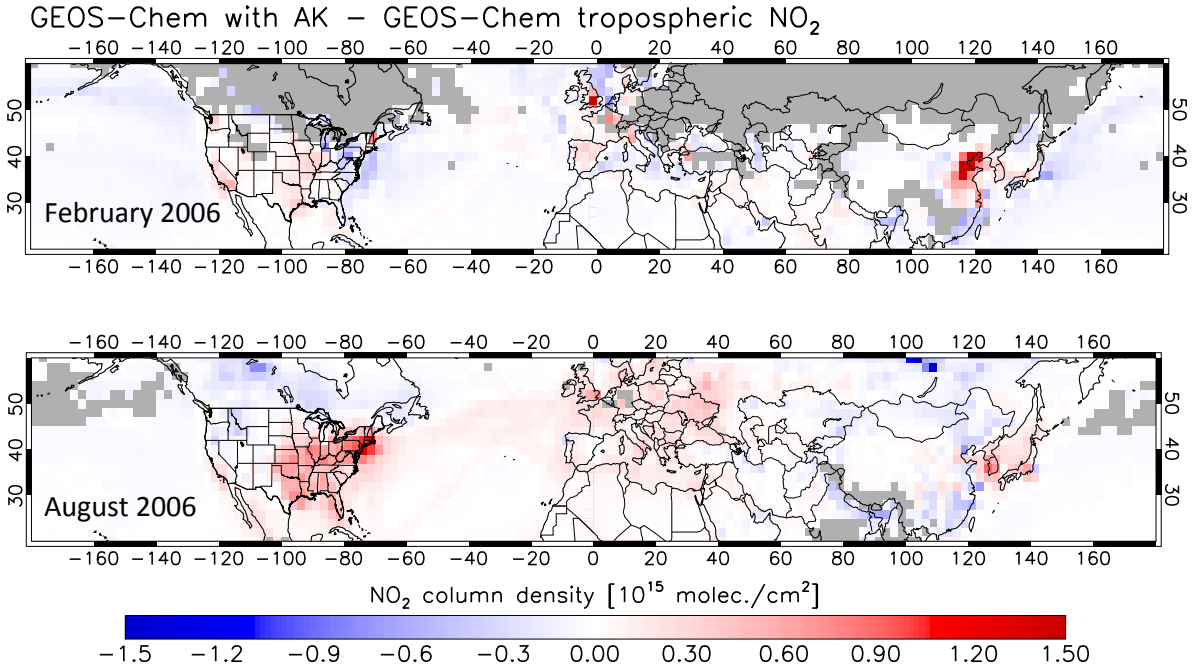


Figure 8. Difference between monthly mean GEOS-Chem with AK (Eq. (6)) and GEOS-Chem tropospheric NO_2 columns without AK for February 2006 (upper panel) and August 2006 (lower panel). Only grid cells with more than 3 days of better than 40% coverage of clear-sky pixels have been selected.

Here we evaluate the level of agreement between the original GEOS-Chem and OMI NO_2 columns, compared to the level of agreement between the kernel-based GEOS-Chem and OMI NO_2 column for the polluted source regions in the northern hemisphere, as the differences provide a measure of the representativeness errors that can be avoided by using the averaging kernel. Figure 9 shows the agreement between OMI and the GEOS-Chem NO_2 columns with and without kernel over Europe in February and August 2006. The upper panels indicate that the spatial correlation between the model and OMI tropospheric columns improves when the kernel is applied on the model NO_2 profiles, especially in Summer when differences between the TM4 a priori and GEOS-Chem NO_2 profile shapes are strong. Application of the kernel also results in geometric mean OMI:GEOS-Chem ratios with smaller uncertainty intervals at values of $1.15^{1.82}_{0.73}$ (February) and $1.24^{1.78}_{0.86}$ (August) compared to $1.13^{1.82}_{0.70}$ and $1.42^{2.14}_{0.94}$. We find similar results over the eastern United States and China (see Table 1 in Sect. 7). Figure 9(d) further supports the notion that application of the kernel

allows for a better-constrained evaluation of the model, as witnessed by the more peaked and narrower histogram of satellite:model ratios. We conclude that sampling the model according to the averaging kernel is especially relevant in Summer, and improves the satellite-model evaluation by removing differences between (TM4 apriori and GEOS-chem) profile shapes contributing to the discrepancies [Boersma et al., 2004]. Neglecting the kernels for GEOS-Chem would lead to up to 15% stronger discrepancies between OMI and GEOS-Chem, and this portion could be wrongfully attributed in a model evaluation to e.g. too low NO_x emissions, or too fast NO_2 removal by chemistry or deposition. Appendix D presents an alternative to the application of the averaging kernel by providing a recipe to replace the a priori profile used in the retrieval by the profile from the CTM under evaluation. Such a recipe results in a modified retrieval that can be directly compared with the CTM under evaluation.

