Response to the editor

Remark: Editor comments are printed in bold, our responses in regular font, explicit changes made in the manuscript in italic font.

After 2nd review, the author response file contains only responses to reviewer #1, but responses to the minor comments and suggestions made by reviewer #2 is missing; no changes appear to have been implemented in the revised paper. The editor considers this a simple oversight, but reviewer #2 should not be ignored.

In fact, the missing responses to reviewer 2 were due to an oversight, we apologize. Please find the responses below.

Both reviewers had raised the likelihood of a connection between PAC and SF < 1, and reviewer #1 had pointed to a missing paper (Ortega et al., 2016) that seems relevant here. Reviewer #2 had recommended to soften language that suggests SF and PAC "can be regarded as independent" (Section 3.4). The editor agrees that such an affirmative statement may be too strong, and can easily be misread; it further seems somewhat at odds with the authors own conclusion that "O4 scaling and PAC were found to have similar impact on the MAX-DOAS AOT results" (Conclusions). Consider to adopt the suggestion made by reviewer #2.

We agree that the statement is too unspecific and strong and should be softened. However, it should clearly be stated that the PAC (the removal of an OEM mathematical artefact) cannot provide new insights on whether aerosol aloft is responsible for the SF (a finding drawn from forward simulations in Ortega 2016) or not. Further, we would like to emphasize that our statement in the conclusions "O4 scaling and PAC were found to have similar impact on the MAX-DOAS AOT results" is not generally valid, as the PAC correction factors (f_tau) are determined by the a priori assumptions. In fact, according to equation 9 and 10 in the manuscript, any f_tau between zero and infinity can be produced applying corresponding a priori assumptions. We therefore changed the corresponding sentences to:

"Thus their findings can in general be regarded as independent from any kind of PAC, even though PAC and SF have similar impact on the MAX-DOAS AOT results with the a priori assumptions applied in this study. Particularly, it shall be pointed out that our findings regarding the PAC have no implications on whether elevated aerosol layers explain the necessity of the SF (as proposed by Ortega, 2016), or not."

"With the a priori settings applied in this study, O4 scaling and PAC were found to have similar impact on the MAX-DOAS AOT results."

Furthermore, as reviewer #1 points out, the Ortega et al. is not only a "source of bias" (as the current revisions state), but establishes that "lack of sensitivity aloft" (read "AOT-PAC = elevated layers aloft") and SF are plausibly connected. This is currently not said clearly enough in the revised paper. A simple solution may be to add a half sentence "..., consistent with earlier findings (Ortega et al., 2016)." behind the above sentence in the Conclusions.

See our response above and our reasoning in the manuscript section 3.4. particularly P30, L14-28. Based on that, we believe that our findings regarding the PAC do not allow to doubt or support findings on the SF in former publications.

Ortega 2016 has been added as a reference in the course of the response above.

Also consider to cite Ortega et al. 2016 together with Wagner et al., 2019 at the end of that same paragraph.

We added Ortega as a second reference here.

Belated responses to reviewer #2

Reviewer comments are printed in bold, our responses in regular font.

Reviewer #2: I have the two minor technical suggestions on wording:
While it is reasonably clear in the context of the response the referee, I think the following passage could be misinterpreted in the text: "In this way, its contribution to the total variance observed among the participants under clear sky conditions can be estimated to 40 % (for AOTs), 85 % (HCHO VCDs), 70 % (HCHO surface concentrations), 50 % (NO2 VCDs), 40 % (NO2 UV surface concentrations) and 20 % (NO2 Vis surface concentrations), respectively. The residual variance can be attributed to the choice and setup of the retrieval algorithm." Could the authors reword to make it more apparent that the reported variance contributions are those that arise from the measurements?

Response: We reworded the sentence accordingly. It reads now:

"For clear-sky conditions, we find that the differences in the measured dSCDs are responsible for approximately 40% (for AOTs), 85% (HCHO VCDs), 70% (HCHO surface concentrations) and 50% (NO2) of the total variance observed among the participants. The residual variance can be attributed to the choice and setup of the retrieval algorithm."

My comment regarding the PAC and SF comparison has been well addressed. Looking at the comments of Referee 1, however, there is a connection between these concepts documented in the literature, even if the the motivations for them are different. This study does not seem to clearly indicate one way or the other whether the concepts are related. Read in context the sentence near the top of page 32: "Thus their findings can be considered independent from any kind of PAC." is phrased as a conclusion and seems to imply that the concepts should be considered separately. While it is important that the comparison of PAC and SF is addressed think even this conditional conclusion is phrased to strongly. It is also possible that despite the different motivations, that the PAC and SF are related and considering them together might be fruitful, we do not know. I would recommend softening the language to something like:

Thus their findings might be considered independent from any kind of PAC.

See our response to the editor above.

Marked-up Manuscript

Intercomparison of MAX-DOAS vertical profile retrieval algorithms: studies on field data from the CINDI-2 campaign

Jan-Lukas Tirpitz[1], Udo Frieß[1], François Hendrick[2], Carlos Alberti[3,a], Marc Allaart[4], Arnoud Apituley[4], Alkis Bais[5], Steffen Beirle[6], Stijn Berkhout[7], Kristof Bognar[8], Tim Bösch[9], Ilya Bruchkouski[10], Alexander Cede[11,12], Ka Lok Chan[3,b], Mirjam den Hoed[4], Sebastian Donner[6], Theano Drosoglou[5], Caroline Fayt[2], Martina M. Friedrich[2], Arnoud Frumau[13], Lou Gast[7], Clio Gielen[2,c], Laura Gomez-Martín[14], Nan Hao[15], Arjan Hensen[13], Bas Henzing[13], Christian Hermans[2], Junli Jin[16], Karin Kreher[18], Jonas Kuhn[1,6], Johannes Lampel[1,19], Ang Li[20], Cheng Liu[21], Haoran Liu[21], Jianzhong Ma[17], Alexis Merlaud[2], Enno Peters[9,d], Gaia Pinardi[2], Ankie Piters[4], Ulrich Platt[1,6], Olga Puentedura[14], Andreas Richter[9], Stefan Schmitt[1], Elena Spinei[12,e], Deborah Stein Zweers[4], Kimberly Strong[8], Daan Swart[7], Frederik Tack[2], Martin Tiefengraber[11,22], René van der Hoff[7], Michel van Roozendael[2], Tim Vlemmix[4], Jan Vonk[7], Thomas Wagner[6], Yang Wang[6], Zhuoru Wang[15], Mark Wenig[3], Matthias Wiegner[3], Folkard Wittrock[9], Pinhua Xie[20], Chengzhi Xing[21], Jin Xu[20], Margarita Yela[14], Chengxin Zhang[21], and Xiaoyi Zhao[8,f]

[1]Institute of Environmental Physics, University of Heidelberg, Heidelberg, Germany
[2]Royal Belgian Institute for Space Aeronomy, Brussels, Belgium
[3]Meteorological Institute, Ludwig-Maximilians-Universität München, Munich, Germany
[4]Royal Netherlands Meteorological Institute (KNMI), De Bilt, The Netherlands
[5]Laboratory of Atmospheric Physics, Aristotle University of Thessaloniki, Thessaloniki, Greece
[6]Max Planck Institute for Chemistry, Mainz, Germany
[7]National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands
[8]Department of Physics, University of Toronto, Toronto, Canada
[9]Institute for Environmental Physics, University of Bremen, Bremen, Germany
[10]Belarusian State University, Minsk, Belarus
[11]LuftBlick Earth Observation Technologies, Mutters, Austria
[12]NASA-Goddard Space Flight Center, USA
[13]Netherlands Organisation for Applied Scientific Research (TNO), Utrecht, The Netherlands
[14]National Institute of Aerospatial Technology (INTA), Madrid, Spain
[15]Remote Sensing Technology Institute, German Aerospace Center (DLR), Oberpfaffenhofen, Germany
[16]Meteorological Observation Centre, China Meteorological Administration, Beijing, China
[17]Chinese Academy of Meteorology Science, China Meteorological Administration, Beijing, China
[18]BK Scientific GmbH, Mainz, Germany
[19]Airyx GmbH, Justus-von-Liebig-Straße 14, 69214 Eppelheim, Germany
[20]Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei, China
[21]School of Earth and Space Sciences, University of Science and Technology of China, 230026, Hefei, China
[22]Department of Atmospheric and Cryospheric Sciences, University of Innsbruck, Innsbruck, Austria
[a]now at Institute of Meteorology and Climate Research (IMK-ASF), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
[b]now at Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, Germany
[c]now at Institute for Astronomy, KU Leuven, Belgium
[d]now at Institute for Protection of Maritime Infrastructures, Bremerhaven, Germany
[e]now at Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

[f]now at Air Quality Research Division, Environment and Climate Change Canada, Canada
Correspondence: Jan-Lukas Tirpitz (jan-lukas.tirpitz@iup.uni-heidelberg.de)

Abstract.

The second Cabauw Intercomparison of Nitrogen Dioxide measuring Instruments (CINDI-2) took place in Cabauw (The Netherlands) in September 2016 with the aim of assessing the consistency of MAX-DOAS measurements of tropospheric species (NO_2, HCHO, O_3, HONO, CHOCHO and O_4). This was achieved through the coordinated operation of 36 spectrometers operated by 24 groups from all over the world, together with a wide range of supporting reference observations (in situ analysers, balloon sondes, lidars, Long-Path DOAS, direct-sun DOAS, sun photometer and meteorological instruments).

In the presented study, the retrieved CINDI-2 MAX-DOAS trace gas (NO_2, HCHO) and aerosol vertical profiles of 15 participating groups using different inversion algorithms are compared and validated against the colocated supporting observations, with the focus on aerosol optical thicknesses (AOTs), trace gas vertical column densities (VCDs) and trace gas surface concentrations. The algorithms are based on three different techniques: six use the optimal estimation method, two use a parametrized approach and one algorithm relies on simplified radiative transport assumptions and analytical calculations. To assess the agreement among the inversion algorithms independent of inconsistencies in the trace gas slant column density acquisition, participants applied their inversion to a common set of slant columns. Further, important settings like the retrieval grid, profiles of O_3, temperature and pressure as well as aerosol optical properties and a priori assumptions (for optimal estimation algorithms) have been prescribed to reduce possible sources of discrepancies.

The profiling results were found to be in good qualitative agreement: most participants obtained the same features in the retrieved vertical trace gas and aerosol distributions, however sometimes at different altitudes and of different magnitude. Under clear sky conditions, the root-mean-square differences (RMSDs) among the results of individual participants vary between $(0.01-0.1)$ for AOTs, $(1.5-15)\times10^{14}\,\mathrm{molec\,cm^{-2}}$ for trace gas (NO_2, HCHO) VCDs and $(0.3-8)\times10^{10}\,\mathrm{molec\,cm^{-3}}$ for trace gas surface concentrations. These values compare to approximate average optical thicknesses of 0.3, trace gas vertical columns of $90\times10^{14}\,\mathrm{molec\,cm^{-2}}$ and trace gas surface concentrations of $11\times10^{10}\,\mathrm{molec\,cm^{-3}}$ observed over the campaign period. The discrepancies originate from differences in the applied techniques, the exact implementation of the algorithms and the user defined settings that were not prescribed.

For the comparison against supporting observations, the RMSDs increase to $(0.02-0.2)$ against AOTs from the sun photometer, $(11-55)\times10^{14}\,\mathrm{molec\,cm^{-2}}$ against trace gas VCDs from direct-sun DOAS observations and $(0.8-9)\times10^{10}\,\mathrm{molec\,cm^{-3}}$ against surface concentrations from the Long-Path DOAS instrument. This increase in RMSDs is most likely caused by uncertainties in the supporting data themselves, spatio-temporal mismatch among the observations and simplified assumptions particularly on aerosol optical properties made for the MAX-DOAS retrieval.

As a side investigation, the comparison was repeated with the participants retrieving profiles from their own dSCDs aquired during the campaign. In this case, the consistency among the participants degrades by about $30\,\%$ for AOTs, by $180\,\%$ ($40\,\%$) for HCHO (NO_2) VCDs and by $90\,\%$ ($20\,\%$) for HCHO (NO_2) surface concentrations.

In former publications and also during this comparison study, it was found that MAX-DOAS vertically integrated aerosol extinction coefficient profiles systematically underestimate the AOT observed by the sun photometer. For the first time it is quantitatively shown that for optimal estimation algorithms this can be largely explained and compensated by considering biases arising from the reduced sensitivity of MAX-DOAS observations to higher altitudes and associated *a priori* assumptions.

1 Introduction

The planetary boundary layer (PBL) is the lowest part of the atmosphere, whose behaviour is directly influenced by its contact with the Earth's surface. Its chemical composition and aerosol load is driven by the exchange with the surface, transport processes and homogeneous and heterogeneous chemical reactions. Monitoring of both, trace gases and aerosols, preferably simultaneous, is crucial for the understanding of the spatio-temporal evolution of the PBL composition and the chemical and physical processes.

Multi-AXis Differential Optical Absorption Spectroscopy (MAX-DOAS) (e.g. Hönninger and Platt, 2002; Hönninger et al., 2004; Wagner et al., 2004; Heckel et al., 2005; Frieß et al., 2006; Platt and Stutz, 2008; Irie et al., 2008; Clémer et al., 2010; Wagner et al., 2011; Vlemmix et al., 2015b) is a widely used ground-based measurement technique for the detection of aerosols and trace gases particularly in the lower troposphere: ultraviolet (UV)- and visible (Vis) absorption spectra of skylight are analysed to obtain information on different atmospheric absorbers and scatterers, integrated over the light path (in fact a superposition of a multitude of light paths). The amount of atmospheric trace gases along the light path is inferred by identifying and analysing their characteristic narrow spectral absorption features, applying differential optical absorption spectroscopy (DOAS, Platt and Stutz, 2008). Gases that have been analysed in the UV and visible spectral range are nitrogen dioxide (NO2), formaldehyde (HCHO) nitrogen dioxide (NO_2), formaldehyde (HCHO), nitrous acid (HONO), water vapour (H_2O), sulfur dioxide (SO_2), ozone (O_3), glyoxal (CHOCHO) and halogen oxides (e.g. BrO, OClO). The oxygen collision induced absorption (in the following treated as if being an additional trace gas species O_4) can be used to infer information on aerosols: since the concentration of O_4 is proportional to the square of the O_2 concentration, its vertical distribution is well known. The O_4 absorption signal can therefore be utilized as a proxy for the light path with the latter being strongly dependent on the atmosphere's aerosol content. An appropriate set of spectra recorded under a narrow field of view (FOV, full aperture angle around $10\,\mathrm{mrad}$) and different viewing elevations ("Multi-Axis") provides information on the trace gas and aerosol vertical distributions. Profiles can be retrieved from this information by applying numerical inversion algorithms, typically incorporating radiative transfer models. These profile retrieval algorithms are the subject of this comparison study.

Today, there are numerous retrieval algorithms in regular use within the MAX-DOAS community which rely on different mathematical inversion approaches. This study involves nine of these algorithms (listed in Table 2), of which six use the optimal estimation method (OEM), two use a parametrized approach (PAR) and one algorithm relies on simplified radiative transport

assumptions and analytical calculations (ANA). The main objective of this study is to assess their consistency and to review strengths and weaknesses of the individual algorithms and techniques. Note that this study is strongly linked to the report by Frieß et al. (2019), who performed similar investigations on nearly the same set of profiling algorithms with synthetic data, whereas the underlying data here was recorded during the second "Cabauw Intercomparison for Nitrogen Dioxide measuring Instruments" (CINDI-2, Apituley et al., 2020 in prep.). The CINDI-2 campaign took place from 25 August to 7 October 2016 on the Cabauw Experimental Site for Atmospheric Research (CESAR, 51.9676° N, 4.9295° E) in the Netherlands, which is operated by the Royal Netherlands Meteorological Institute (KNMI). 36 spectrometers of 24 participating groups from all over the world were synchronously measuring together with a wide range of supporting observations (in situ analysers, balloon sondes, lidars, Long-Path DOAS, direct-sun DOAS, sun photometer and meteorological instruments) for validation. This study compares MAX-DOAS profiles of NO_2 and HCHO concentrations as well as the aerosol extinction coefficient (derived from O_4 observations) from 15 of the 24 groups. The results are compared with each other and validated against CINDI-2 supporting observations. For HONO and O_3 profiling results please refer to Wang et al. (2020) and Wang et al. (2018), respectively. In a recent publication by Bösch et al. (2018), CINDI-2 MAX-DOAS profiles retrieved with the BOREAS algorithm were already compared against supporting observations but regarding a few days only. Finally it shall be mentioned that already in the course of the precedent CINDI-1 campaign in 2009, there were comparisons of MAX-DOAS aerosol extinction coefficient profiles e.g. by Frieß et al. (2016) and Zieger et al. (2011), however also over shorter periods and a smaller group of participants.

The paper is organized as follows: Sect. 2 introduces the campaign setup, the MAX-DOAS dataset with the participating groups and algorithms (Sect. 2.1), the available supporting observations for validation (Sect. 2.2) and the general comparison strategy (Sect. 2.3). The comparison results are shown in Sect. 3. A compact summarizing plot and the conclusions appear in Sect. 4.

2 Instrumentation and methodology

Figure 1 shows an overview of the CINDI-2 campaign setup, including the supporting observations relevant for this study. Instrument locations, pointing (remote sensing instruments) and flight paths (radiosondes) are indicated on the map. Details on the instruments and their data products can be found in the following subsections. For further information refer to Kreher et al. (2019) and Apituley et al. (2020 in prep.).

2.1 MAX-DOAS dataset

2.1.1 Underlying dSCD dataset

Deriving vertical gas concentration/aerosol extinction profiles from scattered skylight spectra can be regarded as a two-step process: the 1st step is the DOAS spectral analysis, where the magnitude of characteristic absorption patterns of different gas species in the recorded spectra is quantified to derive the so called "differential slant column densities" (dSCDs, definition in the following paragraph). These provide information on integrated gas concentrations along the lines of sight. The 2nd step is

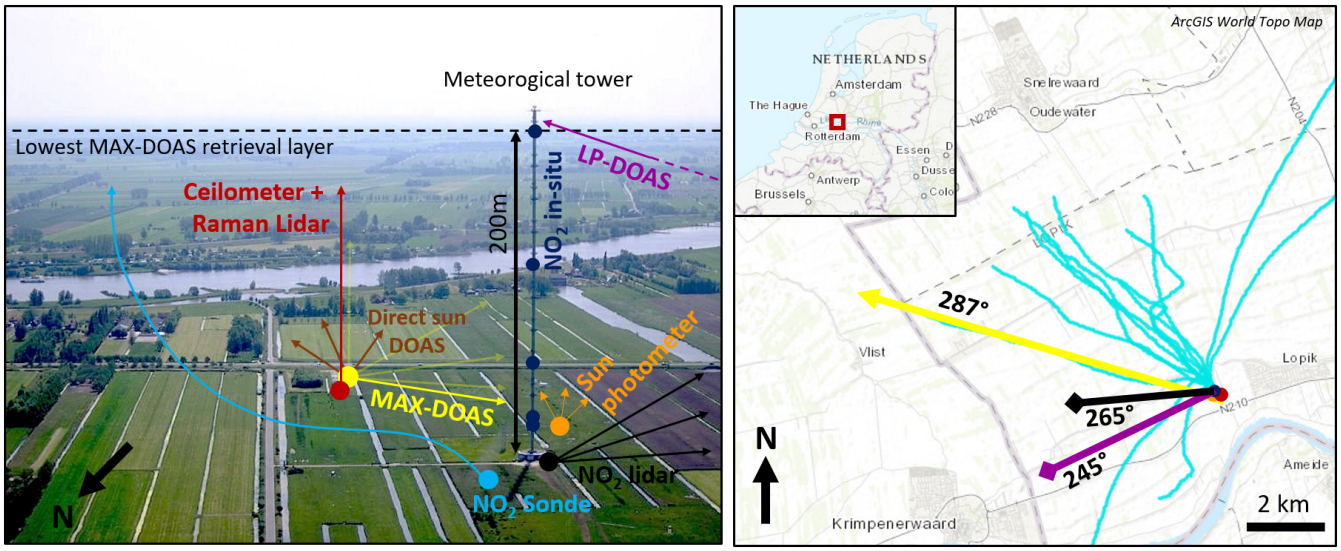

Figure 1. Left: Image of the CESAR site with position and approximate viewing directions of the MAX-DOAS instruments and supporting observations of relevance for this study. Right: Map (Esri et al., 2018) with instrument locations, viewing geometries and sonde flight paths indicated.

the actual profile retrieval, where inversion algorithms incorporating atmospheric radiative transfer models (RTM) are applied to retrieve concentration profiles from the dSCDs derived in the 1st step.

The very initial data in the MAX-DOAS processing chain are intensities of scattered skylight $I_\lambda(\alpha)$ at different wavelengths λ (ultra violet and visible spectral range, typical resolutions of 0.5 to $1.5\,\mathrm{nm}$) recorded under different viewing elevation angles α (ideally the telescope's FOV is negligible compared to the elevation angle resolution). Along the light path l from the top of the atmosphere (TOA) to the instrument on the ground, each atmospheric gas species i imprints its unique spectral absorption pattern (given by the absorption cross section $\sigma_{i,\lambda}$) onto the TOA spectrum $I_{\lambda,TOA}$ with the optical thickness

$$\tau_\lambda(\alpha) = \log\left(\frac{I_{\lambda,TOA}}{I_\lambda(\alpha)}\right) = \sum_i \sigma_{i,\lambda}\, S_i(\alpha) + C \tag{1}$$

$S_i(\alpha)$ is the slant column density (SCD), which is the trace gas concentration integrated along l. C represents terms accounting for other instrumental and physical effects than trace gas absorption (for instance scattering on molecules and aerosols) that will not be further discussed in this context. $S_i(\alpha)$ is inferred by spectrally fitting literature values of $\sigma_{i,\lambda}$ to the observed $\tau_\lambda(\alpha)$. Since normally $I_{\lambda,TOA}$ is not available for the respective instrument, optical thicknesses are instead assessed with respect to the spectrum recorded in zenith viewing direction to obtain

$$\Delta\tau_\lambda(\alpha) = \log\left(\frac{I_\lambda(\alpha = 90°)}{I_\lambda(\alpha)}\right) \tag{2}$$

Then the spectral fit yields the so called differential slant column densities (dSCDs)

$$\Delta S(\alpha) = S(\alpha) - S(\alpha = 90°) \tag{3}$$

which are the typical output of the DOAS spectral analysis when applied to MAX-DOAS data. For further details on the DOAS method refer to Platt and Stutz (2008).

During the CINDI-2 campaign, each participant measured spectra with an own instrument and derived dSCDs applying their preferred DOAS spectral analysis software. The pointings (azimuthal and elevation) of all MAX-DOAS instruments were aligned to a common direction (Donner et al., 2019) and all participants had to comply with a strict measurement protocol, assuring synchronous pointing and spectra acquisition under highly comparable conditions (Apituley et al., 2020 in prep.). A detailed comparison and validation of the dSCD results was conducted by Kreher et al. (2019). In the course of their study, Kreher et al. identified the most reliable instruments to derive a "best" median dSCD dataset. This dataset - in the following referred to as the "median dSCDs" - was distributed among the participants. All participants used the median dSCDs as the input data for their retrieval algorithms and retrieved the profiles that are compared in this study. The "median dSCD" approach was chosen for the following reasons: i) it enables to compare the profiling algorithms independently from differences in the input dSCDs, which is necessary to assess the individual algorithm performances. ii) it makes this study directly comparable to the report by Frieß et al. (2019). Among others, this allows to assess to what extent MAX-DOAS profiling studies on synthetic data (with lower effort) can be used to substitute studies on real data. iii) Two decoupled studies are obtained (Kreher et al. and this study), each confined to a single step in the MAX-DOAS processing chain (the DOAS spectral analysis to obtain dSCDs and the actual profile inversion). A disadvantage of the median dSCD approach is, that the reliability of a typical MAX-DOAS observation undergoing the whole spectra acquisition and processing chain cannot be assessed. Therefore, a comparison of profiles retrieved with the participant's own dSCDs was also conducted, but is not a substantial part of this study. However, these results and a corresponding short discussion can be found in Supplement S10 and Sect. 3.7, respectively. The median dSCDs cover the campaign core period from 12 to 28 September 2016, considering only data from the first 10 minutes of each hour between 7:00 and 16:00 UT, where the CINDI-2 MAX-DOAS measurement protocol scheduled an elevation scan in the nominal 287° azimuth viewing direction with respect to the north. Hence, the total number of processed elevation scans was 170. An elevation scan consisted of ten successively recorded spectra at viewing elevation angles α of 1, 2, 3, 4, 5, 6, 8, 15, 30 and 90°, at an acquisition time of 1 minute each. DSCDs were provided for three chemical species, namely O_4, NO_2 and HCHO. O_4 and NO_2 were each provided for two different spectral fitting ranges, in the ultra-violet (UV) and the visible (Vis) spectral region, resulting in five data products (see Table 1). From the median dSCDs, the participants retrieved profiles for the species listed in Table 1. Not all participants retrieved all species and therefore do not necessarily appear in all plots.