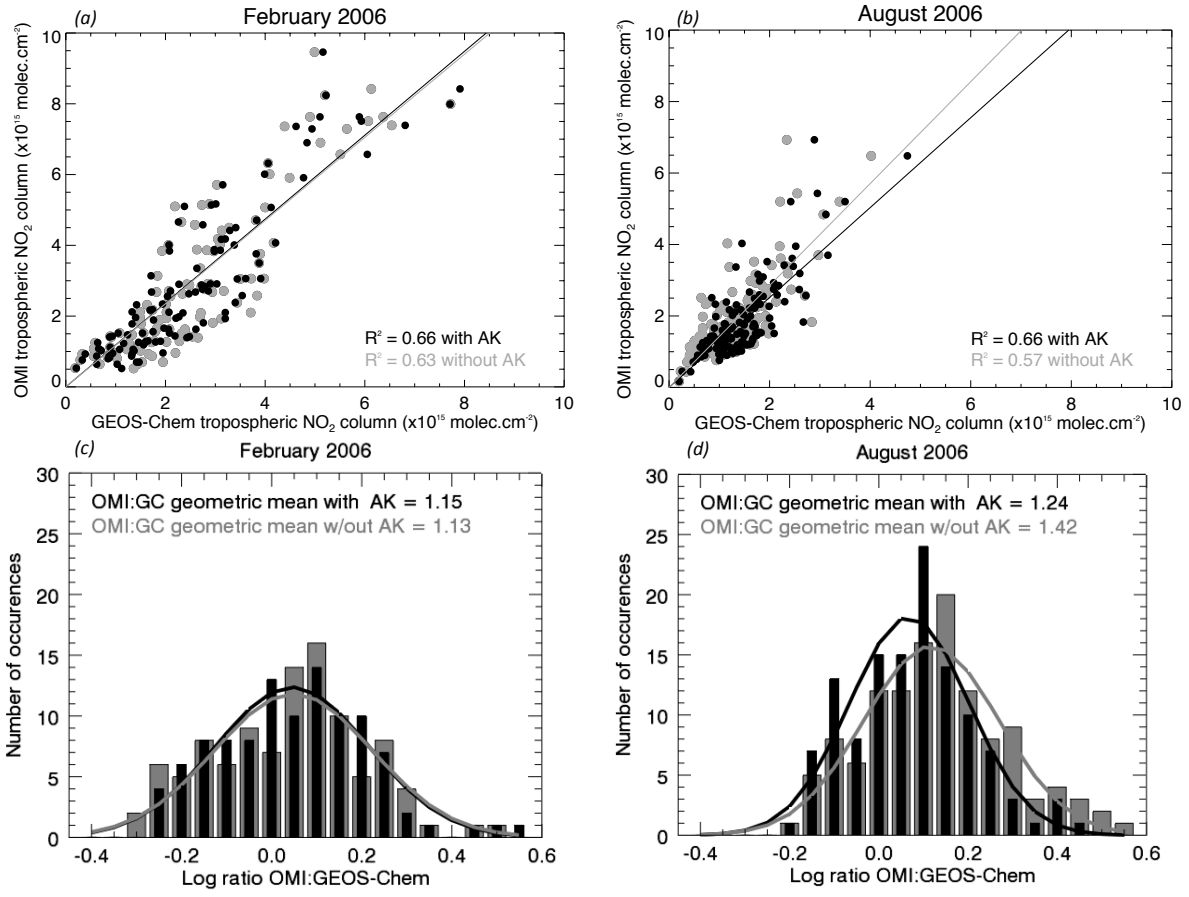


Figure 9. Comparison between monthly average OMI and GEOS-Chem tropospheric NO_2 columns over Europe in February 2006 (left panels) and August 2006 (right panels); (a) scatter diagram of monthly average GEOS-Chem with AK (black circles) and GEOS-Chem without AK (grey circles) vs. OMI tropospheric NO_2 columns for February 2006. The black

and grey lines indicate the geometric mean of the OMI:GEOS-Chem ratio; (b) as for (a) but for August 2006; (c) histogram of per-grid cell OMI-to-GEOS-Chem with AK tropospheric NO_2 column ratios (black bars) and OMI-to-GEOS-Chem without AK ratios (grey bars) for February 2006; (d) as (c) but for August 2006. Only grid cells with more than 3 days of better than 40% coverage of clear-sky pixels have been selected.

6 Combined representativeness errors

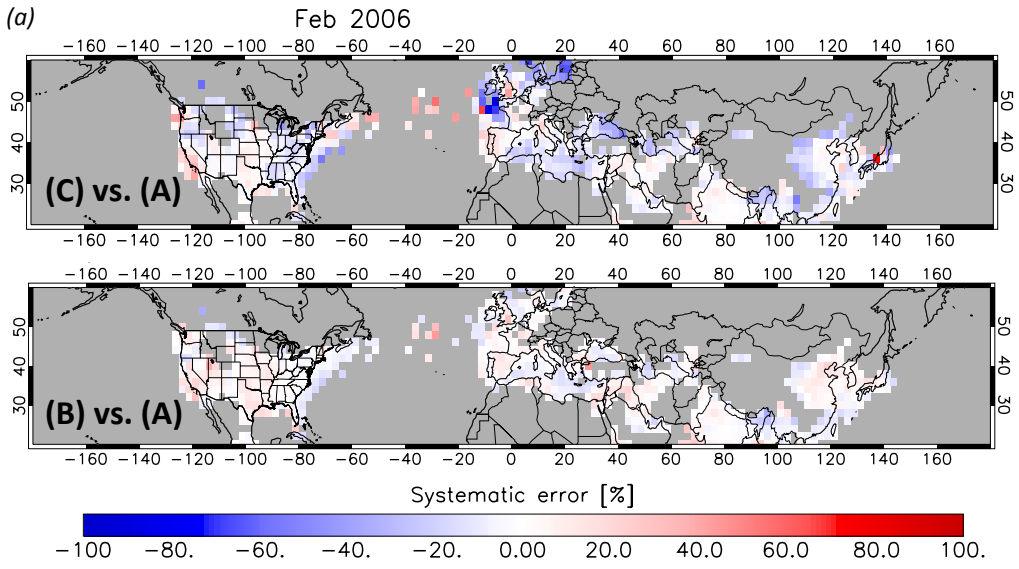
To obtain an estimate of typical, overall representativeness errors in model evaluations with UV/Vis satellite measurements, we define three types of model evaluations, executed with increasing degree of detail. We again evaluate tropospheric NO_2 from the GEOS-Chem model here (with OMI NO_2 retrievals), as this model is sufficiently different from the TM4 model used to provide the a priori profiles in the OMI retrievals. The three types of evaluations can be characterised as advanced, common, and naïve:

(A) advanced evaluation: accounting for sufficient spatial coverage and appropriate temporal representativeness, and also taking into account vertical representativeness,
(B) common evaluation: as (A) but without taking into account vertical sensitivity,
(C) naïve evaluation: no consideration of potential representativeness errors whatsoever,

For evaluation (C), the model monthly average was based on samples from all days of the month (on OMI overpass time), irrespective of cloud coverage, and no kernel was applied (in other words a 31-day, all-sky, without AK monthly mean). We first evaluate the (avoidable) representativeness errors by comparing local OMI:GEOS-Chem ratios evaluated with approaches (A) vs. (C), and approaches (A) vs. (B). Figure 10 shows the relative difference in the local OMI:GEOS-Chem ratios for February and August 2006. We see that the systematic, avoidable errors in the OMI:GEOS-Chem ratio are largest with evaluation approach (C). The blue colours in the upper panel of Figure 10(a) indicate that, in winter, sampling the model on all (including cloudy sky) days leads to too low (by 15-20%) OMI/GEOS-Chem ratios reflecting the too high GEOS-Chem NO_2 values resulting from temporal representativeness errors (cloudy-sky sampling cf. Figure 4).

The similarity between the panels of Figure 10(b) shows that appropriate sampling is not as important in Summer, a season with ample clear-sky days, and, consequently, a smaller sampling error. Figure 10(b) suggests that application of the averaging kernel when sampling the model is the most important step, with the red colours indicating that failure to apply the averaging kernel leads to OMI/GEOS-Chem NO_2 ratios that are too high by up to 30%. We

conclude that appropriate clear-sky sampling is mainly important in winter, but vertical smoothing is less relevant in that season. The reverse holds in Summer: with sufficient clear-sky days available, application of the averaging kernel becomes essential, reflecting the fact NO_2 vertical distributions are especially different between (the TM4 and GEOS-Chem) models in that season.



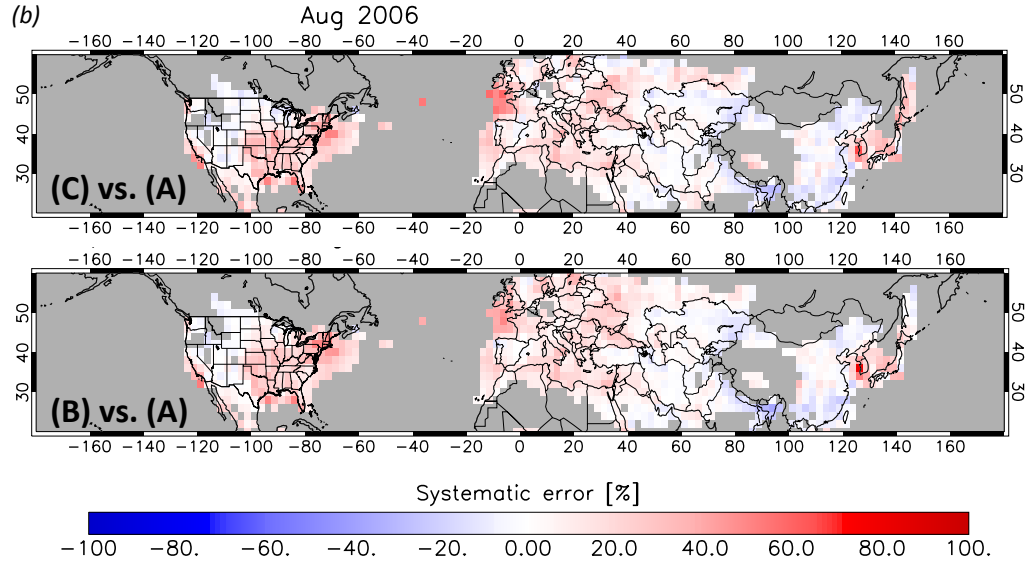


Figure 10. Relative difference between local monthly mean OMI:GEOS-Chem NO_2 column ratio's for (a) February 2006 between method (C) and (A) (upper panel) and between method (B) and (A) (lower panel), and (b) August 2006. Relative difference defined as 100% × ((C/A)-1), and 100% × ((B/A)-1).

Table 1 summarizes the results of the OMI/GEOS-Chem comparisons for the three specific regions of the United States, Europe, and China following the different evaluation approaches. In all cases, the spatial correlation between model and measurements within the regions is highest for evaluation approach (A), and generally lowest for approach (C). Wintertime OMI:GEOS-Chem ratios are too low by 15-20% with approach (C) and too high by 5-10% in Summer. Using the common approach (B), OMI/GEOS-Chem ratios are primarily biased in Summer, by +15-20% for Europe and the United States, and by -5% for China. The results in Table 1 and Figure 9 also indicate that the spread of local OMI/GEOS-Chem ratios is ±30% for approach (A), smaller than for approaches (B) and (C) with spreads of ±35%, corroborating the fact that using the kernel results in a better-defined comparison between satellite measurements and model simulations.

We summarize the contribution of the model sampling errors to the overall representativeness errors for the evaluation of GEOS-Chem simulations with OMI NO_2 in Table 2. The table should not be interpreted as a general recommendation for all applications, but rather as a recommendation for air pollution applications such as model evaluation and inversions to estimate emissions. For instance, for data assimilation and studies of the higher atmosphere, retrievals under cloudy situations can still be used, and the main recommendation there is to apply the averaging kernel. The table shows that naïve comparison strategies (C) that do not account for appropriate temporal or vertical sampling will result in a largely systematic representativeness error of up to 25%. Following the motivated recommendations discussed above however (i.e. comparison strategy (A)) would eliminate temporal and vertical representativeness errors and limit the overall comparison error to not more than 5-10% from imperfect horizontal sampling.

7 Discussion and conclusions

Evaluations of chemistry-transport model simulations with UV/Vis satellite retrievals of short-lived gases, notably NO_2 and HCHO, are strongly influenced by the exact comparison strategy. The characteristics of these satellite retrievals –with ground pixels typically smaller than model grid cells, clear-sky sampling needed for air pollution applications, and reduced vertical sensitivity towards the lower troposphere- require that models and retrievals are sampled as consistently as possible. This pertains to consistent sampling in space (horizontally and vertically) and in time (day-of-week, clear-sky day, time-of-day). Of these

aspects, appropriate horizontal sampling is a relatively minor, but unavoidable concern. In most model-to-satellite comparisons, we recommend using the concept of the superobservation, which has the distinct advantage of providing a grid cell average observed state along with a realistic measurement plus horizontal representativeness error. Depending on the model resolution and the satellite instrument resolution, users can impose a minimum fractional coverage (of the model grid cell area) by the ensemble of pixels to reduce horizontal representativeness errors down to levels where the measurement contribution becomes the dominant term in the superobservation error budget.

Recommendations on and error estimates of the fractional coverage requirement depend on the exact method of comparing model simulations and satellite retrievals and on the spatial variability of the species of interest. Generally speaking, fractional coverage requirements may be rather loose for comparisons over regions with little spatial variability in gas concentration, for coarse-resolution model simulations, and for temporal averages over multiple days (e.g. monthly means). In contrast, total fractional coverage requirements need to be strict for comparisons over regions with strong variability in gas concentrations (i.e. SO_2 and NO_2 source regions), on a high spatial resolution with (regional) CTMs.

In these situations we recommend limiting horizontal representativeness errors to within ±10% because representativeness errors are then still considerably smaller than the satellite observation error $\sigma_{\overline{N},o}$.

A faithful comparison between satellite measurements and model simulations requires that models need to be sampled appropriately in time. Sampling models irrespective of photochemical regime (such as when calculating a 31-day monthly mean without collocating the model with individual measurements) gives rise to systematic temporal representativeness errors on the order of +12% for NO_2 and -8% for HCHO. Such errors should (and can) be avoided, as they may misdirect interpretation of model-satellite differences, for instance by misinforming inversion studies by requiring changes in the rates of emissions, or chemical reactions to better match the observations. Our comparison of OMI O_2-O_2 and co-sampled TM5 cloud information indicated that a strict requirement on the TM5 model to simulate a clear-sky scene along with a mostly clear-sky OMI superobservation has little effect over omitting such a filter. In the case of TM5, driven by ECMWF ERA Interim meteorological fields, the model shows good correlation with OMI-observed cloud fractions, with little probability (<15%) of simulating false positives or negatives.

Larger systematic errors in model-satellite ratios will be introduced when model profiles are not sampled according to the averaging kernel associated with most UV/Vis satellite products. While the exact magnitude effect depends on the model under evaluation and on the a priori profiles and other assumptions used in the retrievals, our analysis showed that for a comparison between OMI and GEOS-Chem NO_2, application of the averaging kernels results in up to 20% lower satellite-to-model ratios, and more coherent values of these ratios within relevant regions such as the eastern United States and Europe. The effect of applying the kernel is most relevant in Summer, when the vertical distribution of species like NO_2 and HCHO is variable, and differences between the model profiles and the profiles used in the retrieval are most prominent. We strongly recommend using averaging kernels in satellite-model evaluations. Use of the averaging kernel allows for a better satellite-to-model comparison, by ensuring that the model is sampled in a manner consistent with the satellite retrievals because identical assumptions are made on vertical sensitivity, and differences between the model and satellite a priori vertical distribution cancel. Here we focused on an evaluation of tropospheric NO_2 simulations from the GEOS-Chem model with retrievals of tropospheric NO_2 columns with substantial vertical sensitivity down to the lower troposphere. However, application of the averaging kernel will be even more relevant for model evaluations of HCHO and SO_2, since these retrievals are less sensitive to the lower troposphere. Recently, retrieval scientists have also made averaging kernel information available along with the HCHO and SO_2 data products (e.g. González Abad et al. [2014]; Theys et al. [2013]).

For future evaluations of chemistry transport models and data assimilation with UV/Vis satellite retrievals (of NO_2, HCHO, CHO-CHO, or SO_2), we advocate the use of the recommendations laid out in this paper, especially with respect to the required clear-sky sampling and appropriate vertical smoothing.

Appendix A

Calculating horizontal representativeness errors

The horizontal representativeness error of an ensemble of satellite measurements (superobservation) for a model grid cell of any size can be calculated as follows:

(1) First compute the distance from one corner coordinate to the two adjacent (not opposite) corners to obtain estimates for the 'base' and the 'height' of the pixel[3]. Then approximate the pixel as a parallelogram, to calculate the pixel area A_i as base × height.