2.1.2 Participating groups and algorithms

Table 2 lists the compared algorithms including the underlying method (OEM, PAR or ANA) and the participating groups with corresponding labels and plotting symbols as they are used throughout the comparison. OEM and PAR algorithms rely on the same idea: a layered horizontally homogeneous atmosphere is set up in a radiative transfer model (RTM) with distinct parameters (aerosol extinction coefficient, trace gas amounts, temperature, pressure, water vapour and aerosol properties) attributed to each layer. This model atmosphere is then used to simulate MAX-DOAS dSCDs under consideration of the

Table 1. List of the retrieved species and fitting ranges. For further details on the spectral analysis, please refer to Kreher et al. (2019).

Species	Retrieved quantity	Retrieved from dSCDs of	spectral fitting window [nm]
Aerosol UV	Extinction coefficient [km^{-1}]	O_4 UV	338 - 370
Aerosol Vis	Extinction coefficient [km^{-1}]	O_4 Vis	425 - 490
NO_2 UV	Number concentration [$molec\,cm^{-3}$]	NO_2 UV	338 - 370
NO_2 Vis	Number concentration [$molec\,cm^{-3}$]	NO_2 Vis	425 - 490
HCHO	Number concentration [$molec\,cm^{-3}$]	HCHO	336.5 - 359

viewing geometries. To retrieve a profile from the measured dSCDs, the model parameters are optimized to minimise the difference between the simulated and measured dSCDs based on a pre-defined cost function.

Regarding profiles, typically only two to four degrees of freedom for signal (DOFS or p) can be retrieved from MAX-DOAS observations, such that general profile retrieval problems with more than p independent retrieved parameters are ill-posed and prior information has to be assimilated to achieve convergence. For OEM algorithms, this is provided in the form of an *a priori* profile and associated *a priori* covariance (Rodgers, 2000), defining the most likely profile and constraining the space of possible solutions according to prior experience. They constitute a portion of the OEM cost function such that with decreasing information contained in the measurements, layer concentrations are drawn towards their *a priori* values. PAR algorithms implement prior assumptions by only allowing predefined profile shapes which can be described by a few parameters.

For OEM algorithms, the radiative transport simulations are performed online in the course of the retrieval whereas the PAR algorithms in this study rely on look-up tables, which are pre-calculated for the parameter ranges of interest. Therefore, PAR algorithms are typically faster than OEM algorithms but also require more memory. The ANA approach by NASA was developed as a quick look algorithm and assumes a simplified radiative transport, based on trigonometric considerations. Since the model equations can be solved analytically for the parameters of interest, neither radiative transport simulation nor the calculation of look-up tables is necessary and an outstanding computational performance is achieved compared to other algorithms (factor of $\approx 10^3$ in processing time, see Frieß et al., 2019).

For further descriptions of the methods and the individual algorithms, please refer to Frieß et al. (2019). Besides the algorithms described therein, our study includes results from the M^3 algorithm by LMU. Its description can be found in Supplement S1. For details, refer to the references given in Table 2.

Note that two versions of aerosol results from the MAPA algorithm with different O_4 scaling factors (SF) are discussed within this paper, referred to as mp-0.8 (retrieved with $SF = 0.8$) and mp-1.0 ($SF = 1.0$), respectively. The scaling factor is applied to the measured O_4 dSCDs prior to the retrieval and was initially motivated by previous MAX-DOAS studies which reported a significant yet debated mismatch between measured and simulated dSCDs (e. g. Wagner et al., 2009; Clémer et al., 2010; Ortega et al., 2016; Wagner et al., 2019, and references therein). Also for MAPA during CINDI-2, a scaling factor of 0.8 was found to improve the dSCD agreement, to enhance the number of valid profiles and to significantly improve the agreement with the sun photometer aerosol optical thickness (Beirle et al., 2019). However, in the course of this study it was found that for

OEM algorithms the disagreement between sun-photometer and MAX-DOAS can largely be explained by smoothing effects (see Sect. 3.4) and that (at least averaged over campaign) there are no clear indications that a SF is necessary (see Supplement S2).

2.1.3 Retrieval settings

To reduce possible sources of discrepancies, all profiles shown in this study were retrieved according to predefined settings similar to those of the intercomparison study by Frieß et al. (2019): pressure, temperature, total air density, and O_3 vertical profiles between 0 and $90\,\mathrm{km}$ altitude were averaged from O_3 sonde measurements performed in De Bilt by KNMI during September months of the years 2013-2015. A fixed altitude grid was used for the inversion, consisting of 20 layers between 0 and $4\,\mathrm{km}$ altitude, each with a height of $\Delta h = 200\,\mathrm{m}$. The results of the parametrized approaches and OEM algorithms where the exact grid could not readily be applied during inversion, were interpolated/averaged accordingly afterwards. Note that, for radiative transfer simulations, the atmosphere was represented by finer (25 to $100\,\mathrm{m}$) layers close to the surface, increasing with altitude) and farther extending (up to 40 to $90\,\mathrm{km}$ altitude) grids, inherently defined by the individual retrieval algorithms. Surface and instruments' altitudes were fixed to $0\,\mathrm{m}$, which is close to the real conditions: the CESAR site and most of the surrounding area lie at 0.7 metres b.s.l., whereas the instruments were installed at 0 to $6\,\mathrm{m}$ above sea level. The model wavelengths were fixed according to Table 3. In the case of the HCHO retrieval, the aerosol profiles retrieved at $360\,\mathrm{nm}$ were extrapolated to $343\,\mathrm{nm}$ using the mean Ångström exponent for the $440 - 675\,\mathrm{nm}$ wavelength range derived from sun photometer measurements (see Sect. 2.2.1) on 14 September 2016 in Cabauw. For the aerosol parameters, the single scattering albedo was fixed to 0.92 and the asymmetry factor to 0.68 for both 360 and 477 nm. These are mean values for 14/09/2016 derived from AERONET measurements at $440\,\mathrm{nm}$ in Cabauw. The standard CINDI-2 trace gas absorption cross-sections were applied (see Kreher et al., 2019). A scaling of the measured O_4 dSCDs prior to the retrieval was not applied. An exception is the parametrized MAPA algorithm for which two datasets, one without and one with a scaling ($SF = 0.8$) were included in this study. The OEM *a priori* profiles for both aerosol and trace gas retrievals were exponentially-decreasing profiles with a scale height of 1 km and aerosol optical thicknesses (AOTs) and vertical column densities (VCDs) as given in Table 3. For the AOTs the mean value at 477 nm for the first days of September 2016 derived from AERONET measurements are used. Trace gas VCDs are mean values derived from OMI observations in September 2006-2015. *A priori* variance and correlation length were set to $50\,\%$ and $200\,\mathrm{m}$, respectively.

2.1.4 Requested dataset

All participants were requested to submit the following results of their retrieval: i) Profiles and profile errors, optionally with errors separated into contributions from propagated measurement noise and smoothing effects. ii) Modelled dSCDs as calculated by the RTM for the retrieved atmospheric state. iii) Averaging Kernels (AVKs) for assessment of information content and vertical resolution (only available for OEM approaches). iv) Optional flags, giving participants the opportunity to mark profiles as invalid. The flagging must be based on inherent quality indicators, which typically are the root-mean-square difference between measured and modelled dSCDs or the general plausibility of the retrieved profiles. Note that only four institutes

submitted flags (INTA/ bePRO, BIRA/ bePRO, KNMI/ MARK and MPIC/ MAPA). It is assumed that an accurate aerosol retrieval is necessary to infer light path geometries, thus trace gas profiles are generally considered invalid if the underlying aerosol retrieval is invalid. A detailed description of the flagging criteria and flagging statistics can be found in Supplement S3.

2.2 Supporting observations

This section introduces the supporting observations, that were used for comparison and validation of the MAX-DOAS retrieved results. It shall be pointed out that a general challenge here was to find compromises between i) using only accurate and representative data with good spatio-temporal overlap and ii) keeping as many supporting data as possible to have a large comparison dataset. Considerations and investigations on this issue (e.g. comparisons between the supporting observations, spatio-temporal variability and overlap) which lead to the decisions finally taken are mentioned in the following subsections and described in more detail in the supplementary material they refer to.

2.2.1 Aerosol optical thickness (AOT)

Independent aerosol optical thickness measurements τ_{aer} were performed with a sun photometer (CE318-T by Cimel) located close to the meteorological tower of the CESAR site (see Fig. 1), which is part of the Aerosol Robotic Network (AERONET, see Holben et al., 1998). AOTs were derived from direct-sun radiometric measurements in $\approx$ 15 minute intervals at 1020, 870, 675 and 440 nm wavelength. The AERONET level 2.0 data was used, which is cloud screened, recalibrated and quality filtered (according to Smirnov et al., 2000). For the extrapolation of τ_{aer} to the DOAS retrieval wavelengths of 360 and 477 nm, a dependency of τ_{aer} on the wavelength λ according to

$$\ln \tau_s(\lambda) = \alpha_0 \, + \, \alpha_1 \cdot \ln \lambda \, + \, \alpha_2 \cdot (\ln \lambda)^2 \tag{4}$$

was assumed, following Kaskaoutis and Kambezidis (2006). The parameters α_i were retrieved by fitting Eq. (4) to the available data points. Note that α_1 corresponds to the Ångström exponent when only the first two (linear) terms on the right hand side are used. The last quadratic term enables to additionally account for a change of the Ångström exponent with wavelength. For the linear temporal interpolation to the MAX-DOAS profile timestamps, the maximum interpolated data gap was set to 30 min, resulting in a data coverage of about $30\,\%$. Smirnov et al. (2000) propose a sun photometer total accuracy in τ_s of 0.02. Each AOT is actually an average over three subsequently performed measurements. In this study, the proposed accuracy of 0.02 was enhanced by the variability between them (typically on the order of 0.008).

2.2.2 Aerosol profiles

Information on the aerosol extinction coefficient profiles (in the following referred to by "aerosol profiles") was obtained by combining the sun photometer AOT with data from a ceilometer (Lufft CHM15k Nimbus). The latter continuously provided vertically resolved information on the atmospheric aerosol content by measuring the intensity of elastically backscattered light from a pulsed laser beam (1064 nm) propagating in zenith direction (see e.g. Wiegner and Geiß, 2012). The raw data are attenuated backscatter coefficient profiles over an altitude range from $180\,\mathrm{m}$ to $15\,\mathrm{km}$, with a temporal and vertical resolution

of 12 s and 10 m, respectively. These were converted to extinction coefficient profiles by scaling with simultaneously measured sun photometer or MAX-DOAS AOTs. This is described in detail in Supplement S4.1. Note that the approach described there presumes a constant extinction coefficient for altitudes $\leq 180\,\mathrm{m}$ and that the aerosol properties like size distribution, single scattering albedo and shape remain constant with altitude. To check plausibility, Supplement S4.1 compares the resulting profiles at 360 nm to a few available extinction coefficient profiles, measured by a Raman lidar at 355 nm (the CESAR Water Vapor, Aerosol and Cloud lidar "CAELI", operated within the European Aerosol Research lidar Network (EARLINET, Bösenberg et al., 2003; Pappalardo et al., 2014) and described in detail in Apituley et al., 2009). The average RMSD between scaled ceilometer and Raman lidar profiles up to $4\,\mathrm{km}$ altitude is $\approx 0.03\,\mathrm{km}^{-1}$. However since there are only few Raman lidar validation profiles available and only for altitudes > 1 km, the ceilometer aerosol profiles should be consulted for qualitative comparison only.

2.2.3 NO_2 profiles

NO_2 profiles were recorded sporadically by two measurement systems: radiosondes (described in Sluis et al., 2010) and an NO_2 lidar (Berkhout et al., 2006). Radiosondes were launched at the CESAR measurement site during the campaign. For this study, only data from sonde ascents through the lowest 4 km (which is the MAX-DOAS profiling retrieval altitude range) were used. A sonde profile was considered temporally coincident to a MAX-DOAS profile, when the middle timestamps of MAX-DOAS elevation scan and sonde flight were less than 30 minutes apart. The horizontal sonde flight paths are indicated in Fig. 1. Typical flight times (lowest 4 km) were of the order of 10 - 15 minutes. Data was recorded at a rate of 1 Hz, typically resulting in a vertical resolution of approximately 10 m at an approximate measurement uncertainty in NO_2 concentration of $5 \times 10^{10}\,\mathrm{molec\,cm}^{-3}$. The horizontal travel distances varied strongly between 4 and 18 km. A detailed overview on the flights is given in Supplement S4.2.

The NO_2 lidar is a mobile instrument setup inside a lorry which was located close to the CESAR meteorological tower. It combines lidar observations at different viewing elevation angles to enhance vertical resolution and to obtain sensitivity close to the ground, despite the limited range of overlap between sending and receiving telescope (see also Sect. 2.2.2). The instrument is sensitive along its line of sight from 300 to 2500 m distance to the instrument. The azimuthal pointing was 265° with respect to the north and the operational wavelength is 413.5 nm. Typical specified uncertainties in the retrieved concentrations are around $2.5 \times 10^{10}\,\mathrm{molec\,cm}^{-3}$. Profiles were provided at a temporal resolution of 28 minutes, each profile consisting of a series of (occasionally overlapping) altitude intervals with constant gas concentration. For an exemplary profile and details on its conversion to the MAX-DOAS retrieval altitude grid, please refer to Supplement S4.3. A lidar profile was considered temporally coincident to a MAX-DOAS profile, when the middle timestamps of MAX-DOAS elevation scan and lidar profile were less than 30 minutes apart. This resulted into 25 suitable Lidar profiles recorded on six different days during the campaign. Example profiles of both radiosonde and NO_2 lidar are shown in the course of a comparison between the two observations in Supplement S4.5.

2.2.4 Trace gas vertical column densities (VCD)

Tropospheric trace gas VCDs were derived from direct-sun DOAS observations, which were performed between minutes 40 and 45 of each hour. NO_2 VCDs were retrieved from combined datasets of two Pandora DOAS instruments (instrument numbers 31 & 32) and calculated based on the Spinei et al. (2014) approach. The reference spectrum was created from the spectra with lowest radiometric error over the whole campaign and the residual NO_2 signal was determined by applying the so-called Minimum Langley Extrapolation (Herman et al., 2009). The temperature dependence of the NO_2 cross sections was used to separate the tropospheric from the stratospheric column.

HCHO VCDs were retrieved from data of the BIRA DOAS instrument (number 4). A fixed reference spectrum acquired on 18 September 2016 at 9:41 UTC and 55.6° SZA was used. DOAS fitting settings were identical to those used for the CINDI-2 HCHO dSCD intercomparison (Kreher et al., 2019). The residual amount of HCHO in the reference spectrum of $(8.8 \pm 1.6) \times 10^{15}\,\mathrm{molec\,cm^{-2}}$ was estimated using a MAX-DOAS profile retrieved on the same day and a geometrical AMF corresponding to 55.6° SZA. Because of that, the HCHO VCDs cannot be considered as a fully independent dataset. VCDs were calculated from total HCHO SCDs using a geometrical AMF including a simple correction for the earth sphericity. Only spectra with DOAS fit residuals $< 5 \times 10^{-4}$ were considered as valid direct-sun data. As for AOTs, these observations can only be performed when the sun is clearly visible, hence the coverage for cloudy scenarios is scarce.

2.2.5 Trace gas surface concentrations

Note that in the following, "surface concentration" will not refer to measurements in the very proximity to the ground but to the average concentration in the lowest 200 m of the atmosphere, as retrieved for the MAX-DOAS first profile layer. Trace gas surface concentrations of HCHO and NO_2 were provided by a long path DOAS system operated by IUP-Heidelberg (LP-DOAS, see Pikelnaya et al., 2007; Pöhler et al., 2010; Merten et al., 2011; Nasse et al., 2019). The LP-DOAS system consists of a light-sending and receiving telescope unit located at 3.8 km horizontal distance to a retro reflecting mirror mounted at the top (207 m altitude) of the meteorological tower (see Supplement S4.4). Light from a UV-Vis light source is sent by the telescope to the retroreflector and the reflected light is again received by the telescope unit and spectrally analysed applying the DOAS method. The fundamental difference to the MAX-DOAS instruments is the well-defined light path which enables very accurate determination of trace gas mixing ratios, averaged along the line of sight. Accordingly, with the retroreflector mounted at 207 m altitude, one obtains average mixing ratios over the lowest MAX-DOAS retrieval layer, as indicated in Fig. 1. Considering DOAS fitting errors and uncertainties in the applied literature cross-sections (Vandaele et al., 1998; Meller and Moortgat, 2000; Pinardi et al., 2013) yields an average accuracy of the LP-DOAS of $\pm 1.5 \times 10^9\,\mathrm{molec\,cm^{-3}} \pm 3\,\%$ ($\pm 5 \times 10^9\,\mathrm{molec\,cm^{-3}} \pm 9\,\%$) for NO_2 (HCHO), respectively. Given the high accuracy, the total vertical coverage of the surface layer and a near-continuous dataset over the campaign period, the LP-DOAS provides the most reliable dataset for the validation of CINDI-2 MAX-DOAS trace gas profiling results.

Further observations for qualitative validation are the surface values of the NO_2 lidar and the radiosondes and also in-situ monitors in the CESAR meteorological tower. Teledyne in situ NO_2 monitors (Teledyne API, model M200E) were located in

the tower basement and were subsequently connected to different inlets located at 20, 60, 120 and 200 m altitude (switching intervals approx. 5 minutes). Further, a CAPS (type AS32M, based on attenuated phase shift spectroscopy, Kebabian et al., 2005) and a CE-DOAS (cavity enhanced DOAS, Platt et al., 2009 and Horbanski et al., 2019) were continuously measuring at 27 m altitude. All the in situ measurements at the tower were combined to obtain another set of surface concentration measurements, more representative for concentrations close to the site. The data were combined by linearly interpolating over altitude between the instruments and subsequently averaging the resulting profile over the retrieval surface layer (0 - 200 m altitude). Note that this method gives a large weight to the uppermost measurements, as they are representative for the majority of the relevant layer.

2.2.6 Meteorology

Meteorological data for the surface layer (pressure, temperature and wind information) routinely measured at the CESAR site were taken from the CESAR database (CESAR, 2018) at a temporal resolution of 10 minutes. Cloud conditions were retrieved from MAX-DOAS data of instruments 4 and 28 according to the cloud classification algorithm developed by MPIC (Wagner et al., 2014; Wang et al., 2015). Basically only two cloud condition states are distinguished in the statistical evaluation: "clear-sky" (green) and "presence of clouds" (red). Only in the overview- and correlation plots, “presence of clouds” is further subdivided into "optically thin clouds" (orange) and "optically thick clouds" (red). According to this classification 72 (98) of the 170 profiles were measured under clear-sky (cloudy) conditions. Over the whole campaign, there was only one rain event (precipitation > 0.01 mm) coinciding with the measurements on 25 September 2016 between 15:00 and 17:00 h UT. At forenoon on 16 September, a heavy fog event strongly limited the visibility (see also Supplement S5).

2.3 Comparison strategy

2.3.1 General approach

Different MAX-DOAS retrieval algorithms were extensively compared in Frieß et al. (2019) using synthetic data. The crucial differences of the presented study are: i) The underlying spectra are not synthetic, but were recorded with real instruments, meaning that real noise and instrument artefacts propagate into the results. ii) Independent information on the real profile can only be inferred from supporting observations with their own uncertainties and an imperfect spatio-temporal overlap with the MAX-DOAS measurements. iii) The real conditions encountered can exceed the model’s scope because horizontal inhomogeneities or the fact that many of the fixed forward model input parameters (such as aerosol properties, surface albedo, temperature and pressure profiles) are averaged quantities of former observations which might be inaccurate for specific days and conditions. iv) In some cases, different participants used the same retrieval algorithms; this allows assessment of the impact of different settings in the remaining parameters, which were not prescribed (see Sect. 2.1.3). The approaches chosen here are therefore limited to the examination of i) the consistency among the participants, ii) the consistency of the results with available supporting observations and iii) inherent quality proxies of the retrieval (described in the next paragraph). Table 4 summarizes the quantities which are compared, together with the corresponding supporting observations if available.

In this study, agreement between different observations are statistically assessed by i) weighted root-mean-square differences (RMSD), ii) weighted "Bias" as introduced below and iii) weighted least-squares regression analysis. Discussions and summary are focussed on RMSD, being the most fundamental quantity as it represents both, statistical and systematic deviations. The Bias was introduced as a general proxy for systematic deviations. Correlation coefficient, slope and offset from the regression analysis are provided and consulted for a more differentiated view.

Consider two time series of length N_T: the retrieval result $x_{p,t}$ of a participant p at time t and some reference observation $x_{ref,t}$ (either MAX-DOAS median results or data from supporting observations, as further described below) with associated uncertainties $\sigma_{p,t}$ and $\sigma_{ref,t}$. Then the RMSD is defined as:

RMSD:
$$\sigma_{rms,p} = \sqrt{\frac{1}{N_T} \cdot \frac{1}{\sum_t w_t} \cdot \sum_t w_t \left(x_{p,t} - x_{ref,t}\right)^2} \tag{5}$$

The weights w_t are defined according to

$$w_t = \frac{1}{\sigma_{p,t}^2 + \sigma_{ref,t}^2} \tag{6}$$

and are also applied for the Bias calculation and regression analysis. The Bias is defined as

Bias:
$$\sigma_{bias,p} = \frac{1}{N_T} \cdot \frac{1}{\sum_t w_t} \cdot \sum_t w_t \left(x_{p,t} - x_{ref,t}\right) \tag{7}$$

Sometimes the term "average RMSD" ("average Bias") is used, which refers to the average over the RMSD (Bias) values of the individual participants. We further introduce the "average Bias magnitude", that averages the absolute values of the Bias. When referring to "relative RMSDs" ("relative Bias"), the underlying RMSD (Bias) value was divided by the average of the investigated quantity. For the linear regression analysis, the vertical distance between the model and the data points is minimised and also here the weights w_t are applied.

To assess the consistency among the participants, the median result over the valid profiles of all participants is inserted as $x_{ref,t}$. The median is used instead of the mean value, since it is less sensitive to (sometimes unphysical) outliers. This comparison shows how far the choice of the retrieval algorithm/ technique affects the results but it does not reveal general systematic MAX-DOAS retrieval errors. Outliers observed for distinct participants and algorithms are therefore not necessarily an indicator for poor performance.

To assess the consistency with supporting observations, the latter are inserted as $x_{ref,t}$. This comparison is a better indicator for the real retrieval performance. However, uncertainties of supporting instruments (see Supplement S4.5), smoothing effects (see Sect. 2.3.2) and imperfect spatial and temporal overlap of the different observations (see Sect. 2.3.3) complicate the interpretation.

An inherent quality indicator for the retrieval algorithms are the consistency of modelled and measured dSCDs. During the inversion, the goal is to minimize the deviation between the RTM simulated dSCDs and the actually measured ones. If strong deviations remain after the final iteration in the minimisation process, this indicates failure of the retrieval.

In a few cases (e.g. Sect. 3.2, where full profiles are compared) the scatter among several participants p (of number N_P) and several retrieval layers h (of number N_H) is of interest. For this purpose, we define the "average standard deviation" (ASDev)

which is the standard deviation observed among the participants for individual profiles averaged over retrieval layers and time, hence:

ASDev:
$$\sigma_{asdev} = \frac{1}{N_T}\sum_t \frac{1}{N_H}\sum_h \sqrt{\frac{1}{N_P - 1}\sum_p \left(x_{p,h,t} - \bar{x}_{h,t}\right)^2} \tag{8}$$

with $\bar{x}_{h,t}$ being the average (over participants) MAX-DOAS retrieved concentration for a given time t and layer h. If not stated otherwise, ASDev values of profiles are calculated considering the lowest five retrieval layers (up to $1\,\mathrm{km}$ altitude).

In the statistical evaluations, clear-sky and cloudy conditions as well as unfiltered and filtered data (according to the flags providede by the participants) are distinguished. The distinction between cloud conditions is of major importance, as particularly in the case of aerosol retrievals under broken clouds, the quality of the results is typically strongly degraded. A consequence of regarding these data subsets is that the number of contributing data points not only depends on the number of submitted profiles and the number of coincident data points from supporting observations but further on the filter settings. Any regression RMSD or Bias value with less than five contributing data points are considered to be statistically unrepresentative and are omitted. If not stated otherwise, numbers given in the text were calculated considering valid data only.