(2) Calculate the fractional coverage f_{cov} of all valid satellite pixels in the model grid cell as the ratio of the area covered by all n valid pixels to the complete area covered by the grid cell A_{cell}:

$$f_{cov} = \frac{\sum_{i=1}^{n} A_i}{A_{cell}} \quad \text{(A1)}$$

(3) Given the fractional coverage f_{cov}, the horizontal representativeness error can be read off from Figure 1 for models with 3° × 2° and 0.5° × 0.5° resolution. Figure 1(b) of Miyazaki et al. [2012] provides a similar figure for a model resolution of 2.5° × 2.5°. For example, a 0.6 fractional coverage for a 3° × 2° model grid cell corresponds to a horizontal representativeness error of ~10%. 0.6 coverage for a 0.5° × 0.5° model corresponds to a representativeness error of ~15%.

Note that the recipe laid out above provides a horizontal representativeness error that is at the high end of the possible range. The variability in the complete ensemble of pixels will often be much smaller than the variability in the ensemble of pixels from Figure 1 (over New York City) or Figure 1(b) from Miyazaki et al. [2012] (which excluded situations with small NO_2 columns).

Appendix B

[3] A simple distance calculation between two latitude, longitude pairs (lat1, lon1) and (lat2, lon2) is provided by the following Fortran90 pseudo-code:

```
real, intent(in)      :: lat1, lon1         ! coordinates of pixel 1
real, intent(in)      :: lat2, lon2         ! coordinates of pixel 2
real, parameter       :: dtkm = 111.32      ! at equator 1deg equals 111.32 km

deg_to_rad = acos(-1.)/180.
angle1 = 0.5 * (lat1-lat2)*deg_to_rad
angle2 = 0.5 * (lon1-lon2)*deg_to_rad
arg = (sin(angle1))2 + cos(lat1*deg_to_rad)*cos(lat2*deg_to_rad) + (sin(angle2))2
y = dtkm * 2. * asin(sqrt(arg))*180./acos(-1.)
```

Derivation of the superobservation error

If the retrieval errors within a superobservation grid cell have some degree of correlation, we cannot simply take the area-weighted average retrieval error σ (calculated as $\frac{\sum_{i=1}^{n} w_i \sigma_i}{\sum_{i=1}^{n} w_i}$) as representative for the superobservation error. The error expectation value for the ensemble of pixels composing a superobservation is written as:

$$\langle \varepsilon_N^2 \rangle = \sum_{ij} w_i w_j \langle \epsilon_i \epsilon_j \rangle \quad \text{(B1)}$$

with ϵ_i the individual retrieval error in pixel *i*, the area weights now normalized ($\sum_i w_i = 1$) to facilitate notation. Now, for a partly correlated error between pixels *i* and *j*, we write:

$$\langle \epsilon_i \epsilon_j \rangle = \begin{cases} \sigma_i^2 \text{ for } i = j \\ c\sigma_i \sigma_j \text{ for } i \neq j \end{cases} \quad \text{(B2)}$$

so that the superobservation error σ_N^2 can be written as follows:

$$\sigma_N^2 = \langle \varepsilon_N^2 \rangle = (1-c) \sum_i w_i^2 \sigma_i^2 + c(\sum_i w_i \sigma_i)^2 \quad \text{(B3)}$$

For $\sigma_i = \sigma$, and $w_i = \frac{1}{n}$ this reduces to Eq. (6): $\sigma_N = \sigma \sqrt{\frac{1-c}{n} + c}$.

Appendix C

Calculating CTM-simulated effective cloud fractions

We can express the modelled cloud properties into a quantity that is comparable to the effective cloud fraction provided by the OMI O_2-O_2 cloud retrieval, and defined as the radiometric equivalent fraction of a viewing scene covered by a Lambertian reflector with an albedo of 0.8 (corresponding to a cloud with an optical thickness of ~40) [Stammes et al., 2008]. Some data products use cloud information retrieved with different approaches, but many UV/Vis trace gas retrievals use the effective cloud fraction approach. The TM5 cloud information (geometric cloud cover, and cloud optical thickness) was converted into an effective cloud fraction in a two-step approach. In the first step the maximum-random overlap assumption is used to calculate the one column-representative geometrical cloud cover $f_{tm5,geo}$ following practical guidelines for similar model evaluations with MODIS clouds by Quaas [2011]. The maximum-random overlap assumption implies maximum overlap for cloud cover in adjacent layers (one cloud layer is exactly on top of the other), and random overlap for (layers of) cloud cover f_l separated by at least one clear-sky layer:

$$f_{tm5,geo} = \prod_{l=1}^{L} \frac{1-\max(f_l, f_{l-1})}{1-\min(f_{l-1}, 1-\epsilon_f)} \quad \text{(C1)}$$

where ϵ_f (here 0.001) is the threshold value for which a layer is considered to be cloud-free. In the second step the albedo of the cloud is determined based on the cloud optical thickness and the sensitivity of cloud spherical albedo to cloud optical thickness modelled by Buriez et al. [2005] for a liquid water cloud[4]. The final step to obtain the effective (OMI equivalent) TM5 cloud fraction $f_{tm5,eff}$ from the geometrical cloud fraction and the obtained cloud albedo a_c proceeds as:

$$f_{tm5,eff} = f_{tm5,geo} \frac{a_c}{0.8} \quad \text{(C2).}$$

Appendix D

Alternatives to the application of the averaging kernel

In a satellite-model comparison, the vertical sensitivity needs to be taken into account, and this can be done in alternatively by replacing the a priori profile $\mathbf{x_a}$ from the CTM used in the retrieval, by the profile $\mathbf{x_m}$ from the CTM used by the modeller, i.e. by re-calculating the air mass factors as follows:

$$M'(\mathbf{x_m}) = M(\mathbf{x_a}) \frac{\sum_{l=1}^{L} A_l x_{m,l}}{\sum_{l=1}^{L} x_{m,l}} \quad \text{(D1)}$$

with $M(\mathbf{x_a})$ the original tropospheric air mass factor used in the retrieval, and A_l the elements of the averaging kernel. The new air mass factors ($M'(\mathbf{x_m})$) need to be applied on the retrieved slant column densities (instead of $M(\mathbf{x_a})$), to generate modified columns $\hat{y}'_o$. These modified columns can be directly compared to the model column $\hat{x}_m$, without the need to explicitly apply the averaging kernel. Such approaches have been shown to improve the consistency of the comparison considerably -- for instance by 10-20% in the case of tropospheric NO_2, see Lamsal et al. [2010], Vinken et al. [2014] and Lamsal et al. [2014].

The modified averaging kernels associated with $\hat{y}'_o$ retrieved with the new a priori profiles $\mathbf{x_m}$ become:

$$\mathbf{A'} = \frac{M(\mathbf{x_a})}{M'(\mathbf{x_m})} \mathbf{A} \quad \text{(D2)}$$

[4] A 6th order polynomial fitted in close approximation to the relationship between cloud albedo a_c and cloud optical thickness τ_c in Figure 2 of Buriez et al. [2005] was used for the conversion: $a_c = \sum_{i=0}^{6} c_i \tau^i$ with c_0=0.00808, c_1=0.11153, c_2=-0.09734, c_3=0.00052, c_4=-0.0000154, c_5=0.00000029, and c_6=-0.0000000013.

Acknowledgements

K. Folkert Boersma acknowledges funding from Vidi Project 'Attributing the sources of tropospheric ozone from space', NWO Grant 864.09.001 and from the FP7 Project Quality Assurance for Essential Climate Variables (QA4ECV), n° 607405. We appreciate the constructive suggestions from reviewers that helped improve this paper.

Author contributions

K. F. B. performed the research, drafted the manuscript, prepared the figures and developed the analysis methods. G. C. M. V. contributed to the development of the averaging kernel code. H. J. E. aided in drafting the manuscript and methods, and supported the development of the analysis methods and interpretation. All authors contributed to discussions of the results and preparation of the manuscript.

References

Acarreta, J. R., J. F. De Haan, J. F., and Stammes, P.: Cloud pressure retrieval using the O_2-O_2 absorption band at 477 nm, J. Geophys. Res., 109, D05204, doi:10.1029/2003JD003915, 2004.

Barkley, M. P., De Smedt, I., Van Roozendael, M., Kurosu, T. P., Chance, K., Arneth, A., Hagberg, D., Guenther, A., Paulot, F., Marais, E., Mao, J.: Top-down isoprene emissions over tropical South America inferred from SCIAMACHY and OMI formaldehyde columns, J. Geophys. Res., 118, 6849–6868, doi:10.1002/jgrd.50552, 2013.

Beirle, S., Platt, U., Wenig, M., and Wagner, T.: Weekly cycle of NO_2 by GOME measurements: a signature of anthropogenic sources, Atmos. Chem. Phys., 3, 2225-2232, doi:10.5194/acp-3-2225-2003, 2003.

Beirle, S., Salzmann, M., Lawrence, M. G., and Wagner, T.: Sensitivity of satellite observations for freshly produced lightning NO_x, Atmos. Chem. Phys., 9, 1077-1094, doi:10.5194/acp-9-1077-2009, 2009.

Beirle, S., Boersma, K.F., Platt, U., Lawrence, M.G., and Wagner, T.: Megacity emissions and lifetimes of nitrogen oxides probed from space. Science, 333(6050), 1737-1739, doi: 10.1126/science.1207824, 2011.

Boersma, K. F., Eskes, H. J., and Brinksma, E. J., Error analysis for tropospheric NO_2 retrieval from space, *J. Geophys. Res.*, 109, D04311, doi:10.1029/2003JD003962, 2004.

Boersma, K. F., Eskes, H. J., Meijer, E. W., and Kelder, H. M.: Estimates of lightning NO_x production from GOME satellite observations, Atmos. Chem. Phys., 5, 2311-2331, doi:10.5194/acp-5-2311-2005, 2005.

Boersma, K. F., Jacob, D. J., Eskes, H. J., Pinder, R. W., Wang, J., and van der A, R. J.: Intercomparison of SCIAMACHY and OMI tropospheric NO_2 columns: Observing the diurnal evolution of chemistry and emissions from space, J. Geophys. Res., 113, D16S26, doi:10.1029/2007JD008816, 2008.

Boersma, K. F., Jacob, D. J., Trainic, M., Rudich, Y., De Smedt, I., Dirksen, R., and Eskes, H. J.: Validation of urban NO_2 concentrations and their diurnal and seasonal variations observed from the SCIAMACHY and OMI sensors using in situ surface measurements in Israeli cities, Atmos. Chem. Phys., 9, 3867-3879, doi:10.5194/acp-9-3867-2009, 2009.

Boersma, K. F., Eskes, H. J., Dirksen, R. J., van der A, R. J., Veefkind, J. P., Stammes, P., Huijnen, V., Kleipool, Q. L., Sneep, M., Claas, J., Leitão, J., Richter, A., Zhou, Y., and Brunner, D.: An improved tropospheric NO_2 column retrieval algorithm for the Ozone Monitoring Instrument, Atmos. Meas. Tech., 4, 1905-1928, doi:10.5194/amt-4-1905-2011, 2011.

Buriez, J. C., F. Parol, C. Cornet, M. Doutriaux-Boucher, An improved derivation of the top-of-atmosphere albedo from POLDER/ADEOS-2: Narrowband albedos, J. Geophys. Res., 110, doi:10:1029/2004JD005243, 2005.

Cerquiera, M. A., Pio, C. A., Gomes, P. A., Matos, J. S., and Nunes, T. V.: Volatile organic compounds in rural atmospheres in central Portugal, The Science of the Total Environment, 313, 49-60, doi:10.1016/S0048-9697(03)00250-X, 2003.

Curier, R. L., Kranenburg, R., Segers, A. J. S., Timmermans, R. M. A., and Schaap, M.: Synergistic use of OMI NO_2 tropospheric columns and LOTOS-EUROS to evaluate the NO_x emissions trends across Europe, Remote Sensing of the Environment, 149(2014), 58-69, dx.doi.org/10.1016/j.rse.2014.03.032, 2014.

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

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., et al.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Q.J. Roy. Meteorol. Soc., 137, 553-597. doi: 10.1002/qj.828, 2011.

Dirksen, R. J., Boersma, K. F., Eskes, H. J., Ionov, D. V., Bucsela, E. J., Levelt, P. F., and Kelder, H. M.: Evaluation of stratospheric NO_2 retrieved from the Ozone Monitoring Instrument: Intercomparison, diurnal cycle, and trending, J. Geophys. Res., 116, D08305, doi:10.1029/2010JD014943, 2011.

Dentener, F., Kinne, S., Bond, T., Boucher, O., Cofala, J., Generoso, S., Ginoux, P., Gong, S., Hoelzemann, J. J., Ito, A., Marelli, L., Penner, J. E., Putaud, J.-P., Textor, C., Schulz, M., van der Werf, G. R., and Wilson, J.: Emissions of primary aerosol and precursor gases in the years 2000 and 1750 prescribed data-sets for AeroCom, Atmos. Chem. Phys., 6, 4321–4344, doi:10.5194/acp-6-4321-2006, 2006.

Dufour, G., Wittrock, F., Camredon, M., Beekmann, M., Richter, A., Aumont, B., and Burrows, J. P.: SCIAMACHY formaldehyde observations: constraint for isoprene emission estimates over Europe?, Atmos. Chem. Phys., 9, 1647-1664, doi:10.5194/acp-9-1647-2009, 2009.

Duncan, B. N., Yoshida, Y., Olson, J. R., Sillman, S., Martin, R. V., Lamsal, L., Hu, Y. Pickering, K. E., Retscher, C., Allen, D. J., Crawford, J. H.: Application of OMI observations to a space-based indicator of NO_x and VOC controls on surface ozone formation: Atmospheric Environment, 44(18), 2213-2223, doi:10.1016/j.atmosenv.2010.03.010, 2010.

Duncan, B. N., Prados, A. I., Lamsal, L. N., Liu, Y., Streets, D. G., Gupta, P., Hilsenrath, E., Kahn, R. A., Nielsen, J. E., Beyersdorf, A. J., Burton, S. P., Fiore, A. M., Fishman, J., Henze, D. K., Hostetlere, C. A., Krotkov, N. A., Lee, P., Lin, M., Pawson, S. Pfister, G., Pickering, K. E., Pierce, R. B., Yoshida, Y., and Ziemba, L.D.: Satellite data of atmospheric pollution for U.S. air quality applications: Examples of applications, summary of data end-use resources, answers to FAQs, and common mistakes to avoid, Atmospheric Environment, 94, 647-662, doi:10.1016/j.atmosenv.2014.05.061, 2014.

Eskes, H. J., van Velthoven, P. F. J., Valks, P. J. M., and Kelder, H. M.: Assimilation of GOME total-ozone satellite observations in a three-dimensional tracer-transport model, Q. J. Roy. Meteorol. Soc., 129, 1663-1681, doi:10.1256/qj.02.14, 2003.

Evans, M. J., and Jacob, D. J.: Impact of new laboratory studies of N_2O_5 hydrolysis on global model budgets of tropospheric nitrogen oxides, ozone, and OH, Geophys. Res. Lett., 32, L09813, doi:10.1029/2005GL022469, 2005.