2.3.2 Smoothing effects

As shown in Sect. 3.1 below, in particular in the UV range, the sensitivity of ground-based MAX-DOAS observations decreases rapidly with altitude, meaning that species above $\approx 2\,\mathrm{km}$ typically cannot be reliably quantified. At higher altitudes, OEM retrieval results are drawn towards the *a priori* profile (according to the definition of the cost-function, see Rodgers (2000)), while the results of parametrized and analytical approaches are driven by the chosen parametrization and their implementation. Further, the vertical resolution is limited (from 100 to several hundred meters, increasing with altitude), which affects the profile shape and - of most importance in this study - the retrieved surface concentration. Both effects cause deviations from the true profile that are in the following referred to as "smoothing effects".

For a meaningful quantitative comparison, they should be considered. This is possible for OEM retrievals, where the information on the vertical resolution and sensitivity is given by the averaging kernel matrix (AVK, see Sect. 3.1 for details). For a meaningful quantitative comparison of an OEM retrieved profile and a validation profile $\boldsymbol{x}$ (assumed here to perfectly represent the true state of the atmosphere), the validation profile resolution and information content has to be degraded by "smoothing" it with the corresponding MAX-DOAS AVK matrix $\mathbf{A}$ according to the following equation (Rodgers and Connor, 2003; Rodgers, 2000):

$$\widetilde{\boldsymbol{x}} = \mathbf{A}\boldsymbol{x} + (\mathbf{1} - \mathbf{A})\boldsymbol{x_a} \tag{9}$$

Here, $\boldsymbol{x_a}$ is the *a priori* profile and $\widetilde{\boldsymbol{x}}$ represents the profile that a MAX-DOAS OEM retrieval (with the resolution and sensitivity described by $\mathbf{A}$) would yield in the respective scenario. For layers with high (low) gain in information, $\widetilde{\boldsymbol{x}}$ is drawn towards $\boldsymbol{x}$ ($\boldsymbol{x_a}$), while vertical resolution is degraded if $\mathbf{A}$ has significant off-diagonal entries (compare to Sect. 3.1). In this study, this has implications not only for the comparison of profiles, but also the comparison of the total columns (AOTs and VCDs, which are derived simply by vertical integration of the corresponding profiles) and surface trace gas concentrations.

For total columns, the dominant issue is the lack of information at higher altitudes. In contrast, there is reasonable information on the surface concentration, however smoothing can have severe impact here in the case of strong concentration gradients close to the surface. The impact on the individual observations is discussed in the corresponding sections below. A particularly important consequence of smoothing effects is the "partial AOT correction" (PAC), which is introduced and discussed in Sect. 3.4.

Finally it shall be pointed out that the sensitivity and spatial resolution is strongly affected by the exact approach that is chosen to solve the ill-posed inversion problem. Frieß et al. (2006) for instance demonstrates, that the sensitivity to higher altitudes can be enhanced by relaxing the prior constraints and by retrieving profiles at several wavelengths simultaneously.

2.3.3 Spatio-temporal variability

It is obvious already from Fig. 1 and Sect. 2.2 that the MAX-DOAS instruments and the various supporting observations sample different air volumes at different times. In addition, the MAX-DOAS horizontal viewing distance (derived in Supplement S5) is highly variable, changing between 2 and 30 km during the campaign for the lowest viewing elevation angles. Similar investigations were already performed by Irie et al. (2011) using CINDI-1 data, however using a different definition of the viewing distance. Table S6 summarizes the spatial and temporal mismatches between MAX-DOAS and supporting observations. Spatial mismatches are of the order of $10\,\mathrm{km}$, temporal mismatches vary between 0 and 20 minutes. Consequently, strong spatio-temporal variations of the observed quantities are expected to induce large discrepancies among the observations, independent of the data quality. Quantitative estimates of the impact on the comparison could only be derived for NO_2 surface concentrations and under strong simplifications (for details see S6) yielding an RMSD of $3.5\times10^{10}\,\mathrm{molec\,cm^{-3}}$. This is indeed of similar magnitude as the average RMSD observed during the comparison (approx. $5\times10^{10}\,\mathrm{molec\,cm^{-3}}$). It shall further be noted, that under strong spatial variability the horizontal homogeneity assumed by the retrieval forward models is inaccurate.

3 Comparison results

3.1 Information content

In the case of OEM retrievals, the gain in information on the atmospheric state can be quantified according to Rodgers (2000). Essentially speaking, this is done by comparing the knowledge before (represented by the *a priori* profile and its uncertainties) and after the profile retrieval. The gain in information for each individual vertical profile can be represented by the averaging kernel matrix (AVK, denoted by $\mathbf{A}$). A_{ij} describes the sensitivity of the measured concentration in the i^{th} layer to small changes in the real concentration in the j^{th} layer. Each row A_i can thus be plotted over altitude providing the following information: (1) the value in the layer i itself (the diagonal element A_{ii} with a value between 0 and 1) gives the gain in information while $1-A_{ii}$ represents the amount of *a priori* knowledge which had to be assimilated to obtain a well defined concentration value. (2) The values in the other layers (off-diagonal elements of $\mathbf{A}$) indicate the cross sensitivity of layer i to layer j. Typically, the cross sensitivity decreases with the distance to the layer i. The length of this decay (note that i can be converted to the corresponding

altitude by multiplication with the retrieval layer thickness Δh) is an indicator for the vertical resolution of the retrieval. The trace of $\mathbf{A}$ equals the degrees of freedom of signal (DOFS), hence the total number of independent pieces of information gained from the measurements compared to the *a priori* knowledge. Figure 2 visualizes the average AVK matrices (median over participants and mean over time) for all five species studied in this work. Note that the AVKs do not necessarily represent the real/ total sensitivity and information content of MAX-DOAS observations as they only consider the gain of information with respect to the *a priori* knowledge. Hence, for stricter *a priori* constraints less gain in information will be indicated by the AVKs.

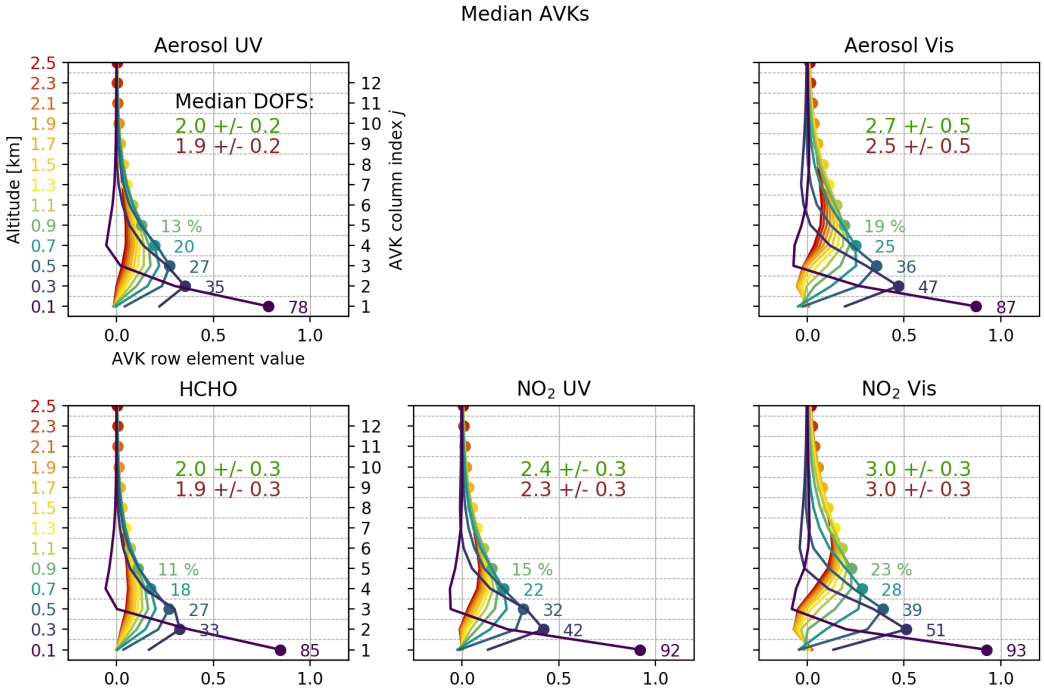

Figure 2. Mean AVKs for the retrieved species (median over participants, mean over time). Their meaning is described in detail in the text. Each altitude and corresponding AVK line A_i are associated with a colour, which is defined by the colour of the corresponding altitude-axis label. The dots mark the AVK diagonal elements. The number next to the dots show the exact value in percent, which corresponds to the amount of retrieved information on the respective layer. In each panel, the numbers indicate the DOFS (median among institutes, average over time) for clear-sky (green) and cloudy conditions (red).

With the *a priori* profiles and covariances used within this study, the sensitivity is limited to about the lowest 1.5 km of the atmosphere for all species. More information is obtained on the Vis species, as the differential light path increases with wavelength resulting in higher sensitivity. The obtained DOFS are generally a bit lower as observed in former studies. This is related to the rather small *a priori* covariance (50 %, see Sect. 2.1.3), which implies a good knowledge on the atmospheric state prior to the retrieval and finally leads to less gain in information from the measurements. Figures S35, S36, S37, S38 and

S39 in Supplement S8.1 show the average AVKs of the individual participants and reveals, that there are significant differences (up to 1 DOFS) between the participants even when using the same algorithm (up to 0.5 DOFS in the case of PRIAM). This indicates that the information content is not assessed consistently. BOREAS for instance states a very low gain in information especially for Aerosol Vis. This is related to an additional Tikhonov term used as a smoother which was also applied during AVK assessment. Furthermore, all BOREAS results were retrieved on another grid and interpolated onto the submission grid, which leads to a decrease in all AVKs and therefore the DOFS. On average, the dependence of the total amount of information on the cloud conditions is small (typically decrease of 0.1 DOFS). Examination of the AVKs of individual profiles (not shown here), indicated that there are two competing effects: (1) the presence of clouds can increase the sensitivity to higher layers due to multiple scattering and thus light path enhancement in the clouds whereas (2) a decrease in the horizontal viewing distance (e.g. due to fog, rain or high aerosol loads) reduces the information content, since the light paths are shorter and their geometry depends less on the viewing elevation.

3.2 Overview plots

Figures 3 to 7 show the retrieved profiles of all participants over the whole semi-blind period. They serve as the basis for a general qualitative comparison. For the trace gases, the altitude ranges (full range is 4 km) were reduced to $0-2.5\,\mathrm{km}$ for better visibility, considering the MAX-DOAS sensitivity range and the occurrence altitude of the respective species.

Table 2. Groups who retrieved and provided profiling results for this study.

Algorithm	Method/ Model	Literature	Participants	Acronym	Sym
bePRO	OEM[o]/ LIDORT	Clémer et al. (2010)	Aristotle University of Thessaloniki, Thessaloniki, Greece	AUTH	
		Hendrick et al. (2014)	Royal Belgian Institute for Space Aeronomy, Brussels, Belgium	BIRA	
			National Institute of Aerospatial Technology, Madrid, Spain	INTA	
PRIAM	OEM[l]/ SCIATRAN	Wang et al.	Anhui Institute Of Optics and Fine Mechanics, Hefei, China	AIOF	
		(2013b, 2013a, 2017)	Belarusian State University, Minsk, Belarus	BSU	
			China Meteorological Administration, Beijing, China	CMA	
			Max-Planck Institute for Chemistry, Mainz, Germany	MPIC	
HEIPRO[x]	OEM[l]/ SCIATRAN	Yilmaz (2012)	Institute of Environmental Physics, University of Heidelberg, Germany	IUPHD	
			Department of Physics, University of Toronto, Canada	UTOR	
BOREAS	OEM[l]/ SCIATRAN	Bösch et al. (2018)	Institute of Environmental Physics, University of Bremen, Germany	IUPB	
M^3	OEM/ LibRadTran	Chan et al.	Ludwig-Maximilians-University, Munich, Germany	LMU	
		(2019, 2017, 2020)			
MMF	OEM[l]/ VLIDORT	Friedrich et al. (2019)	Royal Belgian Institute for Space Aeronomy, Brussels, Belgium	BIRA	
Realtime	ANA[a]/ -		NASA-Goddard, Greenbelt, United States	NASA	
MARK	PAR[p]/ DAK	Vlemmix et al.	Royal Netherlands Meteorological Institute, De Bilt, The Netherlands	KNMI	
		(2011, 2015a)			
MAPA	PAR/ McArtim	Beirle et al. (2019)	Max-Planck Institute for Chemistry, Mainz, Germany	MPIC[y]	
			Max-Planck Institute for Chemistry, Mainz, Germany	MPIC[y]	

[o] OEM: Optimal estimation

[a] ANA: Analytical approach without radiative transfer model

[p] PAR: Parametrized approach

[x] IUPHD and UTOR used different versions of HEIPRO (1.2 and 1.5/1.4, respectively)

[y] Two versions of MAPA (labelled mp-10 and mp08) with different O_4 scaling factors (0.8 and 1.0) are included in the comparison.

[l] Aerosol extinction is retrieved in logarithmic space. This removes negative values and allows larger values.

Table 3. Prescribed settings for the radiative transfer simulation wavelengths and *a priori* total columns (OEM algorithms only).

Species	RTM wavelength [nm]	*A priori* VCD/ AOT
Aerosol UV	360	0.18
Aerosol Vis	477	0.18
NO_2 UV	360	$9 \cdot 10^{15}\,\mathrm{molec\,cm^{-2}}$
NO_2 Vis	460	$9 \cdot 10^{15}\,\mathrm{molec\,cm^{-2}}$
HCHO	343	$8 \cdot 10^{15}\,\mathrm{molec\,cm^{-2}}$

Table 4. Overview on compared quantities and available supporting data.

Species	Quantity	Supporting observations	Result section
Aerosol UV	Profiles	Ceilometer[a] (Sec. 2.2.2)	3.2 & Suppl. S8.2
	Aerosol optical thickness (AOT)	Sun photometer (Sec. 2.2.1)	3.4
Aerosol Vis	Profiles	Ceilometer[a]	3.2 & Suppl. S8.2
	Aerosol optical thickness (AOT)	Sun photometer	3.4
HCHO	Profiles	N.A.	3.2 & Suppl. S8.2
	Vertical column (VCD)	Direct-sun DOAS (Sec. 2.2.4)	3.5
	Surface concentration	Long-path DOAS	3.6
NO_2 UV/Vis	Profiles	NO_2-Lidar & radiosonde[b]	3.2 & Suppl. S8.2
	Vertical column (VCD)	Direct-sun DOAS	3.5
	Surface concentration	Long-path DOAS	3.6
All species	Modelled vs. measured dSCDs	N.A.[c]	3.3

[a] Elastic backscatter profiles scaled with sun photometer or MAX-DOAS AOT.

[b] Scarce data coverage.

[c] Inherent quality proxy.

Figure 3. Aerosol UV extinction profiles. For MAX-DOAS profiles (plots above the wind roses), pink triangles at the top of the corresponding profile indicate invalid data. The lowest row shows AOT scaled ceilometer backscatter profiles, calculated as described in Sect. 2.2.2 (unsmoothed). Backscatter profiles, which were scaled from MAX-DOAS AOTs (and which are therefore not fully independent) are marked by pink triangles. Maximum extinction values reach $20\,\mathrm{km}^{-1}$, exceeding the colour scale. Index letters behind the participant labels indicate whether an OEM (o) or parametrized (p) approach was used and whether aerosol was retrieved in the logarithmic space (l). The wind roses in the lower part of the panel show wind direction (azimuth), wind speed (see colour bar on the right) and occurrences (amplitude). The line close to the panel bottom marked with "CC" indicates the cloud conditions, as described in Sect. 2.2.6.

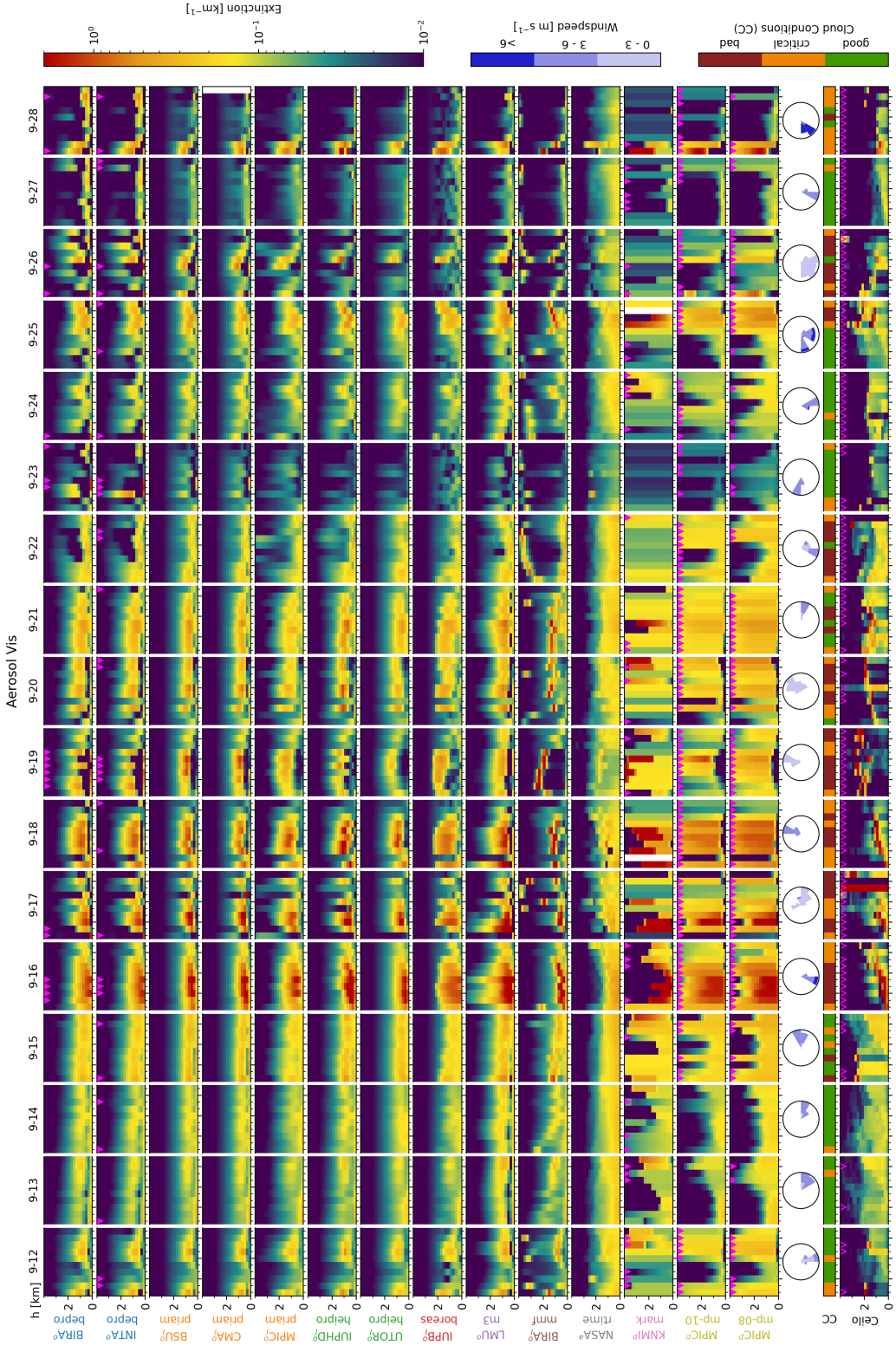

Figure 4. Aerosol Vis extinction profiles. Caption of Fig. 3 applies.

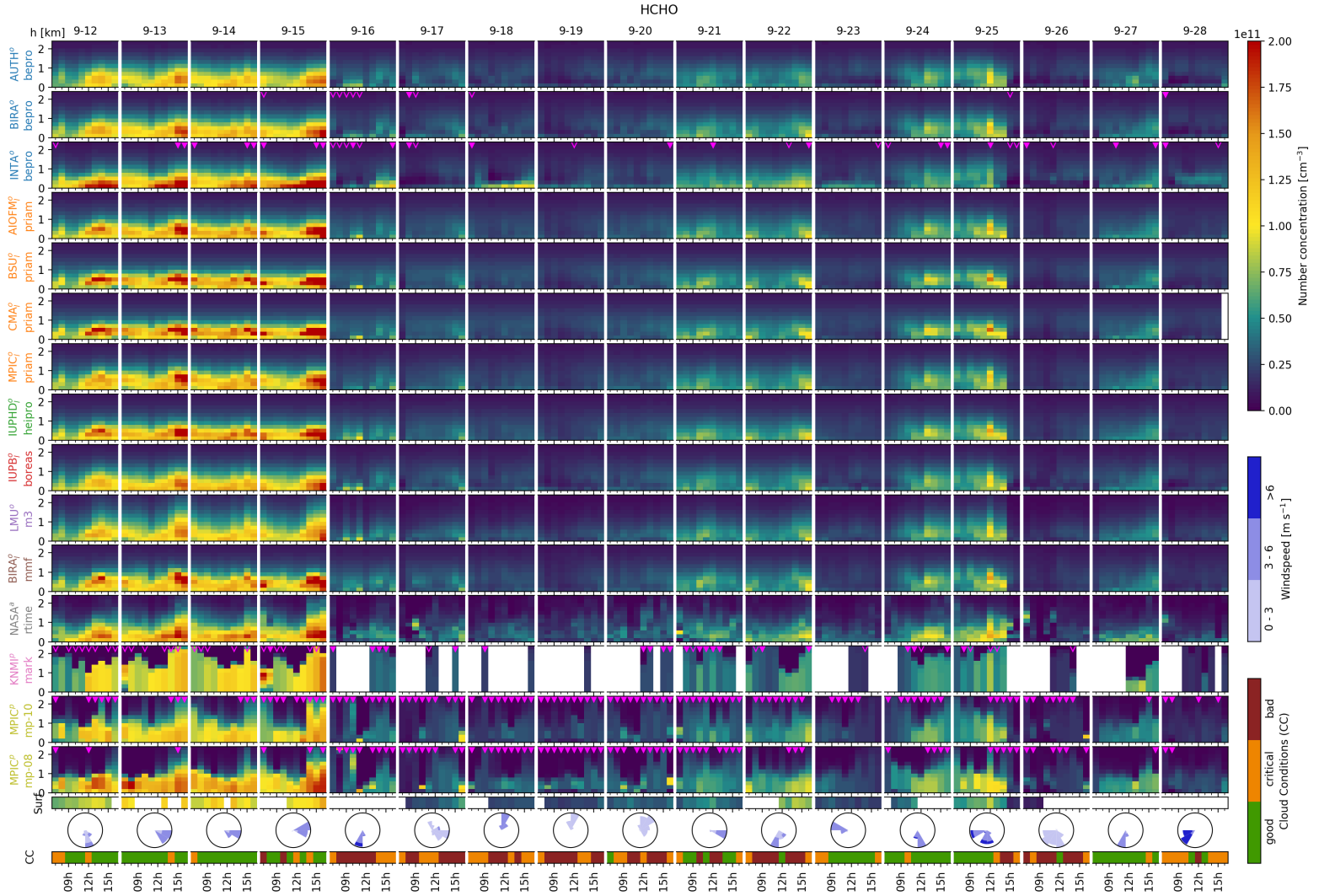

Figure 5. HCHO concentration profiles. The plot is similar to Fig. 3. Open pink triangles at the top of the MAX-DOAS profiles indicate that the underlying aerosol retrieval failed, whereas the trace gas profile retrieval itself was considered successful. The "Surf"-row shows LP-DOAS surface concentrations.