Fried, A., McKeen, S., Sewell, S., Harder, J., Henry, B., Goldan, P., Kuster, W., Williams, E., Baumann, K., Shetter, R., and Cantrell, C.: Photochemistry of formaldehyde during the 1993 Tropospheric OH Photochemistry Experiment, J. Geophys. Res., 102(D5), 6283-6296, 1997.

Gery, M. W., Whitten, G. Z., Killus, J. P., & Dodge, M. C.: A photochemical kinetics mechanism for urban and regional scale computer modeling. J. Geophys. Res., 94(D10), 12925-12956, 1989.

González Abad, G., Liu, X., Chance, K., Wang, H., Kurosu, T. P., and Suleiman, R.: Updated SAO OMI formaldehyde retrieval, Atmos. Meas. Tech. Discuss., 7, 1-31, doi:10.5194/amtd-7-1-2014, 2014.

Herron-Thorpe, Lamb, B. K., Mount, G. H., Vaughan, J. K.: Evaluation of a regional air quality forecast model for tropospheric NO2 columns using the OMI/Aura satellite tropospheric NO2 product, Atmos. Chem. Phys., 10, 8839–8854, doi:10.5194/acp-10-8839-2010, 2010.

Holmes, C. D., Jacob, D. J., Corbitt, E. S., Mao, J., Yang, X., Talbot, R., and Slemr, F.: Global atmospheric model for mercury including oxidation by bromine atoms, Atmos. Chem. Phys., 10, 12037-12057, doi:10.5194/acp-10-12037-2010, 2010.

Hooghiemstra, P. B., M. C. Krol, P. Bergamaschi, A. T. J. deLaat, G. R. van derWerf, P. C. Novelli, M. N. Deeter, I. Aben, and T. Röckmann, Comparing optimized CO emission estimates using MOPITT or NOAA surface network observations, J. Geophys. Res., 117, D06309, doi:10.1029/2011JD017043, 2012.

Huijnen, V., Eskes, H. J., Poupkou, A., Elbern, H., Boersma, K. F., Foret, G., Sofiev, M., Valdebenito, A., Flemming, J., Stein, O., Gross, A., Robertson, L., D'Isidoro, M., Kioutsioukis, I., Friese, E., Amstrup, B., Bergstrom, R., Strunk, A., Vira, J., Zyryanov, D., Maurizi, A., Melas, D., Peuch, V.-H., and Zerefos, C.: Comparison of OMI NO_2 tropospheric columns with an ensemble of global and European regional air quality models, Atmos. Chem. Phys., 10, 3273-3296, doi:10.5194/acp-10-3273-2010, 2010[a].

Huijnen, V., Williams, J. E., van Weele, M., van Noije, T. P. C., Krol, M. C., Dentener, F., Segers, A., Houweling, S., Peters, W., de Laat, A. T. J., Boersma, K. F., Bergamaschi, P., van Velthoven, P. F. J., Le Sager, P., Eskes, H. J., Alkemade, F., Scheele, M. P., Nédélec, P., and Pätz, H.-W.: The global chemistry transport model TM5: description and evaluation of the tropospheric chemistry version 3.0, Geosci. Model Dev. Discuss., 3, 1009-1087, doi:10.5194/gmdd-3-1009-2010, 2010[b].

Ide, K., Courtier, P., Ghil, M., Lorenc, A. C., Unified Notation for Data Assimilation: Operational, Sequential, and Variational, J. Met. Soc. Japan, Special Issue on “Data Assimilation in Meteorology and Oceanography: Theory and Practice, Vol. 75, No. 1B, 181-189, 1997 .

Inness, A., Baier, F., Benedetti, A., Bouarar, I., Chabrillat, S., Clark, H., et al. (2013). The MACC reanalysis: An 8 yr data set of atmospheric composition. *Atmospheric Chemistry and Physics, 13*, 4073-4109. doi:10.5194/acp-13-4073-2013, 2013.

Irie, H., Boersma, K. F., Kanaya, Y., Takashima, H., Pan, X., and Wang, Z. F.: Quantitative bias estimates for tropospheric NO_2 columns retrieved from SCIAMACHY, OMI, and GOME-2 using a common standard for East Asia, Atmos. Meas. Tech., 5, 2403-2411, doi:10.5194/amt-5-2403-2012, 2012.

Jones, D. B. A., Bowman, K. W., Palmer, P. I., Worden, J. D., Jacob, D. J., Hoffman, R. N., Bey, I., and Yantosca, R. M., Potential of observations from the Tropospheric Emission Spectrometer to constrain continental sources of carbon monoxide, J. Geophys. Res., 108(D24), 4789, doi:10.1029/2003JD003702, 2003.

Kleipool, Q. L., Dobber, M. R., de Haan, J., & Levelt, P. F., Earth surface reflectance climatology from 3 years of OMI data. *J. Geophys. Res.*, *113*(D18), D18308, 2008.

Krol, M., Houweling, S., Bregman, B., van den Broek, M., Segers, A., van Velthoven, P., Peters, W., Dentener, F., and Bergamaschi, P.: The two-way nested global chemistry-transport zoom model TM5: algorithm and applications, Atmos. Chem. Phys., 5, 417-432, doi:10.5194/acp-5-417-2005, 2005.

Krol, M., Peters, W., Hooghiemstra, P., George, M., Clerbaux, C., Hurtmans, D., McInerney, D., Sedano, F., Bergamaschi, P., El Hajj, M., Kaiser, J. W., Fisher, D., Yershov, V., and Muller, J.-P.: How much CO was emitted by the 2010 fires around Moscow?, Atmos. Chem. Phys., 13, 4737-4747, doi:10.5194/acp-13-4737-2013, 2013.

Lamsal, L. N., R. V. Martin, R. V., van Donkelaar, A., Celarier, E. A., Bucsela, E. J., Boersma, K. F., Dirksen, R., Luo, C., and Wang, Y.: Indirect validation of tropospheric nitrogen dioxide retrieved from the OMI satellite instrument: Insight into the seasonal variation of nitrogen oxides at northern midlatitudes, J. Geophys. Res., 115, D05302, doi:10.1029/2009JD013351, 2010.

Lamsal, L. N., Krotkov, N. A., Celarier, E. A., Swartz, W. H., Pickering, K. E., Bucsela, E. J., Gleason, J. F., Martin, R. V., Philip, S., Irie, H., Cede, A., Herman, J., Weinheimer, A., Szykman, J. J., and Knepp, T. N.: Evaluation of OMI operational standard NO_2 column retrievals using in situ and surface-based NO_2 observations, Atmos. Chem. Phys., 14, 11587-11609, doi:10.5194/acp-14-11587-2014, 2014.

Lathière, J., Hauglustaine, D. A., Friend, A. D., De Noblet-Ducoudré, N., Viovy, N., and Folberth, G. A.: Impact of climate variability and land use changes on global biogenic volatile

organic compound emissions, Atmos. Chem. Phys., 6, 2129-2146, doi:10.5194/acp-6-2129-2006, 2006.

Lee, C., Martin, R. V., van Donkelaar, A., O'Byrne, G., Krotkov, N., Richter, A., Huey, L. G., and Holloway, J. S., Retrieval of vertical columns of sulfur dioxide from SCIAMACHY and OMI: Air mass factor algorithm development, validation, and error analysis, J. Geophys. Res., 114, D22303, doi:10.1029/2009JD012123, 2009.

Lin, J.-T., and McElroy, M. B.: Detection from space of a reduction in anthropogenic emissions of nitrogen oxides during the Chinese economic downturn, Atmos. Chem. Phys., 11, 8171-8188, doi:10.5194/acp-11-8171-2011, 2011.

Lin, J.-T., Liu, Z., Zhang, Q., Liu, H., Mao, J., and Zhuang, G.: Modeling uncertainties for tropospheric nitrogen dioxide columns affecting satellite-based inverse modeling of nitrogen oxides emissions, Atmos. Chem. Phys., 12, 12255-12275, doi:10.5194/acp-12-12255-2012, 2012[a].

Lin, J.-T.: Satellite constraint for emissions of nitrogen oxides from anthropogenic, lightning and soil sources over East China on a high-resolution grid, Atmos. Chem. Phys., 12, 2881-2898, doi:10.5194/acp-12-2881-2012, 2012[b].

Ma, J. Z., Beirle, S., Jin, J. L., Shaiganfar, R., Yan, P., and Wagner, T.: Tropospheric NO_2 vertical column densities over Beijing: results of the first three years of ground-based MAX-DOAS measurements (2008–2011) and satellite validation, Atmos. Chem. Phys., 13, 1547-1567, doi:10.5194/acp-13-1547-2013, 2013.

Maasakkers, J. D.: Vital improvements to the retrieval of tropospheric NO_2 columns from the Ozone Monitoring Instrument, M.Sc Report, R-1882-A, July 2013, Eindhoven University of Technology, 2013.

Mao, J., Jacob, D. J., Evans, M. J., Olson, J. R., Ren, X., Brune, W. H., Clair, J. M. St., Crounse, J. D., Spencer, K. M., Beaver, M. R., Wennberg, P. O., Cubison, M. J., Jimenez, J. L., Fried, A., Weibring, P., Walega, J. G., Hall, S. R., Weinheimer, A. J., Cohen, R. C., Chen, G., Crawford, J. H., McNaughton, C., Clarke, A. D., Jaeglé, L., Fisher, J. A., Yantosca, R. M., Le Sager, P., and Carouge, C.: Chemistry of hydrogen oxide radicals (HO_x) in the Arctic troposphere in spring, Atmos. Chem. Phys., 10, 5823-5838, doi:10.5194/acp-10-5823-2010, 2010.

Martin, R. V., Fiore, A. M., and van Donkelaar, A.: Space-based diagnosis of surface ozone sensitivity to anthropogenic emissions, Geophys. Res. Lett., 31, L06120, doi:10.1029/2004GL019416, 2004.

McLinden, C. A., Fioletov, V., Boersma, K. F., Kharol, S. K., Krotkov, N., Lamsal, L., Makar, P. A., Martin, R. V., Veefkind, J. P., and Yang, K.: Improved satellite retrievals of NO_2 and SO_2 over the Canadian oil sands and comparisons with surface measurements, Atmos. Chem. Phys., 14, 3637-3656, doi:10.5194/acp-14-3637-2014, 2014.

Migliorini, S., Piccolo, C., and Rodgers, C. D.: Use of the Information Content in Satellite Measurements for an Efficient Interface to Data Assimilation, Mon. Wea. Rev., 136, 2633-2650, doi:10.1175/2007MWR2236.1, 2008.

Millet, D. B., Jacob, D. J., Boersma, K. F., Fu, T. M., Kurosu, T. P., Chance, K., Heald, C. L., and Guenther, A.: Spatial distribution of isoprene emissions from North America derived from formaldehyde column measurements by the OMI satellite sensor, J. Geophys. Res., 113, D02307, doi:10.1029/2007JD008950, 2008.

Miyazaki, K., H. J. Eskes, K. Sudo, M. Takigawa, M. van Weele, and K. F. Boersma, Simultaneous assimilation of satellite NO2, O3, CO, and HNO3 data for the analysis of satellite NO2, O3, CO, and HNO3 data for the analysis of tropospheric chemical composition and emissions, Atmos. Chem. Phys., 12, 9545-9579, doi:10.5194/acp-12-9545-2012, 2012.

Mijling, G., and van der A, R.J.: Using daily satellite observations to estimate emissions of short-lived air pollutants on a mesoscopic scale, J. Geophys. Res., 117, D17302, doi:10.1029/2012JD017817, 2012.

Morcrette, J. J., & Jakob, C., The response of the ECMWF model to changes in the cloud overlap assumption. *Monthly Weather Review*, *128*(6), 1707-1732, 2000.

Müller, J.-F., Stavrakou, T., Wallens, S., De Smedt, I., Van Roozendael, M., Potosnak, M. J., Rinne, J., Munger, B., Goldstein, A., and Guenther, A. B.: Global isoprene emissions estimated using MEGAN, ECMWF analyses and a detailed canopy environment model, Atmos. Chem. Phys., 8, 1329-1341, doi:10.5194/acp-8-1329-2008, 2008.

Olivier, J. G. J. and Berdowski, J. J. M.: Global emissions sources and sinks, in The Climate System, Lisse, Netherlands, 33–78, 2001.

Platt, U., and Stutz, J.: Differential Optical Absorption Spectroscopy: Principles and Applications, Springer, 2008.

Quaas, J., Final versions of CALIPSO-PARASOL observational analysis product and of MODIS simulator, Deliverable D1.2: Final version of MODIS simulator, http://www.euclipse.eu/downloads/D1.2_euclipse_modissimulator.pdf, 2011.

Richter, A., Wittrock, F., & Burrows, J. P., SO_2 measurements with SCIAMACHY, In Proc. Atmospheric Science ConferenceF (pp. 8-12), 2006.

Rodgers, C. D., Inverse methods for atmospheric sounding: Theory and Practice, Series on Atmospheric, Oceanic and Planetary Physics–Vol. 2., Singapore, World Scientific (2000).

Rodgers, C. D., and B. J. Connor, Intercomparison of remote sounding instruments, J. Geophys. Res., 108(D3), 4116, doi:10.1029/2002JD002299, 2003.

Schaub, D., Boersma, K. F., Kaiser, J. W., Weiss, A. K., Folini, D., Eskes, H. J., and Buchmann, B.: Comparison of GOME tropospheric NO_2 columns with NO_2 profiles deduced from ground-based in situ measurements, Atmos. Chem. Phys., 6, 3211-3229, doi:10.5194/acp-6-3211-2006, 2006.

Schaub, D., Brunner, D., Boersma, K. F., Keller, J., Folini, D., Buchmann, B., Berresheim, H., and Staehelin, J.: SCIAMACHY tropospheric NO_2 over Switzerland: estimates of NO_x lifetimes and impact of the complex Alpine topography on the retrieval, Atmos. Chem. Phys., 7, 5971-5987, doi:10.5194/acp-7-5971-2007, 2007.

Stammes, P., Sneep, M., de Haan, J. F., Veefkind, J. P., Wang, P., and Levelt, P. F.: Effective cloud fractions from the Ozone Monitoring Instrument: Theoretical framework and validation, J. Geophys. Res., 113, D16S38, doi:10.1029/2007JD008820, 2008.

Stavrakou, T., Müller, J.-F., Boersma, K. F., van der A, R. J., Kurokawa, J., Ohara, T., and Zhang, Q.: Key chemical NO_x sink uncertainties and how they influence top-down emissions of nitrogen oxides, Atmos. Chem. Phys., 13, 9057-9082, doi:10.5194/acp-13-9057-2013, 2013.