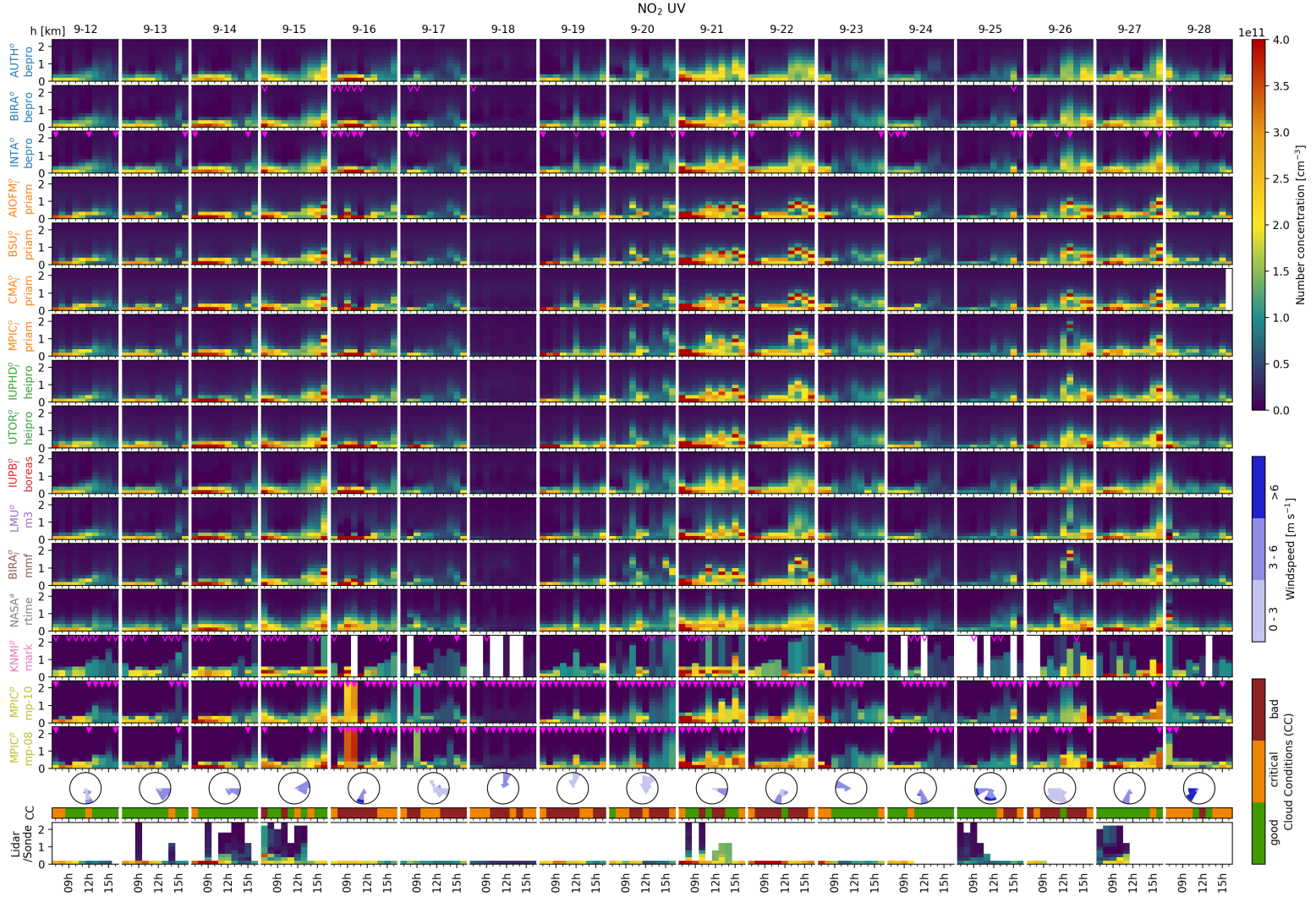

Figure 6. NO_2 UV concentration profiles. The lowest row shows a combined dataset of NO_2 lidar, radiosonde, LP-DOAS and tower in-situ data. Redundant surface concentration measurements were averaged.

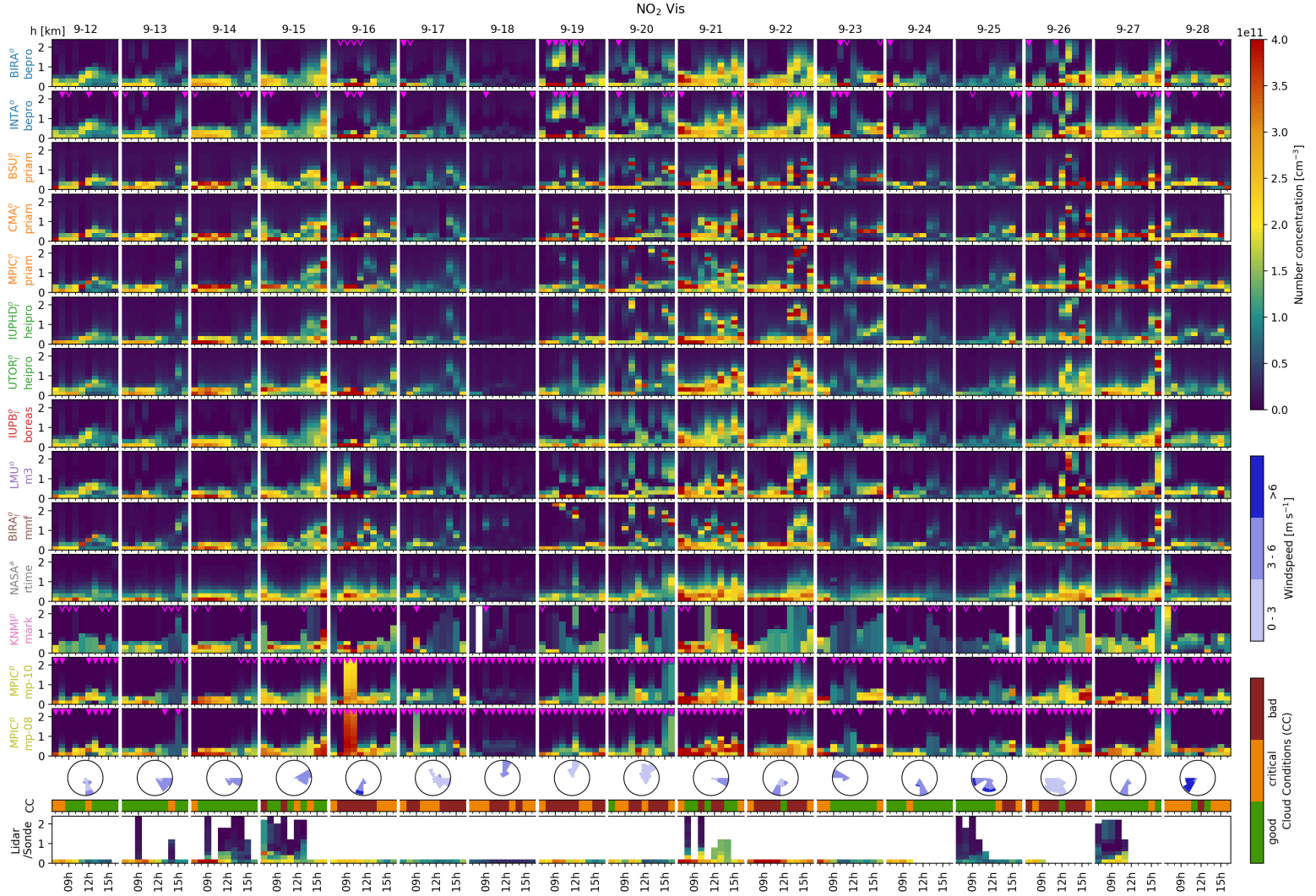

Figure 7. NO_2 Vis concentration profiles. The lowest row shows a combined dataset of NO_2 lidar, radiosonde, LP-DOAS and tower in-situ data. Redundant measurements were averaged here.

Considering valid data only, all algorithms detect similar features in the vertical profiles, but smoothed to different amounts and sometimes detected at different altitudes. For clear sky condition, the observed ASDevs are $3.5 \times 10^{-2}\,\mathrm{km}^{-1}$ for Aerosol UV, $4.0 \times 10^{-2}\,\mathrm{km}^{-1}$ for Aerosol Vis, $1.2 \times 10^{10}\,\mathrm{molec\,cm}^{-3}$ for HCHO, $2.4 \times 10^{10}\,\mathrm{molec\,cm}^{-3}$ for NO_2 UV and $4.4 \times 10^{10}\,\mathrm{molec\,cm}^{-3}$ NO_2 Vis. When regarding participants using the same algorithm, these values are reduced only by about $50\,\%$, indicating that significant discrepancies are caused by differences in the user defined retrieval settings that were not prescribed. The latter are for instance the accuracy criteria for the RTMs, the number of iterations in the inversion, the convergence criteria or the decision at which points of the iteration process the forward model jacobians are (re-)calculated. An example are the discrepancies between UTOR/ HEIPRO and IUPHD/ HEIPRO. In this case the number of applied iteration steps in the aerosol inversion was identified as the main reason: UTOR and IUPHD used 5 and 20 iterations here, respectively. The consequences are evident throughout the comparison. Another example is the aerosol UV retrieval of AUTH/ bePro, where in contrast to other bePRO users oscillations seem to appear. We suspect this to originate from similar reasons, which could not yet been identified.

In general, larger discrepancies appear for the species measured in the Vis spectral range than in the UV. For NO_2 (aerosol) the ASDev increases in the Vis by $50\,\%$ ($90\,\%$). In the case of OEM algorithms, a reason might be that there is lower information content in the UV, meaning that the retrievals are drawn closer to the collectively used *a priori* profile. Further, the larger viewing distance of the Vis retrievals (see Supplement S5) might be problematic, since the exact treatment of the viewing geometries (like the Earth curvature or the treatment of the instrument field of view) gain influence. Note that the worse performance in the Vis was also apparent in the study by Frieß et al. (2019) with synthetic data. The presence of clouds affects ASDevs very differently for different species: for Aerosol UV and Vis it is degraded by a factor of 3 and 4, respectively, which is expected since clouds mostly feature high optical depths > 1 and are detected to very different extent by the individual participants. For HCHO the ASDev decreases by $38\,\%$ which can be well explained by the systematically lower ($-36\,\%$) HCHO concentrations observed under cloudy conditions. ASDevs for NO_2 increase by about $20\,\%$, while the observed concentrations remain similar (increase $< 10\,\%$).

Considering valid data only, the parametrized approaches are mostly in good agreement with the other algorithms. For MAPA, unrealistic results are reliably identified and flagged as invalid, whereas in the case of MARK some valid profiles do not look plausible e.g. for Aerosol Vis on 22 September 2016. For both algorithms a large fraction (30 to $70\,\%$) of the profiles are discarded as invalid or look unrealistic if the retrieval conditions are not ideal (see also flagging statistics in Sect. 4). Gaps in the MARK data appear where no optimum solution could be found at all. For aerosol, OEM algorithms often see elevated layers in the Vis even in clear-sky scenarios that cannot be observed in the UV or the ceilometer profiles. On cloudy days, MMF is capable of detecting clouds as very defined features with a good qualitative agreement with the ceilometer data. In the Vis, even high clouds are detected, e.g. on 17 September and 22 September 2016, which indeed coincide with high-altitude clouds above the retrieval altitude range of $4\,\mathrm{km}$. In contrast to the PAR approaches, OEM and Realtime algorithms yield realistic profiles also under less favourable measurement conditions (e.g. clouds); in particular the OEM results are in qualitative agreement with the ceilometer profiles for many cases.

Regarding HCHO, the agreement of the profiles is exceptionally good considering the particularly low information content of the measurements (due to higher uncertainties in the dSCD data). Probably because observed spatial and temporal concentration gradients are much smaller than for NO_2, which might partly be related to enhanced smoothing by the retrieval, but is also well possible to be real, since HCHO sources (mainly the photolysis of volatile organic compounds) are less localized. High HCHO concentrations coincide with clear-sky conditions and with wind from the continent, which is what would be expected from the current knowledge on the origin and chemistry of atmospheric HCHO. As in the case of aerosol, there are significant discrepancies among the bePRO participants, this time with INTA standing out of the group with slight overestimation.

For NO_2 very shallow layers and large vertical and horizontal gradients might complicate the retrievals. Nevertheless, good ASDev is achieved in the UV. Week-days and weekends (17, 18, 24 and 25 September) can clearly be distinguished. The lowest concentrations are observed on 18 September, where a Sunday coincides with northerly winds from the sea.

The agreement with the supporting observations will be discussed in detail in the following sections.

3.3 Modelled and measured dSCDs

An intrinsic indicator for a successful profile retrieval is a good agreement between the measured and the modelled dSCDs, the latter being the dSCDs obtained from the RTM model for the finally retrieved aerosol and trace gas profiles. Poor agreement might indicate that only a local minimum of the cost function was found (OEM approaches), that inappropriate retrieval settings were chosen (e.g. too small number of iterations in the minimisation) or that the RTM is inaccurate for other reasons, for instance because it cannot describe horizontal inhomogeneities. Figures 8 to 12 show the correlation of measured and modelled dSCDs for all profiles and elevations of each participant. The NASA/ Realtime algorithm is not included since it does not use an RTM and therefore does not provide simulated dSCDs.

For clear-sky conditions, good agreement is achieved by most participants. Only IUPB/ BOREAS, AUTH/ bePRO, BSU/ PRIAM, and KNMI/ MARK exceed relative RMSDs of $10\,\%$ and only for O_4 and NO_2 Vis dSCDs. MMF achieves the best overall performance, being the only algorithm with relative RMSDs $< 5\,\%$ for all species. Regarding HEIPRO, UTOR yields larger RMSD values than IUPHD, which is very likely related to the aforementioned smaller number of iterations applied by UTOR. For the trace gases, small relative RMSD values between $8\,\%$ and $8\,\%$ are achieved for all cloud conditions.

Regarding aerosol, PRIAM and BOREAS feature slightly too low slopes in the UV (approx. 0.9) and more pronounced in the Vis (0.8 to 0.85) interestingly almost exclusively caused by data recorded on the 23 and 27 September where the atmospheric aerosol load is particularly low. RMSDs increase for cloudy scenarios by $10\,\%$ (HCHO), $30\,\%$ (NO_2 UV) and $50\,\%$ (NO_2 Vis, O_4), most likely because the horizontal inhomogeneity cannot be adequately reproduced by the 1D models. This is supported by the comparison results from synthetic data by Frieß et al. (2019), where horizontal homogeneity is inherently assured and the scatter remains similar for all cloud scenarios. KNMI/ MARK has problems to reproduce O_4 dSCDs (relative RMSD > $30\,\%$), while for trace gases the performance is comparable to the other algorithms. Regarding Vis species, M^3 shows outliers under cloudy conditions (while performing excellently in the UV) and bePRO seems to have convergence problems, which was also evident in the synthetic data (Frieß et al., 2019). This problem is overcome by flagging of approx. $10\,\%$ of the data, reducing the RMSD by $> 50\,\%$. PRIAM (except MPIC) shows outliers, in particular for NO_2 Vis. The O_4 scaling factor of 0.8

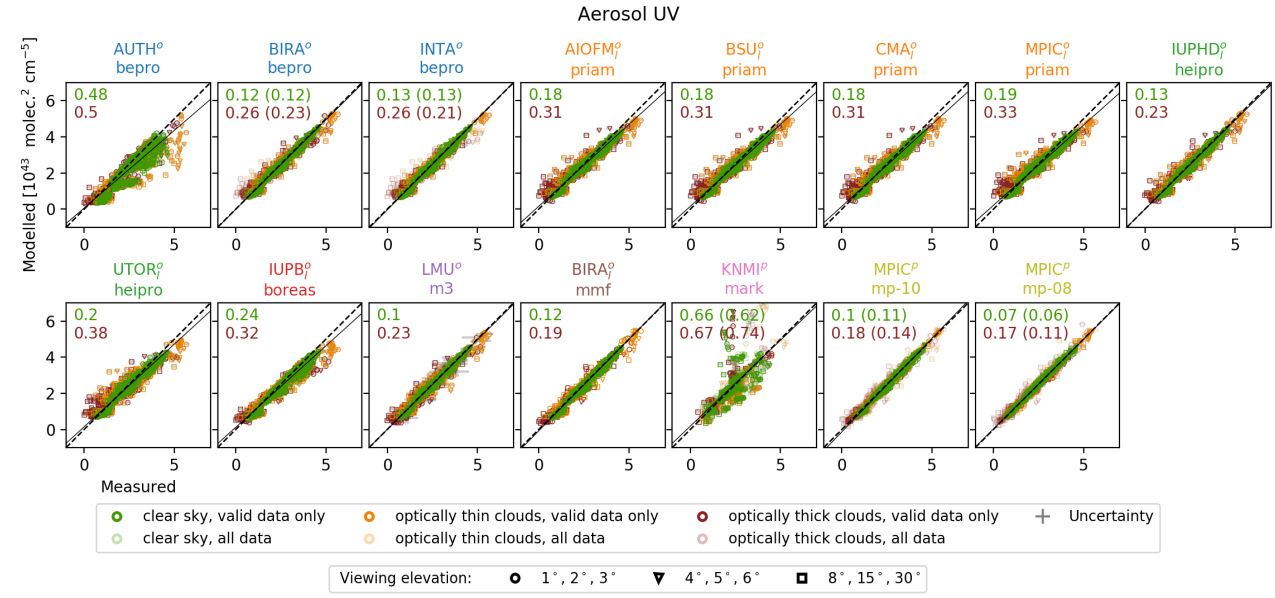

Figure 8. O_4 UV dSCD correlation. Marker colours and marker shapes indicate the cloud conditions and viewing elevation angles, respectively, as indicated in the legend. Numbers represent the measurement-error-weighted RMSD between measured and modelled dSCDs in units of 10^{43} molec^2 cm^{-5} for clear sky (green) and cloudy (red) conditions. Values in brackets were calculated only considering valid data.

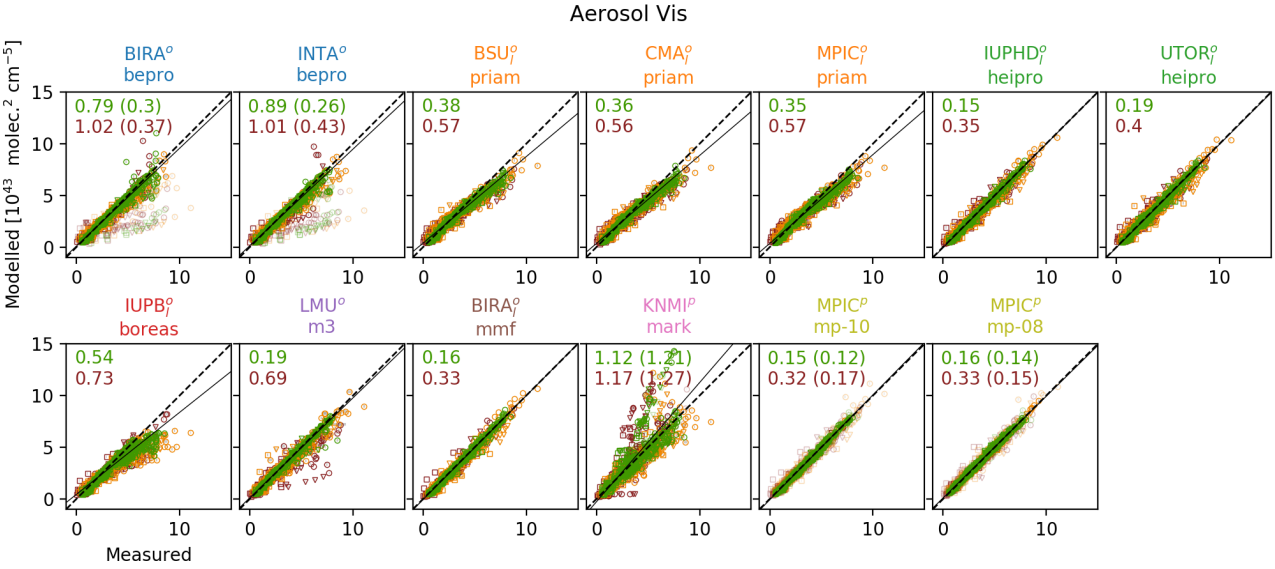

Figure 9. O_4 Vis dSCD correlation. Legends and description of Fig. 8 apply.

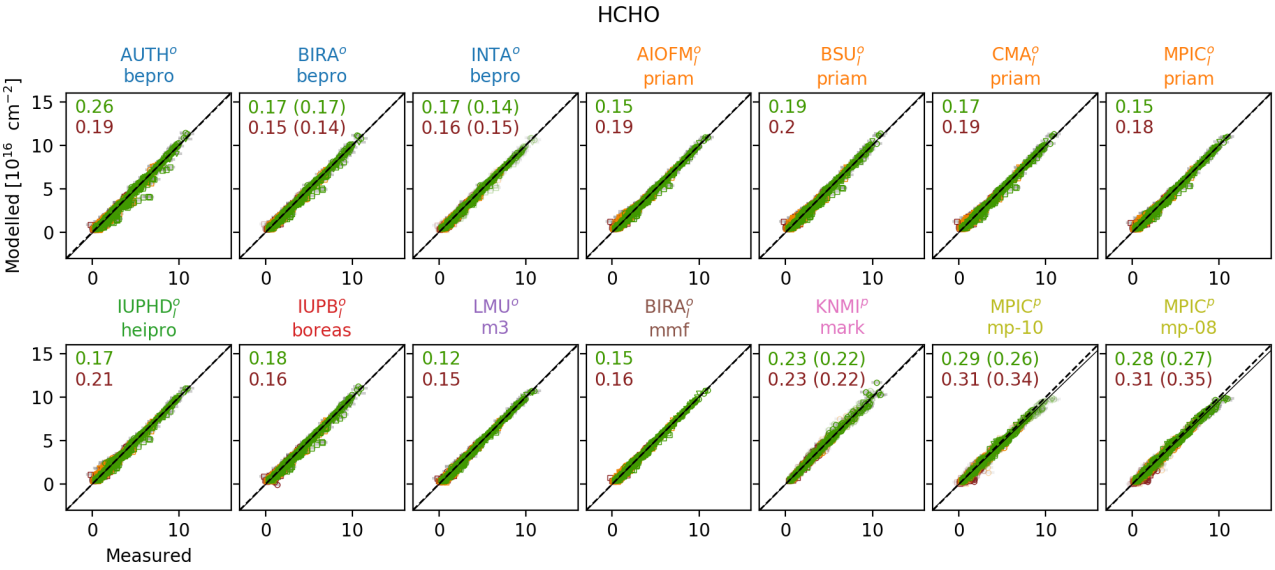

Figure 10. HCHO dSCD correlation. RMSD between measured and modelled dSCDs in units of $10^{16}\,\mathrm{molec\,cm^{-2}}$. Legends and description of Fig. 8 apply.

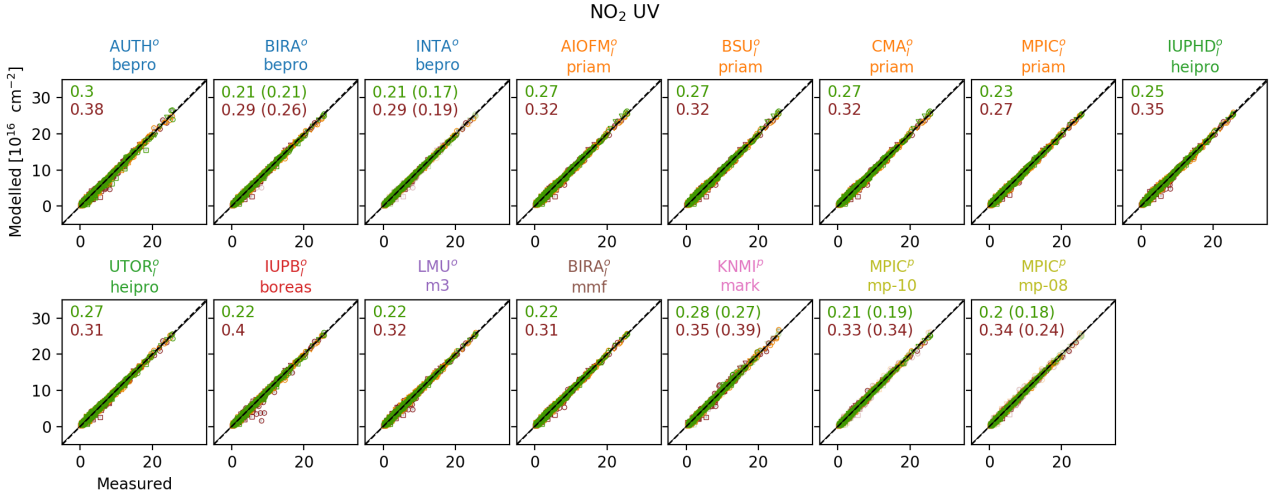

Figure 11. NO_2 UV dSCD correlation. RMSD between measured and modelled dSCDs in units of $10^{16}\,\mathrm{molec\,cm^{-2}}$. Legends and description of Fig. 8 apply.

for MAPA improves O_4 dSCD agreement in the UV by about $35\,\%$ (for clear sky and valid data), but not in the Vis spectral range (see also Supplement S2).