Theys, N., Campion, R., Clarisse, L., Brenot, H., van Gent, J., Dils, B., Corradini, S., Merucci, L., Coheur, P.-F., Van Roozendael, M., Hurtmans, D., Clerbaux, C., Tait, S., and Ferrucci, F.: Volcanic SO_2 fluxes derived from satellite data: a survey using OMI, GOME-2,

IASI and MODIS, Atmos. Chem. Phys., 13, 5945-5968, doi:10.5194/acp-13-5945-2013, 2013.

Uno, I., He, Y., Ohara, T., Yamaji, K., Kurokawa, J.-I.: Systematic analysis of interannual and seasonal variations of model-simulated tropospheric NO_2 in Asia and comparison with GOME-satellite data. Atmos. Chem. Phys., 7, 1671-1681, 2007.

van der Werf, G. R., Randerson, J. T., Giglio, L., Collatz, G. J., Kasibhatla, P. S., and Arellano Jr., A. F.: Interannual variability in global biomass burning emissions from 1997 to 2004, Atmos. Chem. Phys., 6, 3423–3441, doi:10.5194/acp-6-3423-2006, 2006.

van Geffen, J. H. G. M., Boersma, K. F., Van Roozendael, M., Hendrick, F., Mahieu, E., De Smedt, I., Sneep, M., and Veefkind, J. P.: Improved spectral fitting of nitrogen dioxide from OMI in the 405–465 nm window, Atmos. Meas. Tech., 8, 1685-1699, doi:10.5194/amt-8-1685-2015, 2015.

Vinken, G. C. M., Boersma, K. F., Jacob, D. J., and Meijer, E. W.: Accounting for non-linear chemistry of ship plumes in the GEOS-Chem global chemistry transport model, Atmos. Chem. Phys., 11, 11707-11722, doi:10.5194/acp-11-11707-2011, 2011.

Vinken, G. C. M., Boersma, K. F., van Donkelaar, A., and Zhang, L.: Constraints on ship NOx emissions in Europe using GEOS-Chem and OMI satellite NO_2 observations, Atmos. Chem. Phys., 14, 1353-1369, doi:10.5194/acp-14-1353-2014, 2014.

von Hardenberg, J., Vozella, L., Tomasi, C., Vitale, V., Lupi, A., Mazzola, M., van Noije, T. P. C., Strunk, A., and Provenzale, A.: Aerosol optical depth over the Arctic: a comparison of ECHAM-HAM and TM5 with ground-based, satellite and reanalysis data, Atmos. Chem. Phys., 12, 6953-6967, doi:10.5194/acp-12-6953-2012, 2012.

Vrekoussis, M., Wittrock, F., Richter, A., and Burrows, J. P.: Temporal and spatial variability of glyoxal as observed from space, Atmos. Chem. Phys., 9, 4485-4504, doi:10.5194/acp-9-4485-2009, 2009.

Wang, Y., McElroy, M. B., Martin, R. V., Streets, D. G., Zhang, Q., and Fu, T.-M.: Seasonal variability of NO_x emissions over east China constrained by satellite observations: Implications for combustion and microbial sources, J. Geophys. Res., 112, D06301, doi:10.1029/2006JD007538, 2007.

Wang, X., Mallet, V, Berroir, J.-P., and Herlin, I.: Assimilation of OMI NO_2 retrievals into a regional chemistry-transport model for improving air quality forecasts over Europe, Atm. Environm., 45 (2011), 485-492, doi:10.1016/j.atmosenv.2010.09.028, 2011.

Williams, J. E., Scheele, M. P., van Velthoven, P. F. J., Cammas, J.-P., Thouret, V. Galy-Lacaux, C., and Volz-Thomas, A.: The influence of biogenic emissions from Africa on tropical tropospheric ozone during 2006: a global modeling study, Atmos. Chem. Phys., 9, 5729-5749, doi:10.5194/acp-9-5729-2009, 2009.

Williams, J. E., Strunk, A., Huijnen, V., and van Weele, M.: The application of the Modified Band Approach for the calculation of on-line photodissociation rate constants in TM5: implications for oxidative capacity, Geosci. Model Dev., 5, 15-35, doi:10.5194/gmd-5-15-2012, 2012.

Williams, J. E., Le Bras, G., Kukui, A., Ziereis, H., and Brenninkmeijer, C. A. M.: The impact of the chemical production of methyl nitrate from the NO + CH_3O_2 reaction on the global distributions of alkyl nitrates, nitrogen oxides and tropospheric ozone: a global modelling study, Atmos. Chem. Phys., 14, 2363-2382, doi:10.5194/acp-14-2363-2014, 2014.

Zhang, L., Jacob, D. J., Knipping, E. M., Kumar, N., Munger, J. W., Carouge, C. C., van Donkelaar, A., Wang, Y. X., and Chen, D.: Nitrogen deposition to the United States: distribution, sources, and processes, Atmos. Chem. Phys., 12, 4539-4554, doi:10.5194/acp-12-4539-2012, 2012.

Zhang, Q., Streets, D. G., and He, K.: Satellite observations of recent power plant construction in Inner Mongolia, China, Geophys. Res. Lett., 36, L15809, doi:10.1029/2009GL038984, 2009.

Table 1. Summary of tropospheric NO_2 GEOS-Chem model evaluations following recipes (A), (B), and (C) with OMI NO_2 retrievals for February and August 2006. *n* refers to the number of grid cells used in the comparison.

	R^2			Geometric mean			
Model evaluation	(A)	(B)	(C)	(A)	(B)	(C)	*n*
Europe Feb 2006	0.66	0.63	0.54	$1.15^{1.88}_{0.70}$	$1.13^{1.89}_{0.67}$	$0.92^{1.59}_{0.54}$	120
Europe Aug 2006	0.66	0.57	0.59	$1.24^{1.80}_{0.86}$	$1.42^{2.21}_{0.91}$	$1.40^{2.13}_{0.92}$	137
US Feb 2006	0.82	0.79	0.75	$1.12^{1.29}_{0.97}$	$1.08^{1.34}_{0.87}$	$0.95^{1.25}_{0.72}$	41
US Aug 2006	0.83	0.61	0.67	$0.75^{0.95}_{0.59}$	$0.91^{1.21}_{0.68}$	$0.90^{1.16}_{0.70}$	42
China Feb 2006	0.58	0.57	0.54	$1.00^{1.37}_{0.72}$	$0.99^{1.63}_{0.59}$	$0.86^{1.08}_{0.69}$	35
China Aug 2006	0.61	0.58	0.58	$1.13^{1.39}_{0.92}$	$1.07^{1.36}_{0.84}$	$1.06^{1.37}_{0.82}$	44

Table 2. Overview of magnitude and nature of various model sampling errors, their contribution to the overall comparison error budget, and ways to avoid them. Based on the GEOS-Chem evaluation with OMI NO_2 retrievals for February and August 2006. Note that these recommendations hold for air pollution applications of UV/Vis satellite retrievals such as model evaluation and top-down emission estimates.

	Relative error	**Type of error**	**Recommendation**
Horizontal sampling	<5-10%	Inevitable and random	Require at least 40% coverage of model grid cell.
Temporal sampling	10%	Avoidable and systematic	Sample model grid cells exclusively on clear-sky days
Vertical sampling	20%	Avoidable and systematic	Apply averaging kernel on model vertical distribution
Overall representativeness error	10%-25%		Follow recommendations listed above to keep $\sigma_r < \sigma_o$