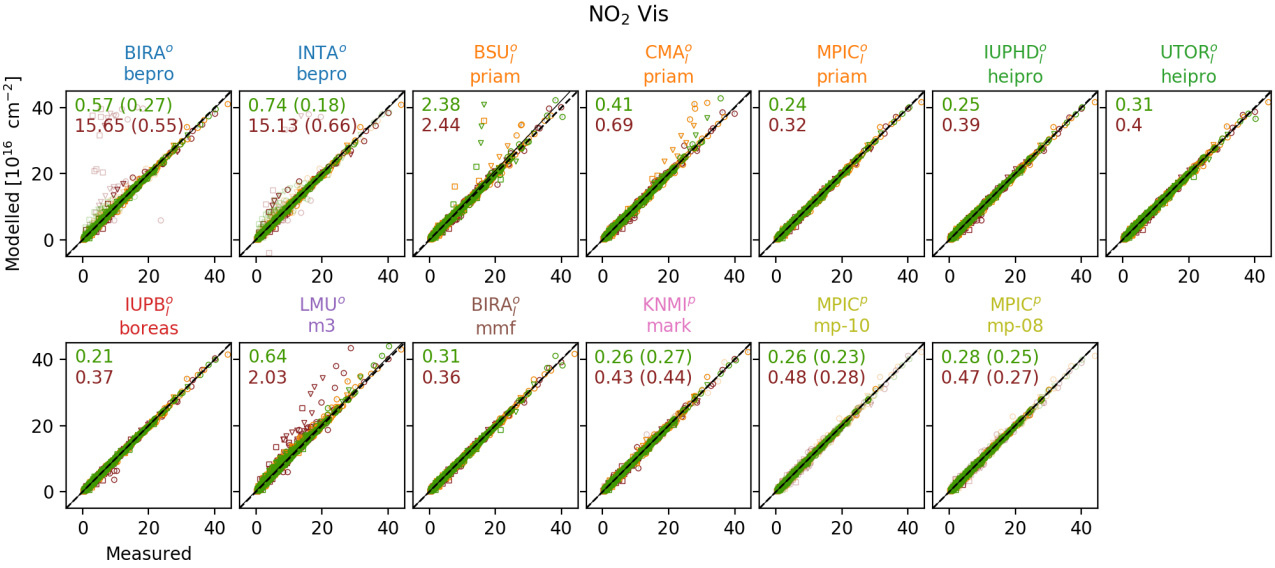

Figure 12. NO_2 Vis dSCD correlation. RMSD between measured and modelled dSCDs in units of $10^{16}\,\mathrm{molec\,cm^{-2}}$. Legends and description of Fig. 8 apply.

3.4 Aerosol optical thickness (AOT)

This section compares vertically integrated MAX-DOAS aerosol extinction profiles with the AOTs observed by the nearby sun photometer. In former publications (e.g. Irie et al., 2008; Clémer et al., 2010; Frieß et al., 2016; Bösch et al., 2018) and also during this comparison study, it was found that MAX-DOAS vertically integrated aerosol profiles systematically underestimate AOTs. It has already been proposed by Irie et al. (2008), Frieß et al. (2016) and Bösch et al. (2018) but not proven that this is related to smoothing effects, namely the reduced sensitivity of MAX-DOAS observations to higher altitudes and associated a priori assumptions. Even though the sensitivity to elevated layers was observed to be increased by the presence of optically thick aerosol layers at the corresponding altitudes (Frieß et al., 2006 and Sect. 3.1 of this study), high-altitude abundances of trace gases and aerosol typically cannot be reliably located and quantified by ground-based MAX-DOAS observations, while aerosol aloft may even introduce systematic errors (Ortega et al., 2016). Integrated profiles rather provide "partial AOTs" which basically only consider low-altitude aerosol and which are additionally biased by *a priori* assumptions on the aerosol extinctions at higher altitudes (for OEM algorithms defined by the *a priori* profile and covariance, for PAR algorithms partly in the form of prescribed profile shapes). Therefore, a comparison between MAX-DOAS and sun photometer is not necessarily meaningful. However, for OEM approaches, information on the true aerosol extinction profile $\boldsymbol{x}$ (which are available from the ceilometer as described in Sect. 2.2.2) and the AVKs $\mathbf{A}$ can be used to account for this effect: inserting $\boldsymbol{x}$ and $\boldsymbol{A}$ into Eq. (9) yields a smoothed profile $\widetilde{\boldsymbol{x}}$ that can be used to estimate which fraction f_τ of the aerosol column is expected to be detected by

the OEM retrievals:

$$f_\tau = \frac{\tau_s'}{\tau_s} = \frac{\sum_i \widetilde{x}_i}{\sum_j x_j} \tag{10}$$

with τ_s' being the actually detectable "partial AOT". The left panel of Fig. 13 shows an example of an extreme case during the campaign from September 15^{th}, 15:00h. Shown are a ceilometer backscatter profile ($\boldsymbol{x}$, black) and the same profile smoothed by the MAX-DOAS median OEM averaging kernels for Aerosol UV and Aerosol Vis ($\boldsymbol{x_{UV}}$ and $\boldsymbol{x_{Vis}}$, blue and green), respectively. In this particular case it is expected that a large fraction of the aerosol above $1\,\mathrm{km}$ altitude will hardly be detected by the MAX-DOAS instruments, resulting in factors $f_\tau = \frac{\tau_s'}{\tau_s}$ of 0.67 and 0.78, for the UV and the Vis AOT, respectively. Note, however, that corresponding information actually seems to be present in the measurements, since part of the high-altitude aerosol appears to be shifted to lower altitudes which are accessible within the constraints of the *a priori* covariance. Multiplying the AOT observed by the sun photometer with f_τ significantly improves the agreement between MAX-DOAS and sun photometer observations in particular in the UV. In the following, this is referred to as "partial AOT correction" (PAC). The right panels in Fig. 13 show information on f_τ and the improvement in the UV and Vis results (2nd and 3rd columns of the figure) over the whole campaign. Average values are $f_\tau = 0.81 \pm 0.16$ in the UV and (0.9 ± 0.13) in the Vis (using the median AVKs of all OEM retrievals). It shall be pointed out that for OEM algorithms the necessity for the PAC can generally be reduced by using improved *a priori* profiles and covariances (e.g. from climatologies, supporting observations and/ or model data). Also the values for f_τ will differ, when other *a priori* profiles and covariances than the ones prescribed for this study (see Sect. 2.1.3) are used. Parametrized and analytical approaches typically do not quantify the sensitivity, the effective resolution or the amount of assimilated *a priori* knowledge. For these algorithms, the correction could not be performed and the total sun photometer AOT τ_s had to be used for the comparison in this section. However, the comparison results and further investigations in Supplement S2 indicate that a scaling of the measured O_4 dSCDs prior to the retrieval with $SF \approx f_\tau$ might be used to at least partly account for the PAC for MAPA and probably other PAR and ANA algorithms (see Supplement S2), even though the motivation for the application of the PAC and the SF are different: the application of the PAC is necessary solely for mathematical reasons related to the concept of OEM and prior constraints applied therein. In contrast, publications that suggest or discuss the application of an SF (e.g. Wagner et al., 2009; Clémer et al., 2010, section 2.2; Ortega et al., 2016; Wagner et al., 2019) directly compare forward modelled O_4 dSCDs (using an atmosphere derived from supporting observations to reproduce the real conditions to best knowledge) to measured O_4 dSCDs. For the determination of the SF, they do not make use of optimal estimation or prior constraints similar to those used in our study. Thus their findings can be ~~considered~~ in general regarded as independent from any kind of PAC~~.~~, even though PAC and SF have similar impact on the MAX-DOAS AOT results with the *a priori* assumptions applied in this study. Particularly, it shall be pointed out that our findings regarding the PAC have no implications on whether elevated aerosol layers explain the necessity of the SF (as proposed by Ortega et al., 2016), or not.

Figure 14 shows time series of the MAX-DOAS retrieved AOTs in comparison to their median and the sun photometer data. For the sun photometer, both the total AOT τ_s and the partial AOT τ_s' are shown. For the calculation of τ_s' in Fig. 14, the median AVKs of all OEM participants were used for the smoothing according to Eq. (9). In the correlation analysis (Fig. 15), AVKs of the individual participants and the individual profiles were applied. Keep in mind that the non-OEM approaches (NASA/

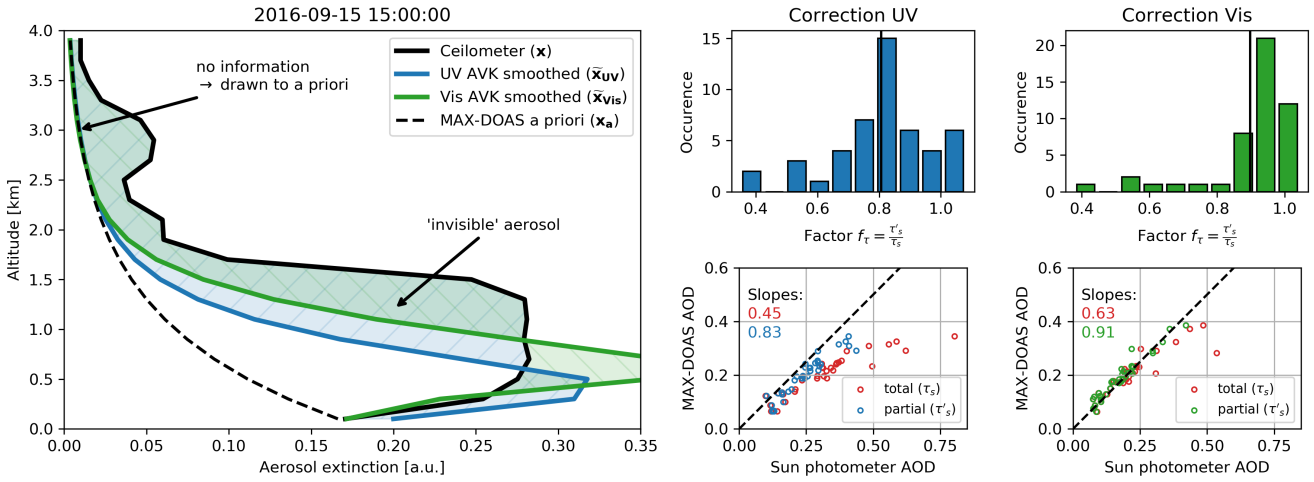

Figure 13. Left panel: example for the smoothing of a ceilometer backscatter profile $\boldsymbol{x}$ (according to Eq. (9)) with particularly heavy aerosol load at high-altitudes retrieved in the UV and Vis, respectively. Right panel: distribution and impact of the correction factor $f_\tau = \tau'_s/\tau_s$ for the UV and the Vis retrieval. Top plots show the distributions of f_τ with the solid lines indicating the mean values. At the bottom the correlation plots between sun photometer and MAX-DOAS median AOTs are shown. Red circles represent sun photometer total AOTs, other dots represent the partial AOT $\tau_s = f_\tau \cdot \tau'_s$.

Realtime, KNMI/ MARK and MPIC/ MAPA) are correlated against τ_s and are therefore expected to generally achieve worse agreement. For correlations of OEM algorithms against τ_s please refer to Supplement S8.3. Correlation parameters, RMSD and Bias values were derived as described in Sect. 2.3.

Under clear sky conditions, average RMSD values against the MAX-DOAS median are 0.028 in the UV and 0.032 in the Vis. In the presence of clouds they increase by about 30 % and 80 %, respectively, which is to mainly due to the periods of particularly large scatter between 16 and 19 September 2016. As already shown in Sect. 3.2, different algorithms detect clouds to very different extent. Especially in the presence of optically thick clouds (AOT > 10), this easily induces discrepancies of several orders of magnitudes. The observed average RMSDs are similar to the specified uncertainties (average is 0.025) that are derived from propagated measurement noise and smoothing effects. Keeping in mind that the retrievals were performed on a common dSCD dataset, this indicates that the choice of the retrieval algorithm and the remaining free settings have severe impact on the results.

For the comparison to the sun photometer, it shall be noted that the PAC induces further uncertainties, as it incorporates the extinction profiles derived from the ceilometer and the algorithms' AVKs, both being error-prone. Further, the comparison to sun photometer data under cloudy conditions might not be very meaningful as (1) there are only 13 measurements available in the presence of clouds and (2) as it is very likely that these measurements were made by looking through very local cloud holes, such that they will not be representative for the MAX-DOAS retrieved AOTs with a typical horizontal sensitivity range of several kilometres (see Supplement S5). The following discussion of the sun photometer comparison therefore refers to

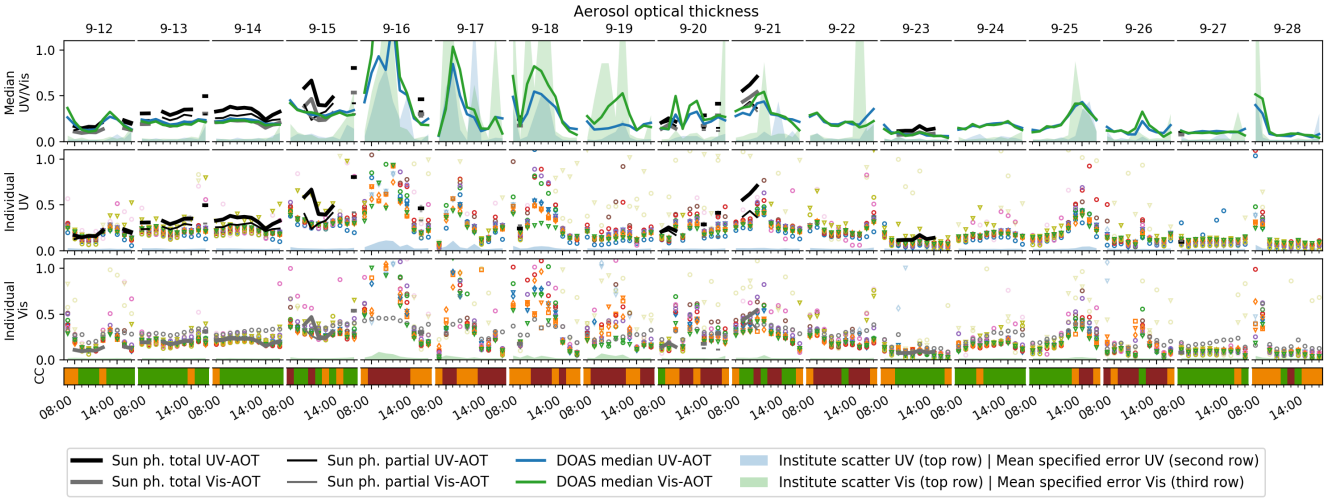

Figure 14. MAX-DOAS retrieved AOTs in comparison to sun photometer data. Symbol and symbol colours are chosen according to Table 2. Transparent symbols indicate data flagged as invalid. Top row: MAX-DOAS median results vs. the available supporting observations, according to the legend below the plot. The "institute scatter" areas show the scattering among the participants in terms of standard deviation with valid data considered only. Two lower rows: Comparison of the individual participants for the two spectral retrieval ranges. Here the coloured area is the average retrieval error, as specified by the participants.

clear-sky conditions and valid data only. In general, there is reasonable agreement of the MAX-DOAS retrieved AOT with the sun photometer, with average observed RMSDs of 0.08 (0.06) for Aerosol UV (Vis). Best performance in the UV is observed for IUPHD/ HEIPRO and LMU/ M^3 with RMSDs around 0.05, in the Vis it is the participants using the bePRO (BIRA and INTA), the HEIPRO (IUPHD and UTOR) and the BOREAS (IUPB) algorithm. For all participants except MPIC-0.8/ MAPA, negative Biases < -0.03 in the UV remain, even though the PAC has been applied for the OEM algorithms. The average Bias in the UV is -0.06, indicating that the systematic underestimation dominates over random deviations here. Note that the slopes and intercepts vary significantly among the participants, however, in an anti-correlated manner, finally resulting into similar Bias values.

The average Bias in the Vis is only 0.02. Bias magnitudes are much smaller than RMSDs for many participants here, indicating that in these cases Vis AOTs mainly suffer from random discrepancies. BePRO suffers the aforementioned convergence problems during inversion in the Vis (see Sect. 3.3) but the affected results are reliably flagged. KNMI/ MARK, NASA/ Realtime and MPIC-1.0/ MAPA feature the highest RMSDs around 0.1 and strongest Biases below -0.1 in the UV. A particular case is KNMI/ Aerosol Vis with RMSD> 0.2, with and without flagging being applied.

As described in Supplement S2, the PAC and the application of an O_4 dSCD scaling factor of $SF \approx f_\tau$ have very similar impact on the AOT correlation. Consequently, the application of $SF = 0.8$ in the case of MPIC-0.8/ MAPA significantly improves the agreement to the sun photometer total AOT in the UV ($f_\tau \approx 0.8$) whereas in the Vis ($f_\tau \approx 0.9$) it leads to an overcompensation with a Bias of about 0.05.

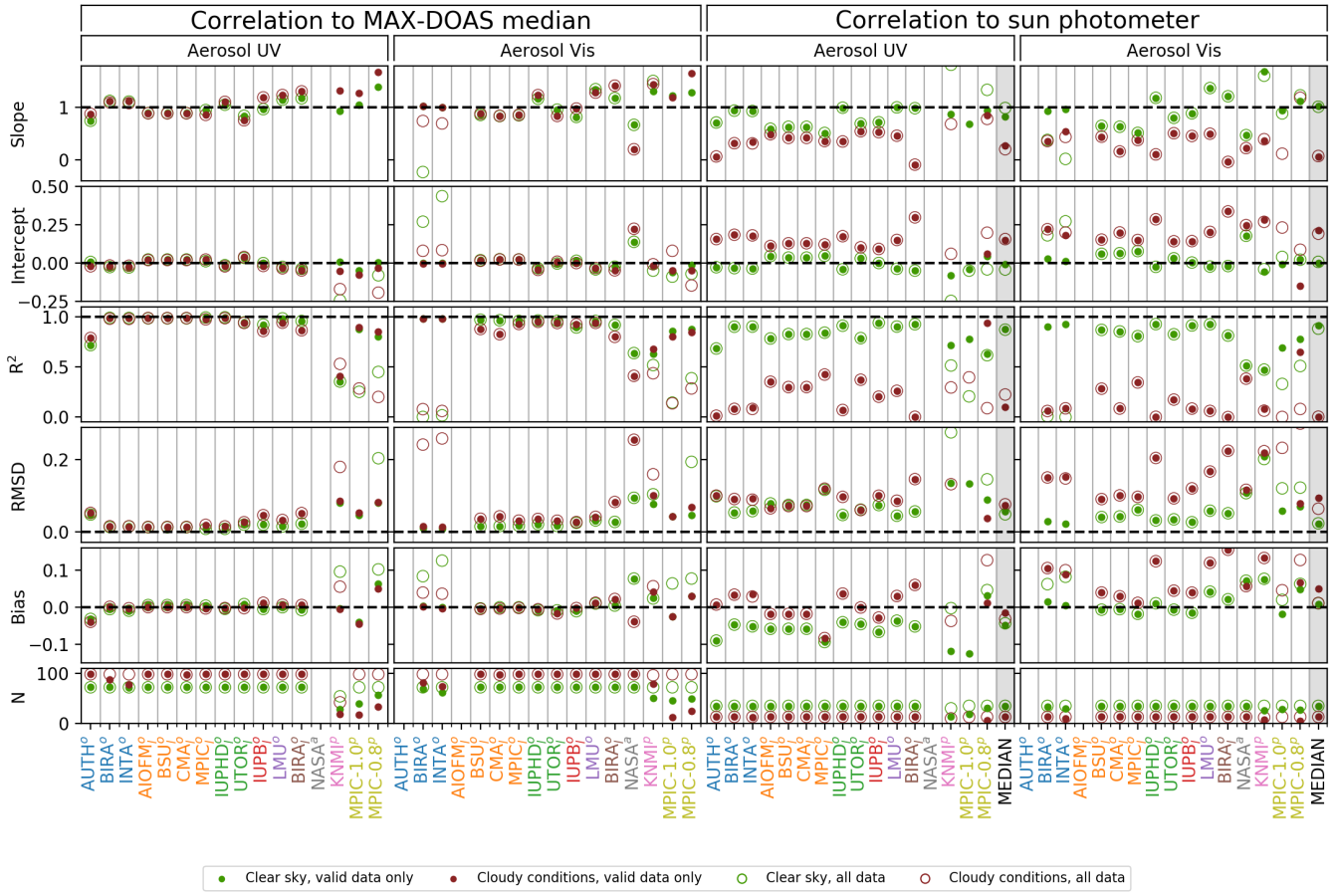

Figure 15. Correlation statistics for AOTs. The two left columns give an impression on the agreement among the institutes, as they show the correlation of the individual participant's retrieved AOT (ordinate of the underlying correlation plot) against the median (abscissa). The two right columns show the correlation against the sun photometer AOT (partial AOT in the case of OEM retrievals) instead of the median. Green and red symbols represent cloud-free and cloudy conditions, respectively. Hollow circles represent values for all submitted data, the dots only consider data points flagged as valid. N is the number of profiles which contributed to the respective data points above. The total number of submitted profiles per participant and species were 170. On the right also the correlation between the MAX-DOAS median results and supporting observations are included (grey shaded columns). The correlation plots are shown in Supplement S8.3.

3.5 Trace gas vertical column densities

This section assesses the consistency of the VCDs for each of the trace gases HCHO and NO_2. Independent observations of VCDs are the direct-sun DOAS observations, but also integrated columns of radiosonde and lidar profiles (NO_2 only). Time series comparisons of all observations are shown in Fig. 16 and 17. For the statistical evaluation in Fig. 18, from the supporting observations only direct-sun observations were considered, as they provide the most complete dataset.

As for AOTs, smoothing effects potentially affects the comparability of MAX-DOAS and direct-sun observations. In contrast to aerosol, only scarce (NO_2) or no (HCHO) information on the true profile is available and a correction similar to the PAC cannot be performed. However for NO_2 the available radiosonde profiles could be used for an impact estimate. Ignoring one problematic radiosonde profile on 09-27 07:00:00 (where NO_2 concentration was close to the radiosonde detection limit and thus instrumental offsets became particularly apparent), correction factors of 1.06 ± 0.05 in the UV and 1.03 ± 0.03 in the Vis are obtained, indicating that the MAX-DOAS retrieved tropospheric NO_2 VCD is affected by smoothing effects to only a few percent. This is expected since NO_2 mostly appears close to the ground. Also in Fig. 6 and 7, NO_2 appears to be confined to the lowermost retrieval layers with concentrations dropping to around zero already at altitudes where MAX-DOAS sensitivity is still significant. Profiles from the NO_2 lidar were not used in this investigation as they often suffer from artefacts at higher altitudes. Regarding HCHO, the MAX-DOAS profiling results on some days show large concentrations over the whole altitude range where the information content of the measurements is significant (compare Fig. 2 and 5), indicating that there might be "invisible" HCHO at even higher altitudes. This is supported by Fig. 16, where MAX-DOAS observations tend to yield smaller VCDs than the direct-sun observations in particular in scenarios with high HCHO abundance.

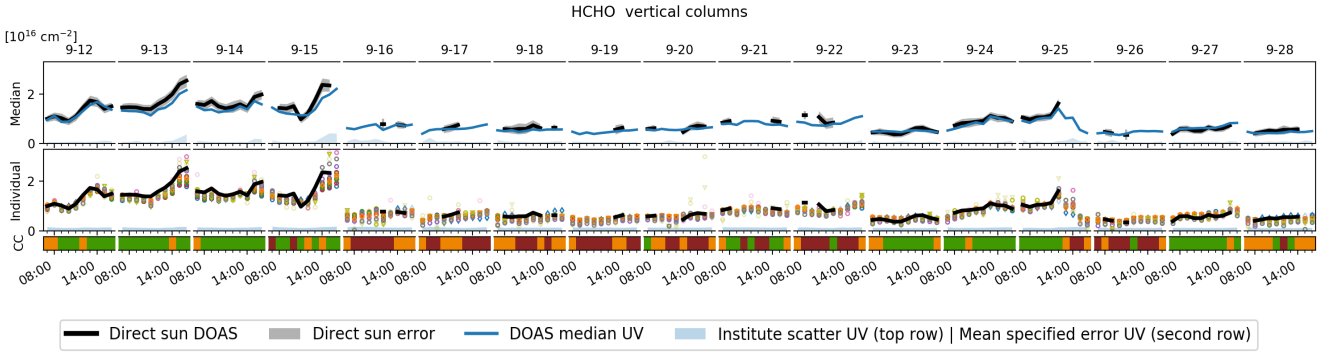

Figure 16. Comparison of MAX-DOAS retrieved HCHO VCDs vs. direct-sun DOAS. Basic descriptions of Fig. 14 apply.

Under clear sky conditions, average RMSD values against the MAX-DOAS median are $5 \times 10^{14}\,\mathrm{molec\,cm^{-2}}$ for HCHO and $7 \times 10^{14}\,\mathrm{molec\,cm^{-2}}$ for NO_2 (both UV and Vis). In contrast to AOTs, these values do not increase significantly ($< 15\,\%$) in the presence of clouds. For HCHO it is even reduced by $25\,\%$ for the same reasons as discussed already in Section 3.2. Bias values are approximately of half the magnitude of RMSDs for all trace gases.

For HCHO, the comparison against the direct-sun DOAS observations yields an average RMSD of $1.4 \times 10^{15}\,\mathrm{molec\,cm^{-2}}$. Note however that the two observations are not fully independent, as for the direct-sun data, the residual HCHO amount in the reference spectrum was adapted from the MAX-DOAS VCD (see Sect. 2.2.4). Bias values are of the order of $35\,\%$ of the RMSDs, indicating that the deviations are mostly random.

For NO_2 UV (Vis) the comparison to the direct-sun DOAS yields an average RMSD of $3.7 \times 10^{15}\,\mathrm{molec\,cm^{-2}}$ ($3.8 \times 10^{15}\,\mathrm{molec\,cm^{-2}}$), which is about five times the average RMSD of the MAX-DOAS median comparison. Between 12 and 14 September the direct sun VCDs but also most radiosonde and lidar observation are systematically lower than the MAX-DOAS

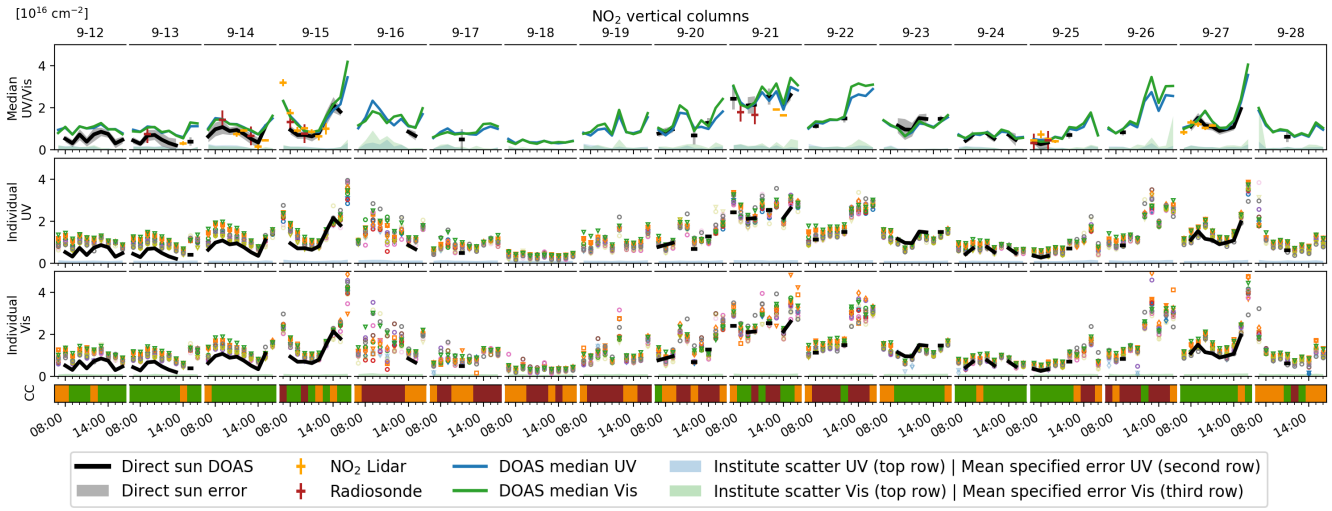

Figure 17. Comparison of MAX-DOAS retrieved NO_2 VCDs vs. direct-sun DOAS, NO_2 lidar and radiosonde. Basic descriptions of Fig. 14 apply.

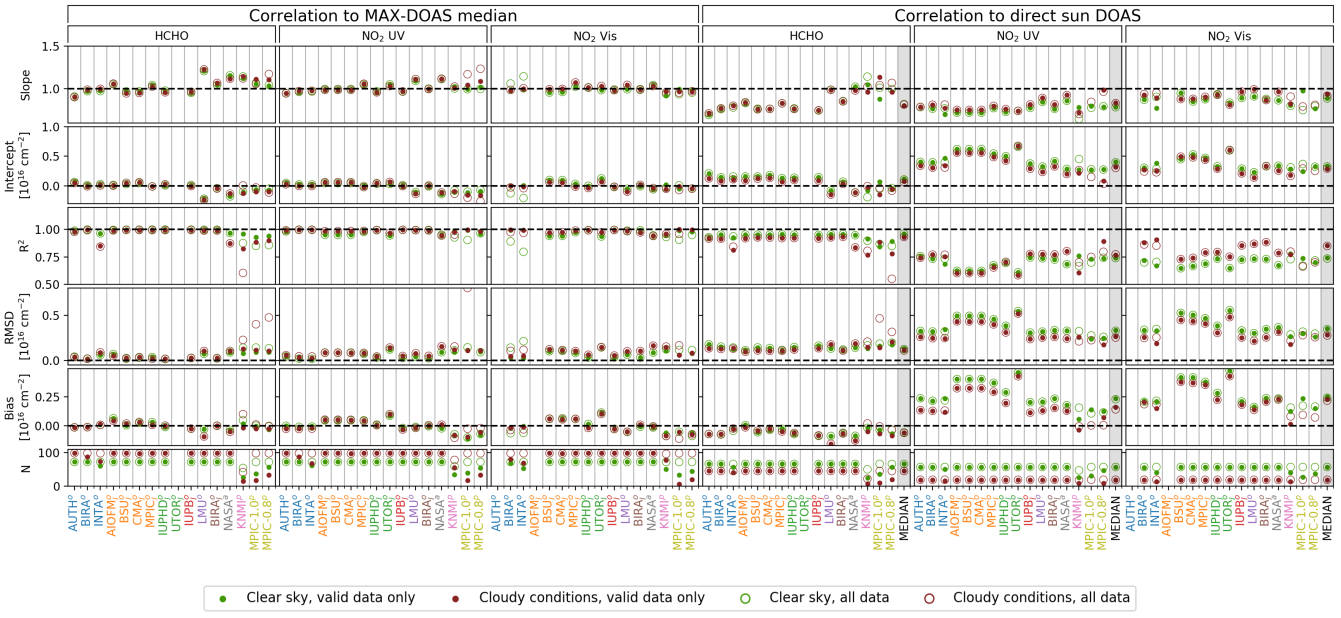

Figure 18. Correlation statistics of trace gas VCDs. The plot is similar to Fig. 15. In the underlying correlation plots, ordinates are MAX-DOAS VCDs of individual participants and abscissas are the MAX-DOAS median and direct-sun VCDs, respectively. The correlation plots are shown in Supplement S8.3.

VCDs. This is also reflected in the correlation statistics: RMSDs and Bias values of different participants appear strongly correlated in Fig. 18 and Bias magnitudes are $> 70\,\%$ of the RMSDs for both UV and Vis. The reason could not yet be identified. Interestingly, this contrasts with findings on the surface concentration in the following section, where discrepancies to the LP-DOAS are dominated by random deviations.

In contrast to the AOTs, the RMSDs against the MAX-DOAS median here are smaller than the specified retrieval errors, which are $1.3 \times 10^{15}\,\mathrm{molec\,cm^{-2}}$ for HCHO, $1.3 \times 10^{15}\,\mathrm{molec\,cm^{-2}}$ for NO_2 UV and $1.2 \times 10^{15}\,\mathrm{molec\,cm^{-2}}$ for NO_2 Vis. On the other hand NO_2 RMSDs against the direct-sun observations are about three times larger. For the less abundant HCHO, the signal-to-noise ratio in the median dSCDs is smaller than for other species, such that the specified uncertainties derived from the dSCD noise are larger and more representative for the actual retrieval accuracy.

3.6 Trace gas surface concentrations

This section compares the number concentration of NO_2 and HCHO observed at the surface. Note that in this paper "surface concentration" refers to the average concentration in the lowest MAX-DOAS retrieval layer extending from 0 to 200 m altitude. Independent observations are the LP-DOAS (NO_2 and HCHO), and the surface values of radiosonde and lidar profiles (NO_2), as well as integrated values of in situ measurements in the tower (described in Sect. 2.2.5). Comparisons of all observations are shown in Fig. 19 and 20. For the statistical evaluation (Fig. 21) only LP-DOAS data were considered since they provide a very accurate, representative and complete dataset (see Section 2.2.5). The impact of profile smoothing during the retrieval on the retrieved surface concentration was estimated for NO_2 in Supplement S9 from available radiosonde and lidar NO_2 profiles and was found to be around $5.5 \times 10^9\,\mathrm{molec\,cm^{-3}}$ ($4 \times 10^9\,\mathrm{molec\,cm^{-3}}$) in the UV (Vis). Typical RMSD values in the comparison with the LP-DOAS are about one order of magnitude larger, indicating that the impact of smoothing on the NO_2 surface concentration is negligible in this study.

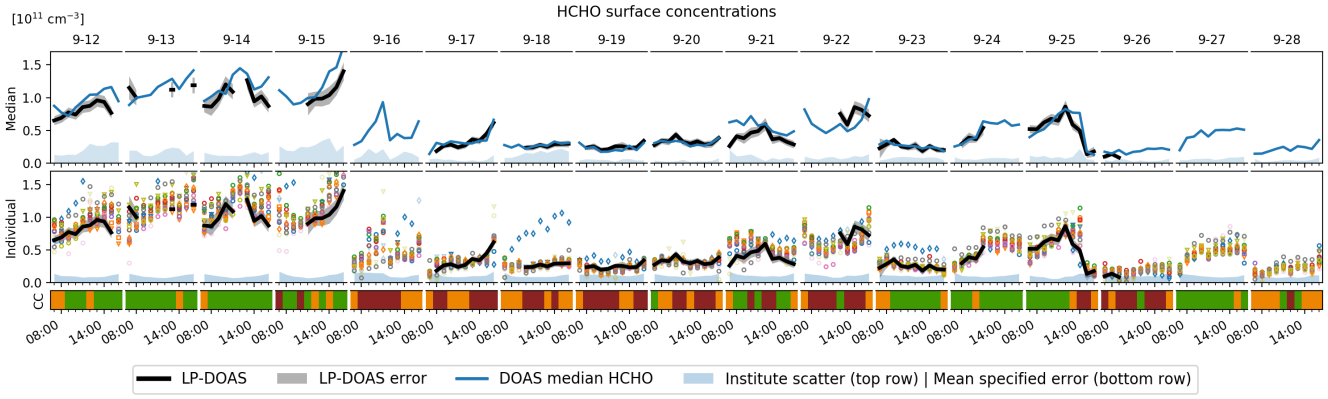

Figure 19. Comparison of MAX-DOAS retrieved HCHO surface concentrations. Basic descriptions of Fig. 14 apply. Note that the mean specified uncertainties in the two lower rows of the figure are very small and thus barely visible.

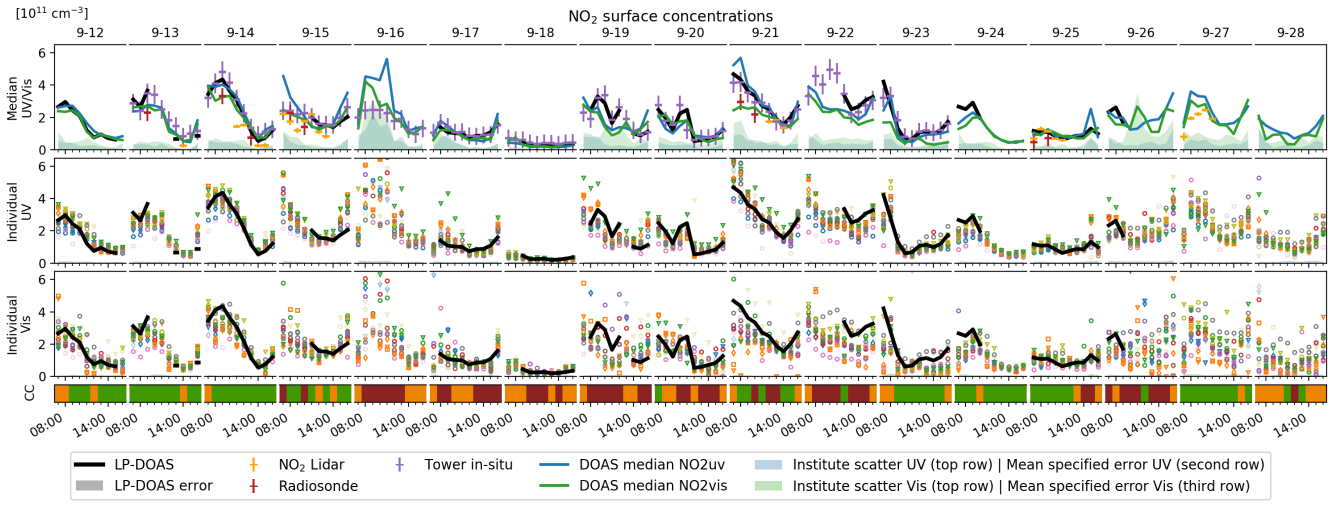

Figure 20. Comparison of MAX-DOAS retrieved NO_2 surface concentrations. Basic descriptions of Fig. 14 apply. Note that the mean specified uncertainties in the two lower rows of the figure are very small and thus barely visible.

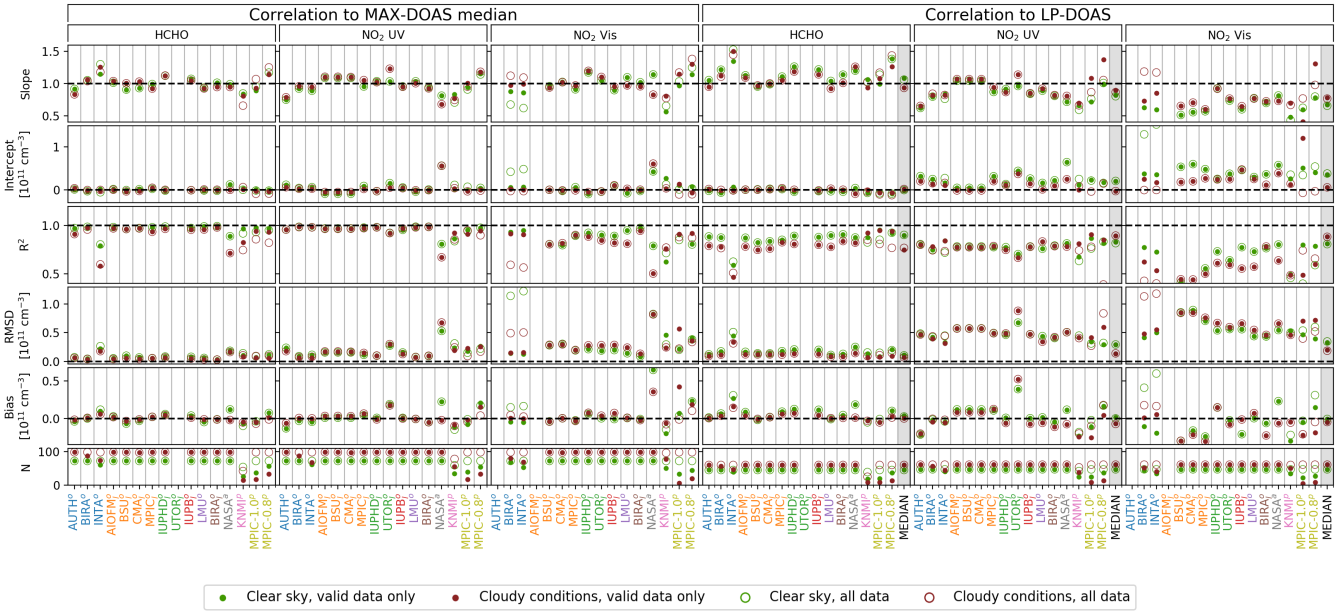

Figure 21. Correlation statistics of trace gas surface concentrations. The plot is similar to Fig. 15. In the underlying correlation plots, ordinates are MAX-DOAS surface concentrations of individual participants and abscissas are the MAX-DOAS median and direct-sun VCDs, respectively. The correlation plots are shown in Supplement S8.3.

The comparisons of surface concentrations are particularly useful, because the largest set of validation data is available here and because in contrast to the comparison of AOT and VCDs, the surface concentration comparison requires an isolation of the

surface layer from the layers above and therefore reflects the MAX-DOAS' ability to actually resolve vertical profiles at least close to the surface.

Figures 19 and 20 show good qualitative agreement between all observations most of the time, even in the presence of clouds. Apparent exceptions for NO_2 are the fog event on 16 September (strong scatter among the participants) and at forenoon on 22 September (MAX-DOAS median shows large deviations compared to the tower measurements probably due to a very local NO_2 emission event close to the tower).

Under clear sky conditions average RMSDs observed for the comparison to the MAX-DOAS median results are $8.8 \times 10^9\,\mathrm{molec\,cm^{-3}}$ for HCHO, $1.8 \times 10^{10}\,\mathrm{molec\,cm^{-3}}$ for NO_2 UV and $2.7 \times 10^{10}\,\mathrm{molec\,cm^{-3}}$ for NO_2 Vis. For the comparison to the LP-DOAS, these values increase to $1.8 \times 10^{10}\,\mathrm{molec\,cm^{-3}}$, $4.7 \times 10^{10}\,\mathrm{molec\,cm^{-3}}$ and $5.6 \times 10^{10}\,\mathrm{molec\,cm^{-3}}$, respectively. For the median comparison, Biases magnitudes are about $40\,\%$ of the RMSD values. In contrast to the VCDs, deviations to the supporting observations (LP-DOAS) seem to be random to large part, as Bias magnitudes are about three times smaller than RMSDs. Significant Biases are only observed for some participants, e.g. UTOR/ HEIPRO in the UV.

Clouds have very different impact on the results: the average RMSD to the median increases by 15 for HCHO, 26 for NO_2 UV and $38\,\%$ for NO_2 Vis, whereas the average RMSD to the LP-DOAS is even reduced by 4, 15 and $17\,\%$, respectively. A large fraction of the scatter in the comparison to the LP-DOAS might be related to the spatio-temporal variability of the gas concentrations, in particular in the Vis spectral range, where the MAX-DOAS viewing distance is large. The good agreement of the surface concentrations with the supporting observations during the first days is opposite to the VCD comparison, which at least for NO_2 points to a problem with the retrieval results in higher layers or the direct-sun data. For NO_2 Vis, the agreement is generally worse than for NO_2 UV. Convergence problems of bePRO appear again in the form of outliers (see in particular the RMSD values), which are efficiently removed by flagging. INTA shows strong systematic outliers over whole days (e.g. on 18 September), which are not observed for other bePRO users and are very likely produced by technical problems. Again, as for AOTs and VCDs, the scatter among the participants is similar or larger than the specified errors even for clear-sky conditions (factors of about one for HCHO, two for NO_2 UV and three for NO_2 Vis, see Fig. 19 and Fig. 20).

3.7 Retrieval from dSCDs of individual participants

As described in Sect. 2.1.1, the results compared so far were retrieved from a common set of median dSCDs. Thus, the results only illustrate the performance of the different retrieval techniques. However, it is also interesting to compare collocated MAX-DOAS measurements which are fully independent, to obtain an estimate of the reliability of a typical MAX-DOAS profile measurement undergoing the whole spectra acquisition and data processing chain. Therefore, the study above was once more conducted with each participant using their own measured dSCDs (see Kreher et al., 2019, for dataset details). Supplement S10 shows further details by means of figures that are equivalent to those shown before in the course of the median dSCD comparison. A summary is given in Table 5 which shows the increase in average RMSD and average Bias magnitude for the most important comparisons (as described in the precedent subsections for the median dSCDs) when participants use their own instead of the median dSCDs. Only valid data of participants appearing in both studies were considered and BIRA/ bePRO and

KNMI were excluded because in contrast to the median dSCD study BIRA/ bePRO and KNMI did not submit flags for the own dSCD study, which heavily impacted the results.

Table 5. Relative increase in average RMSD (first value) and average Bias magnitude (values in brackets) when participants retrieve profiles from their own dSCDs instead of using the median dSCDs. Values are given for clear sky and cloudy conditions separately. Further the comparisons among the participants (to the MAX-DOAS median) and the comparisons to the supporting observations (sun photometer AOTs, direct-sun DOAS VCDs and LP-DOAS surface concentrations) are distinguished.

		Clear sky		Cloudy	
		To median [%]	To supp. obs. [%]	To median [%]	To supp. obs. [%]
Observation	Species				
AOT	Aerosol UV	29 (37)	-10 (-16)	32 (48)	45 (58)
	Aerosol Vis	29 (55)	18 (15)	26 (110)	21 (37)
VCD	HCHO	175 (187)	66 (109)	152 (113)	46 (32)
	NO_2 UV	45 (52)	-8 (-18)	45 (31)	-8 (-30)
	NO_2 Vis	43 (8)	6 (13)	27 (-8)	3 (-2)
Surface	HCHO	87 (64)	16 (34)	120 (129)	37 (82)
	NO_2 UV	28 (53)	10 (64)	25 (76)	1 (45)
	NO_2 Vis	13 (11)	6 (37)	-9 (-42)	-13 (-12)

Regarding only the increase in RMSD in the MAX-DOAS median comparison (hence, the degradation of consistency among the participants) is qualitatively consistent with what one would expect from the findings by Kreher et al. (2019) on the CINDI-2 dSCD consistency: for NO_2, almost all participating instruments were able to deliver good quality dSCDs suitable for profile inversion, while for HCHO the quality was much more variable, resulting in the stronger degradation given in Table 5. Kreher et al. (2019) identified instrumental characterisation (e.g. detector non-linearity and stray-light in the spectrometer) and pointing issues as the main sources of discrepancy between the participant's own dSCD datasets. The degradation is smaller for the surface concentrations than for the trace gas VCDs and is very similar for different cloud conditions.

For the comparison to the supporting observations, the increase in average RMSD is smaller (second and fourth column of Table 5). This means, that even though using the own dSCDs induces differences among the participants, the average quality of the dSCDs is basically maintained or at least small compared to the discrepancies induced by the retrieval techniques. Interestingly, the RMSD and Bias values for the UV AOT and NO_2 VCD even decrease, indicating that the median dSCDs suffer from systematic errors. Under clear sky conditions, low impact ($\leq 10\,\%$) was found for Aerosol UV AOTs and NO_2 data products. Particularly large impact is observed for HCHO VCDs ($66\,\%$). Under cloudy conditions, the impact on NO_2 products remains small (again $< 10\,\%$), whereas for all other products, the increase in average RMSD exceeds $20\,\%$.

It is also of interest to explicitly estimate which fractions of the total observed discrepancies among MAX-DOAS observations are caused either by the use of different retrieval algorithms or by inconsistencies in the dSCD acquisition. Note that the RMSD values from the median dSCD comparison represent the error arising solely from using different algorithms while the

RMSD values from the own dSCD comparison represent the combined effect of both aspects. For simplicity, we assume that the contributions of both aspects are random and independent so that the effect of using own dSCDs can be isolated by simple RMSD error calculations. ~~In this way, its contribution to the total variance observed among the participants under clear sky conditions can be estimated to~~ For clear-sky conditions, we find that the differences in the measured dSCDs are responsible for approximately 40% (for AOTs), 85% (HCHO VCDs), 70% (HCHO surface concentrations), 50% (NO_2 VCDs), 40% (NO_2 UV surface concentrations) and 20% (NO_2 Vis surface concentrations) ~~, respectively~~of the total variance observed among the participants. The residual variance can be attributed to the choice and setup of the retrieval algorithm. The residual variance can be attributed to the choice and setup of the retrieval algorithm.

4 Conclusions

Within this study, 15 participants used 9 different profiling algorithms with 3 different technical approaches (optimal estimation (OEM), parametrized (PAR) and analytical (ANA) approach) to retrieve aerosol and trace gas (NO_2, HCHO) vertical profiles from a common set of dSCDs which was recorded during the CINDI-2 campaign. The results were compared and validated against colocated supporting observations with the focus on aerosol optical thicknesses (AOTs), trace gas vertical column densities (VCDs) and trace gas surface concentrations. Data from some supporting observations were used for qualitative comparison only (Ceilometer, NO_2 radiosondes, NO_2-Lidar, NO_2 in-situ instruments) while for a statistical assessment AOTs from the sun photometer, VCDs from direct-sun DOAS observations and surface concentrations from the LP-DOAS were used.

Figure 22 shows an overview of RMSD and Bias values for the correlation between measured and modelled dSCDs and the comparisons to supporting observations. General strengths and weaknesses of different algorithms become particularly apparent here. Very good overall performance without the need for validity flagging is achieved by the MMF and the M^3 algorithm. Note that the results for aerosol are of very similar quality, even though in contrast to M^3, MMF retrieves aerosol in the logarithmic space. For valid data (about 20% discarded) INTA also shows good overall performance apart from the outliers in the HCHO surface concentration, which are very likely related to technical problems. Very good performance for aerosol is observed for IUPHD/ HEIPRO over the full dataset. For NO_2, best performance is achieved by MAPA. The AOT comparison looks generally worse for parametrized approaches which is expected since no partial AOT correction can be performed and thus - with the MAX-DOAS integrated extinction profile and the sun photometer total AOT - basically two different quantities are compared. Finally, the Realtime algorithm by NASA (being the only ANA algorithm) shall be pointed out: despite its simplified radiative transport and the associated outstanding computational performance it provides reasonable results for trace gases (RMSD/ Average RMSD around unity).

Parametrized approaches appear to be less stable in the sense that for less favourable conditions no convergence is achieved or inconsistent results are returned (30 to 70% of all profiles). For MAPA, these cases are reliably identified and flagged as invalid such that the remaining results achieve very good RMSD and Bias values. In contrast for MARK, even some profiles considered valid do not look plausible. The instability of parametrized algorithms is likely related to the approach: in reality, a vertical profile can be described by an arbitrarily large set of parameters and the information on those contained in a MAX-

DOAS measurement depends on the atmospheric conditions, hence the profiles themselves. For parametrized approaches, the number of retrieved parameters is reduced to the number of typically observed DOFs by describing the profile by a few prescribed (not necessarily orthogonal) parameters. Lack of information in those due to particular atmospheric conditions (also if information is available but only on parameters not covered by the chosen parametrization) leads to an under-determined problem with ambiguous solution and the inversion fails. For OEM approaches, the information can be dynamically distributed to a larger number of parameters (20 in this study, namely the species abundances in the retrieval layers) while parameters of few or no information are constrained by *a priori* information. This is why OEM inversions converge under a broader range of atmospheric conditions even when information from the measurement is reduced or shifted between retrieved parameters. On the other hand, this means that OEM algorithms even provide plausibly looking profiles (basically the a priori profile) when few/no information is contained in the measurements. Even though such cases can be identified by examining the AVKs, this makes OEM retrievals prone to misinterpretations particularly by inexperienced users.

Regarding full profiles, the overview plots in Sect. 3.2 and figures in Supplement S8.2 show a good qualitative agreement between the algorithms for valid data and clear-sky conditions. In most cases they detect the same features, however sometimes at different altitudes and of different magnitude. Under clear-sky conditions, the RMSDs between individual participants and the MAX-DOAS median results range between $(0.01 - 0.1)$ for AOTs, $(1.5 - 15) \times 10^{14}\,\mathrm{molec\,cm^{-2}}$ for trace gas VCDs and $(0.3 - 8) \times 10^{10}\,\mathrm{molec\,cm^{-3}}$ for trace gas surface concentrations. These values compare to approximate average AOTs of 0.3, trace gas VCDs of $90 \times 10^{14}\,\mathrm{molec\,cm^{-2}}$ and trace gas surface concentrations of $11 \times 10^{10}\,\mathrm{molec\,cm^{-3}}$ observed over the campaign period. Note that profiles were retrieved from a common set of dSCDs and thus these discrepancies solely arise from the choice of the retrieval algorithm and detailed settings, that were not prescribed according to Sect. 2.1.3. Obvious source of discrepancies is the use of different techniques (OEM, PAR and ANA). Further, differences among the two PAR approaches are expected as they use different parametrizations. Note also that the compared algorithms have different priorities: the NASA/ Realtime algorithm for instance is optimised for computational performance rather than accuracy. Discrepancies among the different OEM algorithms are expected as they retrieve aerosol extinction either in logarithmic or linear space and since the exact implementation might differ (consider for instance the Thikonov regularisation approach used by BOREAS). Interestingly, discrepancies among participants using the same OEM algorithm are only about 50% smaller (regarding ASDevs of profiles as defined in Sect. 2.3) than the average discrepancies among all participants. This indicates that user defined retrieval settings that were not prescribed within this study (e.g. number of applied iteration steps in the optimisation process and RTM accuracy options) also have significant impact. An example appearing in this study are the differences between IUPHD and UTOR (both using HEIPRO) that were found to mainly be caused by differences in the number of applied iteration steps in the optimisation process of the aerosol inversions.

As discussed in more detail below and in Sect. 3.7, the discrepancies among the participants are of very similar order of magnitude as discrepancies that are induced when participants retrieve profiles from their own measured dSCDs. It is an important finding that, at least for CINDI-2, the choice of the algorithm/settings has similar impact on the profiling results as the inconsistencies in the dSCD acquisition.

For the comparison against supporting observations (see Fig. 22) RMSDs increase to $(0.02-0.2)$ against AOTs from the sun photometer, $(11-55)\times10^{14}\,\mathrm{molec\,cm^{-2}}$ against trace gas VCDs from the direct-sun DOAS and $(0.8-9)\times10^{10}\,\mathrm{molec\,cm^{-3}}$ against trace gas surface concentrations from the LP-DOAS. For Vis AOTs and trace gas surface concentrations discrepancies are mostly random (average Bias magnitude smaller than half the average RMSD) while for AOT UV and trace gas VCDs systematic deviations are dominant (compare Fig. 22). The average uncertainties of the supporting observations themselves are 0.022, $19\times10^{14}\,\mathrm{molec\,cm^{-2}}$ and $0.74\times10^{10}\,\mathrm{molec\,cm^{-3}}$, respectively, and can therefore be regarded as major RMSD contributors at least in cases where RMSD values are low. Errors in the median dSCDs used as the input for the retrievals are also likely to significantly contribute (see discussion on the own dSCD comparison below). Further, investigations on the spatio-temporal variability (see Sect. 2.3.3 and Supplement S6) indicate that a significant fraction of the RMSD observed between MAX-DOAS and supporting observations is caused by imperfect spatio-temporal overlap. For NO_2 surface concentrations the RMSD resulting from this could roughly be estimated to be around $3\times10^{10}\,\mathrm{molec\,cm^{-3}}$ (using strong simplifications though) which is indeed of the order of magnitude of the average RMSDs observed. Finally, simplified assumptions on the fixed RTM atmosphere were made (compare Sect. 2.1.3). While the choice of pressure and temperature profiles has little impact on the overall agreement with supporting observations ($<5\%$, see Supplement S7), the assumptions on the aerosol optical properties (Henyey-Greenstein approximation with constant single scattering albedo and asymmetry parameter over the whole campaign) are a likely source of error.

The consistency of Aerosol Vis and NO_2 Vis products (in particular the agreement among the participants) is typically worse in comparison to their UV counterparts by up to several ten percent. Only the agreement with the sun photometer AOT improves when going from the UV to the Vis spectral range. This might also be related to the reliability of the sun photometer AOTs τ_s: while in the Vis the MAX-DOAS retrieval wavelength $(477\,\mathrm{nm})$ is close to the lowest sun photometer wavelength channel $(440\,\mathrm{nm})$, in the UV extrapolation of τ_s down to $360\,\mathrm{nm}$ is required (see Sect. 2.2.1).

The presence of clouds strongly affects the agreement of aerosol retrieval results particularly in the visible spectral range. For AOTs the increase in average RMSD against the median is around $30\,\%$ in the UV and $80\,\%$ in the Vis while RMSDs against the sun photometer are degraded by $10\,\%$ and $130\,\%$, respectively. This is expected as i) high aerosol optical thicknesses at altitudes of low MAX-DOAS sensitivity make the results extremely susceptible to even small changes in the retrieval strategy and ii) the few sun photometer observations under cloudy conditions are likely recorded through local cloud holes and therefore not representative for MAX-DOAS measurements integrating horizontally over several kilometres. In contrast, the impact of clouds on average RMSDs for trace gas VCDs is $<15\,\%$. Surface concentration RMSDs against the median are degraded by around $25\,\%$, whereas average RMSDs to supporting observations even decrease.

It could be shown that, in the case of CINDI-2, the average impact of smoothing effects on the NO_2 surface concentration is negligible (Supplement S9). In contrast to that, smoothing has a strong impact on the agreement of MAX-DOAS observations with AOTs and probably HCHO VCDs from supporting observations (Sect. 2.3.2). In particular it was shown for the first time, that formerly observed systematic discrepancies between MAX-DOAS integrated aerosol profiles and sun photometer AOTs can be largely explained and compensated by considering biases arising from the reduced sensitivity of MAX-DOAS observations to higher altitudes and associated *a priori* assumptions (see Sect. 3.4).

For CINDI-2 data, there is no clear indication that an O_4 dSCD scaling is necessary. On the one hand for OEM algorithms the MAX-DOAS AOT is in good agreement with the sun photometer partial AOT and in contrast to Beirle et al. (2019), we find that a scaling factor of 0.8 is too small (Supplement S2) at least when applied to the whole campaign. On the other hand a less extreme scaling ($0.8 < SF < 1.0$) potentially removes remaining biases (see Fig. S3) and improves the agreement between forward model and reality (see Fig. S4). With the *a priori* settings applied in this study, O_4 scaling and PAC were found to have similar impact on the MAX-DOAS AOT results. Scaling might therefore be used to at least partly replace the PAC in the case of retrieval approaches that do not quantify their sensitivity or the assimilated *a priori* information. At last we think for this study the prescribed scaling factor of 1.0 is justified. Even though it might not be ideal, it is the most straightforward approach and yields reasonable and consistent results within the uncertainties introduced by other factors. To draw more concise conclusions, further studies as performed e.g. by Wagner et al. (2019) and Ortega et al. (2016) are necessary.

In most comparisons, RMSDs of individual participants against the MAX-DOAS median results (even when using the same algorithm) was of the order or larger than the uncertainties specified by the algorithms themselves (up to a factor of three for NO_2 Vis surface concentrations), indicating that the choice of the retrieval algorithm has severe impact on the results. It shows further, that the specified uncertainties (which typically take propagated measurement noise and smoothing errors into account but neglect other effects like model errors) are too optimistic as a measure for the MAX-DOAS retrieval accuracy and have to be regarded with care.

If the profiles are retrieved from the participant's individually measured dSCDs instead of using a common median dSCD dataset (see Sect. 3.7), the agreement of MAX-DOAS results with supporting observations (average RMSD) is degraded by very different amounts, depending on species and data product. Low impact ($\leq 10\,\%$) was found for Aerosol UV AOTs and NO_2 data products. For Aerosol UV AOTs and NO_2 UV VCDs even improvements were observed, hinting to potential systematic errors in the median dSCDs. A particularly strong degradation was observed for HCHO VCDs ($65\,\%$). Further, we estimated what fractions of the observed discrepancies among the MAX-DOAS participants are caused either by the use of different retrieval algorithms or by inconsistencies in the dSCD acquisition. In average the impact of both aspects is very similar: the effect of using own dSCDs can be estimated to contribute 40% (for AOTs), 85% (HCHO VCDs), 70% (HCHO surface concentrations), 50% (NO_2 VCDs), 40% (NO_2 UV surface concentrations) and 20% (NO_2 Vis surface concentrations) to the total variance introduced by both aspects. The high values for HCHO are expected, since according to Kreher et al. (2019) the acquisition of dSCDs was particular challenging and here and they varied widely among the participants.

We summarize our major findings as follows: besides the quality of the spectral data, the applied inversion strategy has significant impact on the accuracy of MAX-DOAS retrieval results. Nevertheless, partial AOTs, VCDs and surface concentrations can be retrieved with good accuracy, if algorithm, settings and quality filters are chosen carefully and ideally by experienced users. For the future, we therefore suggest to put focus on further harmonisation of MAX-DOAS retrievals, in particular with regard to their application by the broader scientific community.

For future campaign and comparison exercises, fixed model parameters (particularly aerosol optical properties) and prior constraints might be improved. Further we suggest putting enhanced focus on the coordinated operation of all (not only MAX-

DOAS) instruments and to incorporate techniques with more appropriate spatial kernels, e.g. limb DOAS measurements from unmanned aerial vehicles, to reduce the spatio-temporal mismatch between different observations.

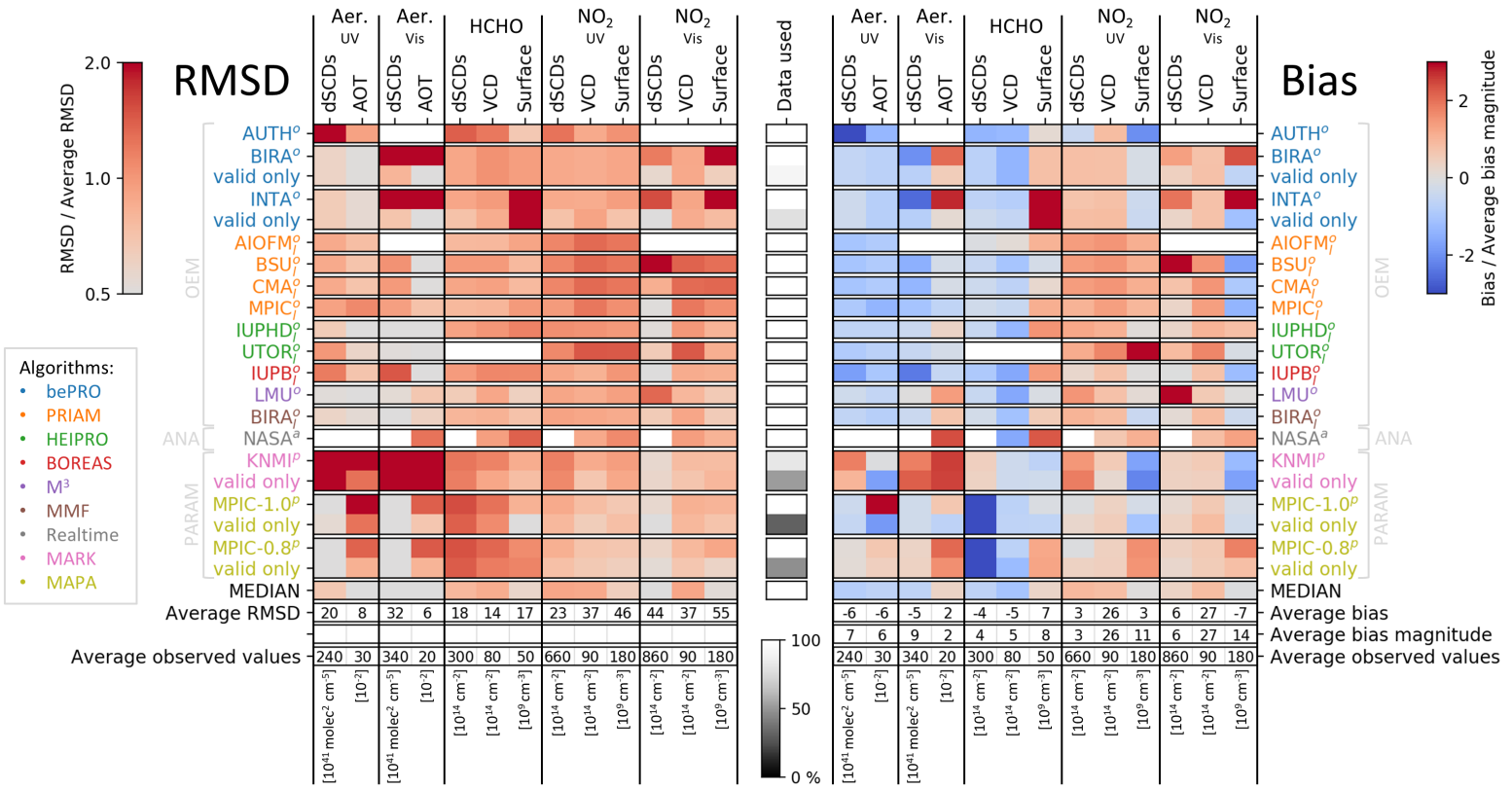

Figure 22. Summary of the comparisons in Sect. 3 for clear-sky conditions. Left panel shows RMSD, right panel shows Bias. Average values of RMSD (Bias) define the colour scale of each column of the left (right) panel as indicated by the color bars on the top left (top right) of the figure. Values of AOT, VCD and surface concentration are given with respect to the corresponding supporting observations (sun photometer, direct-sun DOAS and LP-DOAS). White spaces indicate no data. Average observed values (bottom row) are rounded campaign averages of the supporting observations. Average Bias and Average Bias magnitude values (third last and second last row of right panel) represent the averages over the signed and the absolute Bias values, respectively. The "data used"-column in the center indicates which fraction of the maximum number (170) of available profiles has been used. Participants who submitted flags are represented by two rows: one considering all data and one using only those flagged as valid ("valid only").

Author contributions. JLT performed the comparison and the associated investigations as described in the paper and wrote the first draft. UF was involved in the planning of the campaign and the profiling activities, operated the IUPHD instrument, evaluated its data, supervised the comparison activities and contributed in scientific discussions and the manuscript revision. FH was involved in the planning of the campaign and the profiling activities, retrieved profiles for BIRA and contributed in scientific discussions and the manuscript revision. FH, GP, MVR, AA, AP, AR, TW, KK, UF, JL designed, planned and organized the CINDI-2 campaign. AL/JX/PX, AP, CF/CH/AM/FT/GP/MVR, CZ/KLC/NH/ZW, EP/FW/TB, ES, IB, JJ/JM, KB/XZ, KLC, MY/OPu, SD and TD/AB prepared and operated the MAX-DOAS instrument(s) of AIOFM, KNMI, BIRA, USTC, IUPB, Pandora, BSU, CMA, UTOR, DLR, INTA, MPIC and AUTH, respectively. AL/JX/PX, AP/TV, CA, CF/MVR, CG/FH, CX/HL/KLC, EP/TB/ FW/ AR, ES, IB, JJ/JM, KB/XZ, KLC, LGM/MY/OPu, MMF, SBei, SD, TD/AB, YW and ZW evaluated the MAX-DOAS data for AIOFM, KNMI/MARK, LMU, BIRA, BIRA/bePRO, USTC, IUPB, NASA/Realtime, BSU, CMA, UTOR, DLR/M3, INTA, BIRA/MMF, MPIC/MAPA, MPIC, AUTH, MPIC/PriAM and DLR/bePRO. AB, AR, CL, KS, MWe, NH and TW supervised the activities of AUTH, Bremen, USTC, UTOR, LMU, DLR and MPIC, respectively. KK as the campaign referee was involved in the actual running of the campaign and the data evaluation up to dSCDs. TW and JK planned and performed the common MAX-DOAS pointing calibration. NH coordinated the cooperation between DLR and USTC. Installation, operation and data evaluation of in-situ NO_x instrumentation was performed by AF/AH (in-situ profile instrumentation in the tower), AM/FT (CAPS) and JL (ICAD/CE-DOAS). BH calibrated and operated the CIMEL sun photometer that is part of AERONET. DSw, LGa, RVH and SBer operated the NO_2 lidar and processed its data into NO_2 profiles. SS installed, operated and evaluated the data of the LP-DOAS instrument. DSZ, MA and MDH operated and evaluated the data of the NO_2 radiosondes. AC and MT provided and installed the Pandora instruments from which NO_2 direct-sun and NASA/Realtime profiling data were deduced. AA, AF, AH, AR, ES, JH, KB, KLC, MMF, MWi, SBei, SS, TB, TW, UP and YW contributed to the scientific discussion and interpretation. AM, AR, CF, CH, FT, GP, JH, JV, MMF, MVR, MWi, SBei, SS, TB, TW, UP and YW revised and contributed to the manuscript. All authors read and approved the submitted version.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We gratefully acknowledge the KNMI staff at Cabauw for their excellent technical and infrastructural support during the campaign. Further we acknowledge EARLINET, CESAR and AERONET for providing data for this study. We acknowledge the authors of the QDOAS package (Caroline Fayt, Michel van Roozendael, Thomas Dankaert). Pandora instrument deployment was supported by Luftblick through ESA Pandonia Project and NASA Pandora Project at Goddard Space Flight Center under NASA Headquarters' Tropospheric Composition Program. We like to thank Airyx GmbH and Dr. Denis Pöhler for supporting measurements with the Airyx GmbH / EnviMeS MAX-DOAS and in-situ instruments. We kindly acknowledge further CINDI-2 participants, who indirectly contributed to the median dSCD dataset and a successful campaign: Abishek Mishra Kumar, Alexander Borovski, Alfonso Saiz-Lopez, Andre Seyler, Andrea Pazmino, Anja Schönhardt, Ermioni Dimitropoulou, Fahim Khokhar, Henning Finkenzeller, Hitoshi Irie, Jeron van Gent, Junaid Khayyam Butt, Manuel Pinharanda, Mareike Ostendorf, Martin Tiefengraber, Mihalis Vrekoussis, Monica Anguas, Monica Navarro-Comas, Moritz Müller, Nader Abuhassan, Nuria Benavent, Paul Johnston, Rainer Volkamer, Richard Querel, Shanshan Wang, Stefan F. Schreier, Syedul Hoque, Theodore K. Koenig, Vinayak Sinha, Vinod Kumar, Xin Tian. We gratefully acknowledge the efforts taken by the two anonymous reviewers and the editor (Rainer Volkamer) to read and revise this extensive manuscript.

Funding for this study was provided by ESA through the CINDI-2 (ESA Contract No. 4000118533/16/I-Sbo) and FRM4DOAS (ESA Contract No. 4000118181/16/I-EF) projects and partly within the EU 7th Framework Programme QA4ECV project (Grant Agreement no. 607405). The AIOFM group acknowledges the support by the NSFC under project No. 41530644. The participation of the University of Toronto team was supported by the Canadian Space Agency (through the AVATARS project) and the Natural Sciences and Engineering Research Council of Canada (through the PAHA project). The instrument was funded by the Canada Foundation for Innovation and is usually operated at the Polar Environment Atmospheric Research Laboratory (PEARL) by the Canadian Network for the Detection of Atmospheric Change (CANDAC). The activities of the IUP Heidelberg were supported by the DFG project RAPSODI (grant No. PL 193/17-1). INTA acknowledges support from the National funding projects HELADO (CTM2013-41311-P) and AVATAR (CGL2014-55230-R). CMA group acknowledges the support by the NSFC under project Nos. 41805027. The participation of the LMU team was made possible by the DFG Major Research Instrumentation Programme (INST 86/1499-1 FUGG). KLC has received funding from the Marie Curie Initial Training Network of the European 7th Framework Programme (Grant No. 607905) and the European Union's Horizon 2020 research and innovation programme (Grant No. 654109). Support was received from ACTRIS-2 H2020 Grant Agreement Nr. 654109. The CINDI-2 campaign received funding from the Dutch Space Office (NSO).

References

Apituley, A., Wilson, K., Potma, C., Volten, H., and de Graaf, M.: Performance Assessment and Application of Caeli–A highperformance Raman lidar for diurnal profiling of Water Vapour, Aerosols and Clouds, in: Proceedings of the 8th International Symposium on Tropospheric Profiling, pp. 19–23, S06-O10-1-4, Delft/KNMI/RIVM Delft, Netherlands, 2009.

Apituley, A., Hendrick, F., van Roozendael, M., Richter, A., Wagner, T., Frieß, U., Kreher, K., and et al.: Second Cabauw Intercomparison of Nitrogen Dioxide Measuring Instruments (CINDI-2) – Campaign Overview, Atm. Meas. Tech., 2020 in prep.

Beirle, S., Dörner, S., Donner, S., Remmers, J., Wang, Y., and Wagner, T.: The Mainz profile algorithm (MAPA), Atmospheric Measurement Techniques, 12, 1785–1806, https://doi.org/10.5194/amt-12-1785-2019, https://www.atmos-meas-tech.net/12/1785/2019/, 2019.

Berkhout, S., van der Hoff, R., Swart, D., and Bergwerff, J.: The RIVM mobile lidar—Design and operation of a versatile system for measuring atmospheric trace gases, in: Reviewed and Revised Papers of the 23rd International Laser Radar Conference (ILRC), 2006.

Bösch, T., Rozanov, V., Richter, A., Peters, E., Rozanov, A., Wittrock, F., Merlaud, A., Lampel, J., Schmitt, S., de Haij, M., Berkhout, S., Henzing, B., Apituley, A., den Hoed, M., Vonk, J., Tiefengraber, M., Müller, M., and Burrows, J. P.: BOREAS – a new MAX-DOAS profile retrieval algorithm for aerosols and trace gases, Atmospheric Measurement Techniques, 11, 6833–6859, https://doi.org/10.5194/amt-11-6833-2018, https://www.atmos-meas-tech.net/11/6833/2018/, 2018.

Bösenberg, J., Matthias, V., Amodeo, A., Amoiridis, V., Ansmann, A., Baldasano, J. M., Balin, I., Balis, D., Böckmann, C., Boselli, A., Carlsson, G., Chaikovsky, A., Chourdakis, G., Comeron, A., Tomasi, F. D., Eixmann, R., Freudenthaler, V., Giehl, H., Grigorov, I., Hagard, A., Iarlori, M., Kirsche, A., Kolarov, G., Komguem, L., S. Kreipl, W. K., Larcheveque, G., Linné, H., Matthey, R., Mattis, I., Mekler, A., Mironova, I., Mitev, V., Mona, L., Müller, D., Music, S., Nickovic, S., Pandolfi, M., Papayannis, A., Pappalardo, G., Pelon, J., Perez, C., Perrone, R., Persson, R., Resendes, D. P., Rizi, V., Rocadenbosch, F., Rodrigues, J. A., Sauvage, L., Schneidenbach, L., Schumacher, R., Shcherbakov, V., Simeonov, V., Sobolewski, P., Spinelli, N., Stachlewska, I., Stoyanov, D., Trickl, T., Tsaknakis, G., Vaughan, G., Wandinger, U., Wang, X., Wiegner, M., Zavrtanik, M., and Zerefos, C.: EARLINET: A European Aerosol Research Lidar Network to establish an aerosol climatology, Report 348, ISSN 0937-1060, 192 pp., Max-Planck-Institut für Meteorologie, 2003.

CESAR: Cabauw Experimental Site for Atmospheric Research Homepage, http://www.cesar-observatory.nl/index.php?pageID=1002, 2018.

Chan, K. L., Wiegner, M., Wenig, M., and Pöhler, D.: Observations of tropospheric aerosols and NO_2 in Hong Kong over 5 years using ground based MAX-DOAS, Science of The Total Environment, 619, 1545–1556, https://doi.org/10.1016/j.scitotenv.2017.10.153, 2017.

Chan, K. L., Wang, Z., Ding, A., Heue, K.-P., Shen, Y., Wang, J., Zhang, F., Hao, N., and Wenig, M.: MAX-DOAS measurements of tropospheric NO_2 and HCHO in Nanjing and the comparison to OMI observations, Atmospheric Chemistry and Physics Discussions, 2019, 1–25, https://doi.org/10.5194/acp-2018-1266, https://www.atmos-chem-phys-discuss.net/acp-2018-1266/, 2019.

Chan, K. L., Wiegner, M., van Geffen, J., De Smedt, I., Alberti, C., Cheng, Z., Ye, S., and Wenig, M.: MAX-DOAS measurements of tropospheric NO_2 and HCHO in Munich and the comparison to OMI and TROPOMI satellite observations, Atmospheric Measurement Techniques, 13, 4499–4520, https://doi.org/10.5194/amt-13-4499-2020, https://amt.copernicus.org/articles/13/4499/2020/, 2020.

Clémer, K., Van Roozendael, M., Fayt, C., Hendrick, F., Hermans, C., Pinardi, G., Spurr, R., Wang, P., and De Mazière, M.: Multiple wavelength retrieval of tropospheric aerosol optical properties from MAXDOAS measurements in Beijing, Atmospheric Measurement Techniques, 3, 863–878, https://doi.org/10.5194/amt-3-863-2010, https://www.atmos-meas-tech.net/3/863/2010/, 2010.

Donner, S., Kuhn, J., Van Roozendael, M., Bais, A., Beirle, S., Bösch, T., Bognar, K., Bruchkousky, I., Chan, K. L., Drosoglou, T., Fayt, C., Frieß, U., Hendrick, F., Hermans, C., Jin, J., Li, A., Ma, J., Peters, E., Pinardi, G., Richter, A., Schreier, S. F., Seyler, A., Strong, K., Tirpitz, J.-L., Wang, Y., Xie, P., Xu, J., Zhao, X., and Wagner, T.: Evaluating different methods for elevation calibration of MAX-DOAS

instruments during the CINDI-2 campaign, Atmospheric Measurement Techniques Discussions, 2019, 1–51, https://doi.org/10.5194/amt-2019-115, https://www.atmos-meas-tech-discuss.net/amt-2019-115/, 2019.

Esri, EsriNL, Rijkswaterstaat, Intermap, NASA, NGA, Kadaster, U. ., Esri, HERE, Garmin, P, I., and METI: arcGIS World Topo Map, 2018.

Friedrich, M. M., Rivera, C., Stremme, W., Ojeda, Z., Arellano, J., Bezanilla, A., García-Reynoso, J. A., and Grutter, M.: NO_2 vertical profiles and column densities from MAX-DOAS measurements in Mexico City, Atmospheric Measurement Techniques Discussions, 2018, 1–34, https://doi.org/10.5194/amt-2018-358, https://www.atmos-meas-tech-discuss.net/amt-2018-358/, 2019.

Frieß, U., Monks, P., Remedios, J., Rozanov, A., Sinreich, R., Wagner, T., and Platt, U.: MAX-DOAS O4 measurements: A new technique to derive information on atmospheric aerosols: 2. Modeling studies, Journal of Geophysical Research: Atmospheres, 111, 2006.

Frieß, U., Klein Baltink, H., Beirle, S., Clémer, K., Hendrick, F., Henzing, B., Irie, H., de Leeuw, G., Li, A., Moerman, M. M., van Roozendael, M., Shaiganfar, R., Wagner, T., Wang, Y., Xie, P., Yilmaz, S., and Zieger, P.: Intercomparison of aerosol extinction profiles retrieved from MAX-DOAS measurements, Atmospheric Measurement Techniques, 9, 3205–3222, https://doi.org/10.5194/amt-9-3205-2016, https://www.atmos-meas-tech.net/9/3205/2016/, 2016.

Frieß, U., Beirle, S., Alvarado Bonilla, L., Bösch, T., Friedrich, M. M., Hendrick, F., Piters, A., Richter, A., van Roozendael, M., Rozanov, V. V., Spinei, E., Tirpitz, J.-L., Vlemmix, T., Wagner, T., and Wang, Y.: Intercomparison of MAX-DOAS vertical profile retrieval algorithms: studies using synthetic data, Atmospheric Measurement Techniques, 12, 2155–2181, https://doi.org/10.5194/amt-12-2155-2019, https://www.atmos-meas-tech.net/12/2155/2019/, 2019.

Heckel, A., Richter, A., Tarsu, T., Wittrock, F., Hak, C., Pundt, I., Junkermann, W., and Burrows, J. P.: MAX-DOAS measurements of formaldehyde in the Po-Valley, Atmospheric Chemistry and Physics, 5, 909–918, https://doi.org/10.5194/acp-5-909-2005, https://www.atmos-chem-phys.net/5/909/2005/, 2005.

Hendrick, F., Müller, J.-F., Clémer, K., Wang, P., De Mazière, M., Fayt, C., Gielen, C., Hermans, C., Ma, J. Z., Pinardi, G., Stavrakou, T., Vlemmix, T., and Van Roozendael, M.: Four years of ground-based MAX-DOAS observations of HONO and NO_2 in the Beijing area, Atmospheric Chemistry and Physics, 14, 765–781, https://doi.org/10.5194/acp-14-765-2014, https://www.atmos-chem-phys.net/14/765/2014/, 2014.

Herman, J., Cede, A., Spinei, E., Mount, G., Tzortziou, M., and Abuhassan, N.: NO_2 column amounts from ground-based Pandora and MFDOAS spectrometers using the direct-Sun DOAS technique: Intercomparisons and application to OMI validation, Journal of Geophysical Research: Atmospheres, 114, 2009.

Hönninger, G. and Platt, U.: Observations of BrO and its vertical distribution during surface ozone depletion at Alert, Atmospheric Environment, 36, 2481 – 2489, https://doi.org/https://doi.org/10.1016/S1352-2310(02)00104-8, http://www.sciencedirect.com/science/article/pii/S1352231002001048, air/Snow/Ice Interactions in the Arctic: Results from ALERT 2000 and SUMMIT 2000, 2002.

Holben, B. N., Eck, T. F., Slutsker, I., Tanre, D., Buis, J., Setzer, A., Vermote, E., Reagan, J., Kaufman, Y., Nakajima, T., et al.: AERONET—A federated instrument network and data archive for aerosol characterization, Remote sensing of environment, 66, 1–16, 1998.

Hönninger, G., von Friedeburg, C., and Platt, U.: Multi axis differential optical absorption spectroscopy (MAX-DOAS), Atmospheric Chemistry and Physics, 4, 231–254, https://doi.org/10.5194/acp-4-231-2004, https://www.atmos-chem-phys.net/4/231/2004/, 2004.

Horbanski, M., Pöhler, D., Lampel, J., and Platt, U.: The ICAD (iterative cavity-enhanced DOAS) method, Atmospheric Measurement Techniques, 12, 3365–3381, https://doi.org/10.5194/amt-12-3365-2019, https://www.atmos-meas-tech.net/12/3365/2019/, 2019.

Irie, H., Kanaya, Y., Akimoto, H., Iwabuchi, H., Shimizu, A., and Aoki, K.: First retrieval of tropospheric aerosol profiles using MAX-DOAS and comparison with lidar and sky radiometer measurements, Atmospheric Chemistry and Physics, 8, 341–350, https://doi.org/10.5194/acp-8-341-2008, https://www.atmos-chem-phys.net/8/341/2008/, 2008.

Irie, H., Takashima, H., Kanaya, Y., Boersma, K. F., Gast, L., Wittrock, F., Brunner, D., Zhou, Y., and Van Roozendael, M.: Eight-component retrievals from ground-based MAX-DOAS observations, Atmospheric Measurement Techniques, 4, 1027–1044, https://doi.org/10.5194/amt-4-1027-2011, https://www.atmos-meas-tech.net/4/1027/2011/, 2011.

Kaskaoutis, D. G. and Kambezidis, H. D.: Investigation into the wavelength dependence of the aerosol optical depth in the Athens area, Quarterly Journal of the Royal Meteorological Society, 132, 2217–2234, https://doi.org/10.1256/qj.05.183, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/qj.05.183, 2006.

Kebabian, P. L., Herndon, S. C., and Freedman, A.: Detection of Nitrogen Dioxide by Cavity Attenuated Phase Shift Spectroscopy, Analytical Chemistry, 77, 724–728, https://doi.org/10.1021/ac048715y, https://doi.org/10.1021/ac048715y, pMID: 15649079, 2005.

Kreher, K., Van Roozendael, M., Hendrick, F., Apituley, A., Dimitropoulou, E., Frieß, U., Richter, A., Wagner, T., Abuhassan, N., Ang, L., Anguas, M., Bais, A., Benavent, N., Bösch, T., Bognar, K., Borovski, A., Bruchkouski, I., Cede, A., Chan, K. L., Donner, S., Drosoglou, T., Fayt, C., Finkenzeller, H., Garcia-Nieto, D., Gielen, C., Gómez-Martín, L., Hao, N., Herman, J. R., Hermans, C., Hoque, S., Irie, H., Jin, J., Johnston, P., Khayyam Butt, J., Khokhar, F., Koenig, T. K., Kuhn, J., Kumar, V., Lampel, J., Liu, C., Ma, J., Merlaud, A., Mishra, A. K., Müller, M., Navarro-Comas, M., Ostendorf, M., Pazmino, A., Peters, E., Pinardi, G., Pinharanda, M., Piters, A., Platt, U., Postylyakov, O., Prados-Roman, C., Puentedura, O., Querel, R., Saiz-Lopez, A., Schönhardt, A., Schreier, S. F., Seyler, A., Sinha, V., Spinei, E., Strong, K., Tack, F., Tian, X., Tiefengraber, M., Tirpitz, J.-L., van Gent, J., Volkamer, R., Vrekoussis, M., Wang, S., Wang, Z., Wenig, M., Wittrock, F., Xie, P. H., Xu, J., Yela, M., Zhang, C., and Zhao, X.: Intercomparison of NO_2, O_4, O_3 and HCHO slant column measurements by MAX-DOAS and zenith-sky UV-Visible spectrometers during the CINDI-2 campaign, Atmospheric Measurement Techniques Discussions, 2019, 1–58, https://doi.org/10.5194/amt-2019-157, https://www.atmos-meas-tech-discuss.net/amt-2019-157/, 2019.

Meller, R. and Moortgat, G. K.: Temperature dependence of the absorption cross sections of formaldehyde between 223 and 323 K in the wavelength range 225–375 nm, Journal of Geophysical Research: Atmospheres, 105, 7089–7101, https://doi.org/10.1029/1999JD901074, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/1999JD901074, 2000.

Merten, A., Tschritter, J., and Platt, U.: Design of differential optical absorption spectroscopy long-path telescopes based on fiber optics, Applied optics, 50, 738–754, 2011.

Nasse, J.-M., Eger, P. G., Pöhler, D., Schmitt, S., Frieß, U., and Platt, U.: Recent improvements of Long-Path DOAS measurements: impact on accuracy and stability of short-term and automated long-term observations, Atmospheric Measurement Techniques Discussions, 2019, 1–36, https://doi.org/10.5194/amt-2019-69, https://www.atmos-meas-tech-discuss.net/amt-2019-69/, 2019.

Ortega, I., Berg, L. K., Ferrare, R. A., Hair, J. W., Hostetler, C. A., and Volkamer, R.: Elevated aerosol layers modify the O2–O2 absorption measured by ground-based MAX-DOAS, Journal of Quantitative Spectroscopy and Radiative Transfer, 176, 34 – 49, https://doi.org/https://doi.org/10.1016/j.jqsrt.2016.02.021, http://www.sciencedirect.com/science/article/pii/S0022407315301746, 2016.

Pappalardo, G., Amodeo, A., Apituley, A., Comeron, A., Freudenthaler, V., Linné, H., Ansmann, A., Bösenberg, J., D'Amico, G., Mattis, I., Mona, L., Wandinger, U., Amiridis, V., Alados-Arboledas, L., Nicolae, D., and Wiegner, M.: EARLINET: towards an advanced sustainable European aerosol lidar network, Atmospheric Measurement Techniques, 7, 2389–2409, https://doi.org/10.5194/amt-7-2389-2014, https://www.atmos-meas-tech.net/7/2389/2014/, 2014.

Pikelnaya, O., Hurlock, S. C., Trick, S., and Stutz, J.: Intercomparison of multiaxis and long-path differential optical absorption spectroscopy measurements in the marine boundary layer, Journal of Geophysical Research: Atmospheres, 112, https://doi.org/10.1029/2006JD007727, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2006JD007727, 2007.

Pinardi, G., Van Roozendael, M., Abuhassan, N., Adams, C., Cede, A., Clémer, K., Fayt, C., Frieß, U., Gil, M., Herman, J., et al.: MAX-DOAS formaldehyde slant column measurements during CINDI: intercomparison and analysis improvement., Atmospheric Measurement Techniques, 6, 2013.

Platt, U. and Stutz, J.: Differential Optical Absorption Spectroscopy, vol. 1, Springer Berlin Heidelberg, https://doi.org/10.1007/978-3-540-75776-4, 2008.

Platt, U., Meinen, J., Pöhler, D., and Leisner, T.: Broadband cavity enhanced differential optical absorption spectroscopy (CE-DOAS)–applicability and corrections, Atmospheric Measurement Techniques, 2, 713–723, 2009.

Pöhler, D., Vogel, L., Frieß, U., and Platt, U.: Observation of halogen species in the Amundsen Gulf, Arctic, by active long-path differential optical absorption spectroscopy, Proceedings of the National Academy of Sciences, 107, 6582–6587, https://doi.org/10.1073/pnas.0912231107, https://www.pnas.org/content/107/15/6582, 2010.

Rodgers, C. D.: Inverse methods for atmospheric sounding : theory and practice, World Scientific Publishing, 2000.

Rodgers, C. D. and Connor, B. J.: Intercomparison of remote sounding instruments, Journal of Geophysical Research: Atmospheres, 108, https://doi.org/10.1029/2002JD002299, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002JD002299, 2003.

Sluis, W. W., Allaart, M. A. F., Piters, A. J. M., and Gast, L. F. L.: The development of a nitrogen dioxide sonde, Atmospheric Measurement Techniques, 3, 1753–1762, https://doi.org/10.5194/amt-3-1753-2010, http://www.atmos-meas-tech.net/3/1753/2010/, 2010.

Smirnov, A., Holben, B., Eck, T., Dubovik, O., and Slutsker, I.: Cloud-Screening and Quality Control Algorithms for the AERONET Database, Remote Sensing of Environment, 73, 337 – 349, https://doi.org/https://doi.org/10.1016/S0034-4257(00)00109-7, http://www.sciencedirect.com/science/article/pii/S0034425700001097, 2000.

Spinei, E., Cede, A., Swartz, W. H., Herman, J., and Mount, G. H.: The use of NO_2 absorption cross section temperature sensitivity to derive NO_2 profile temperature and stratospheric–tropospheric column partitioning from visible direct-sun DOAS measurements, Atmospheric Measurement Techniques, 7, 4299–4316, https://doi.org/10.5194/amt-7-4299-2014, 2014.

Vandaele, A., Hermans, C., Simon, P., Carleer, M., Colin, R., Fally, S., Mérienne, M., Jenouvrier, A., and Coquart, B.: Measurements of the NO2 absorption cross-section from 42 000 cm^{-1} to 10 000 cm^{-1} (238–1000 nm) at 220 K and 294 K, Journal of Quantitative Spectroscopy and Radiative Transfer, 59, 171 – 184, https://doi.org/https://doi.org/10.1016/S0022-4073(97)00168-4, http://www.sciencedirect.com/science/article/pii/S0022407397001684, atmospheric Spectroscopy Applications 96, 1998.

Vlemmix, T., Piters, A. J. M., Berkhout, A. J. C., Gast, L. F. L., Wang, P., and Levelt, P. F.: Ability of the MAX-DOAS method to derive profile information for NO_2: can the boundary layer and free troposphere be separated?, Atmospheric Measurement Techniques, 4, 2659–2684, https://doi.org/10.5194/amt-4-2659-2011, https://www.atmos-meas-tech.net/4/2659/2011/, 2011.

Vlemmix, T., Eskes, H. J., Piters, A. J. M., Schaap, M., Sauter, F. J., Kelder, H., and Levelt, P. F.: MAX-DOAS tropospheric nitrogen dioxide column measurements compared with the Lotos-Euros air quality model, Atmospheric Chemistry and Physics, 15, 1313–1330, https://doi.org/10.5194/acp-15-1313-2015, https://www.atmos-chem-phys.net/15/1313/2015/, 2015a.

Vlemmix, T., Hendrick, F., Pinardi, G., De Smedt, I., Fayt, C., Hermans, C., Piters, A., Wang, P., Levelt, P., and Van Roozendael, M.: MAX-DOAS observations of aerosols, formaldehyde and nitrogen dioxide in the Beijing area: comparison of two profile retrieval approaches, Atmospheric Measurement Techniques, 8, 941–963, https://doi.org/10.5194/amt-8-941-2015, https://www.atmos-meas-tech.net/8/941/2015/, 2015b.

Wagner, T., Dix, B. v., Friedeburg, C. v., Frieß, U., Sanghavi, S., Sinreich, R., and Platt, U.: MAX-DOAS O_4 measurements: A new technique to derive information on atmospheric aerosols—Principles and information content, Journal of Geophysical Research: Atmospheres, 109, 2004.

Wagner, T., Deutschmann, T., and Platt, U.: Determination of aerosol properties from MAX-DOAS observations of the Ring effect, Atmospheric Measurement Techniques, 2, 495–512, https://doi.org/10.5194/amt-2-495-2009, https://www.atmos-meas-tech.net/2/495/2009/, 2009.

Wagner, T., Beirle, S., Brauers, T., Deutschmann, T., Frieß, U., Hak, C., Halla, J. D., Heue, K. P., Junkermann, W., Li, X., Platt, U., and Pundt-Gruber, I.: Inversion of tropospheric profiles of aerosol extinction and HCHO and NO_2 mixing ratios from MAX-DOAS observations in Milano during the summer of 2003 and comparison with independent data sets, Atmospheric Measurement Techniques, 4, 2685–2715, https://doi.org/10.5194/amt-4-2685-2011, https://www.atmos-meas-tech.net/4/2685/2011/, 2011.

Wagner, T., Apituley, A., Beirle, S., Dörner, S., Friess, U., Remmers, J., and Shaiganfar, R.: Cloud detection and classification based on MAX-DOAS observations, Atmospheric Measurement Techniques, 7, 1289–1320, 2014.

Wagner, T., Beirle, S., Benavent, N., Bösch, T., Chan, K. L., Donner, S., Dörner, S., Fayt, C., Frieß, U., García-Nieto, D., Gielen, C., González-Bartolome, D., Gomez, L., Hendrick, F., Henzing, B., Jin, J. L., Lampel, J., Ma, J., Mies, K., Navarro, M., Peters, E., Pinardi, G., Puentedura, O., Puķīte, J., Remmers, J., Richter, A., Saiz-Lopez, A., Shaiganfar, R., Sihler, H., Van Roozendael, M., Wang, Y., and Yela, M.: Is a scaling factor required to obtain closure between measured and modelled atmospheric O_4 absorptions? An assessment of uncertainties of measurements and radiative transfer simulations for 2 selected days during the MAD-CAT campaign, Atmospheric Measurement Techniques, 12, 2745–2817, https://doi.org/10.5194/amt-12-2745-2019, https://www.atmos-meas-tech.net/12/2745/2019/, 2019.

Wang, Y., Li, A., Xie, P.-H., Chen, H., Mou, F.-S., Xu, J., Wu, F.-C., Zeng, Y., Liu, J.-G., and Liu, W.-Q.: Measuring tropospheric vertical distribution and vertical column density of NO_2 by multi-axis differential optical absorption spectroscopy, Acta Physica Sinica, 62, 200705, https://doi.org/10.7498/aps.62.200705, http://wulixb.iphy.ac.cn/EN/abstract/article_56201.shtml, 2013a.

Wang, Y., Li, A., Xie, P.-H., Chen, H., Xu, J., Wu, F.-C., Liu, J.-G., and Liu, W.-Q.: Retrieving vertical profile of aerosol extinction by multi-axis differential optical absorption spectroscopy, Acta Physica Sinica, 62, 180705, https://doi.org/10.7498/aps.62.180705, http://wulixb.iphy.ac.cn/EN/abstract/article_55526.shtml, 2013b.

Wang, Y., Penning de Vries, M., Xie, P. H., Beirle, S., Dörner, S., Remmers, J., Li, A., and Wagner, T.: Cloud and aerosol classification for 2.5 years of MAX-DOAS observations in Wuxi (China) and comparison to independent data sets, Atmospheric Measurement Techniques, 8, 5133–5156, https://doi.org/10.5194/amt-8-5133-2015, https://www.atmos-meas-tech.net/8/5133/2015/, 2015.

Wang, Y., Lampel, J., Xie, P., Beirle, S., Li, A., Wu, D., and Wagner, T.: Ground-based MAX-DOAS observations of tropospheric aerosols, NO_2, SO_2 and HCHO in Wuxi, China, from 2011 to 2014, Atmospheric Chemistry and Physics, 17, 2189–2215, https://doi.org/10.5194/acp-17-2189-2017, https://www.atmos-chem-phys.net/17/2189/2017/, 2017.

Wang, Y., Puķīte, J., Wagner, T., Donner, S., Beirle, S., Hilboll, A., Vrekoussis, M., Richter, A., Apituley, A., Piters, A., Allaart, M., Eskes, H., Frumau, A., Van Roozendael, M., Lampel, J., Platt, U., Schmitt, S., Swart, D., and Vonk, J.: Vertical Profiles of Tropospheric Ozone From MAX-DOAS Measurements During the CINDI-2 Campaign: Part 1—Development of a New Retrieval Algorithm, Journal of Geophysical Research: Atmospheres, 123, 10,637–10,670, https://doi.org/10.1029/2018JD028647, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018JD028647, 2018.

Wang, Y., Apituley, A., Bais, A., Beirle, S., Benavent, N., Borovski, A., Bruchkouski, I., Chan, K. L., Donner, S., Drosoglou, T., Finkenzeller, H., Friedrich, M. M., Frieß, U., Garcia-Nieto, D., Gómez-Martín, L., Hendrick, F., Hilboll, A., Jin, J., Johnston, P., Koenig, T. K., Kreher, K., Kumar, V., Kyuberis, A., Lampel, J., Liu, C., Liu, H., Ma, J., Polyansky, O. L., Postylyakov, O., Querel, R., Saiz-Lopez, A., Schmitt, S., Tian, X., Tirpitz, J.-L., Van Roozendael, M., Volkamer, R., Wang, Z., Xie, P., Xing, C., Xu, J., Yela, M., Zhang, C., and Wagner, T.: Inter-comparison of MAX-DOAS measurements of tropospheric HONO slant column densities and vertical profiles

during the CINDI-2 Campaign, Atmospheric Measurement Techniques Discussions, 2020, 1–44, https://doi.org/10.5194/amt-2019-464, https://www.atmos-meas-tech-discuss.net/amt-2019-464/, 2020.

Wiegner, M. and Geiß, A.: Aerosol profiling with the Jenoptik ceilometer CHM15kx, Atmospheric Measurement Techniques, 5, 1953–1964, https://doi.org/10.5194/amt-5-1953-2012, https://www.atmos-meas-tech.net/5/1953/2012/, 2012.

Yilmaz, S.: Retrieval of atmospheric aerosol and trace gas vertical profiles using multi-axis differential optical absorption spectroscopy, Ph.D. thesis, Heidelberg, Univ., Diss., 2012, http://archiv.ub.uni-heidelberg.de/volltextserver/volltexte/2012/13128, 2012.

Zieger, P., Weingartner, E., Henzing, J., Moerman, M., de Leeuw, G., Mikkilä, J., Ehn, M., Petäjä, T., Clémer, K., van Roozendael, M., Yilmaz, S., Frieß, U., Irie, H., Wagner, T., Shaiganfar, R., Beirle, S., Apituley, A., Wilson, K., and Baltensperger, U.: Comparison of ambient aerosol extinction coefficients obtained from in-situ, MAX-DOAS and LIDAR measurements at Cabauw, Atmospheric Chemistry and Physics, 11, 2603–2624, https://doi.org/10.5194/acp-11-2603-2011, https://www.atmos-chem-phys.net/11/2603/2011/, 2011.