Please note that the required information has already been uploaded as author comments to the public discussion forum at https://www.atmos-meas-tech-discuss.net/amt-2019-456/#discussion in 4 separate files:

- **Answers to referee 1**
  To be found in: "Reply to anonymous referee #1" → supplement → ANSWERS_REVIEWER_1.pdf

- **Answers to referee 2**
  To be found in: "Reply to anonymous referee #2" → supplement → ANSWERS_REVIEWER_2.pdf

- **Marked-up version of manuscript**
  Either "Reply to anonymous referee #1" or "Reply to anonymous referee #2" → supplement → Latexdiff_Manuscript.pdf

- **Marked-up version of supplements**
  Either "Reply to anonymous referee #1" or "Reply to anonymous referee #2" → supplement → Latexdiff_Supplement.pdf

This document is a merged version of the files listed above.

Answers to anonymous referee 1

General information

First of all, we would like to gratefully acknowledge the efforts taken by the reviewers to read and revise this extensive manuscript. We are convinced that their comments helped to significantly improve the manuscript regarding comprehensibility and completeness, particularly in the conclusions.

Document formatting

- The reviewer's comments are reprinted here in bold face.
- Our answers are given in regular font
- Explicit changes made in the manuscript are in italic font
- Page-, Line-, Section-, etc. numbers refer to the initially submitted (unrevised) manuscript unless stated otherwise.

Summary on the changes

Major changes on the manuscript were made regarding abstract, conclusions and section 2.3.1 (on the description of the statistical approaches; most changes were made in the course of the introduction of the "Bias" as described below). Further, Section 3.7 (the comparison of NO2 UV and NO2 Vis results) was completely eliminated and Supplement S2 (on the partial AOT correction) is now embedded into Section 3.4 in the main text (on the comparison of AOTs).

Answering some of the comments required minor revisions throughout the manuscript, of which not all are explicitly mentioned here. For an overview on all the changes taken, please refer to the Latexdiff_Manuscript.pdf and Latexdiff_Supplements.pdf files.

Answers

Tirpitz et al presents trace gas concentration (NO2 & HCHO) and aerosol extinction profiles of 15 participating groups derived from MAX-DOAS measurements and implementing different retrieval algorithms during the CINDI-2 campaign. The authors attempt to validate profiles/partial columns using collocated observations. This is an important effort since there are several retrieval approaches using MAX-DOAS measurements, and even though MAX-DOAS measurements started a while ago still there are not harmonized approaches to retrieve gases and aerosols. Hence, this is an important work and likely suitable for the journal. However, I have major comments and foremost revisions are warranted before publication. In my opinion, the quality of the paper needs to be improved before publication.

- According with the manuscript the main goal "is to assess their consistency with respect to different conditions and to review strengths and weaknesses of the individual algorithms and techniques" and they use supporting collocated measurements to "validate" the retrieval algorithms. However, authors include primarily results of retrievals using "median dSCDs" obtained in a separate study (Kreher et al., 2019). I do completely understand the value of using the "median dSCDs" but I also see an extreme value in including detailed results using each participant's dSCDs. The current approach seems quite unusual in a validation point of view. So far, section 3.8 describes briefly results using dSCDs of individual participant but needs to be expanded in the main body, abstract, and conclusions.

Response:
We fully agree with the reviewer's statement, that the retrieval results from the own dSCDs are of importance. But as mentioned by the reviewer in the specific comments below, discussing both in detail in a single paper goes beyond its scope, so the focus should be on one of the two. As our focus

was on the comparison exclusively of the retrieval algorithms, we consider the median dSCDs to be the better choice.

Nevertheless, we extended the information on the own dSCD comparison in the following ways:

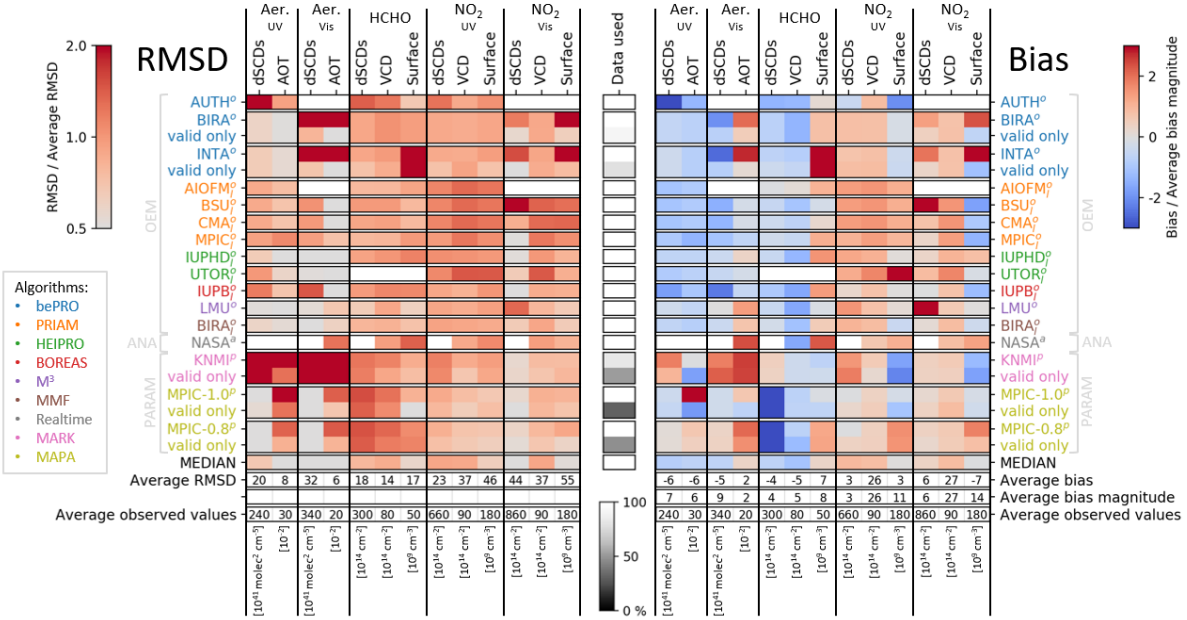
1. A summarising figure similar to Fig. 23 was created also for the own dSCD comparison and is contained in the supplementary material
2. In the corresponding Section (3.8) in the main text, "Bias" values (description below) were added in Table 5. Further, we now directly compare the impact of the use of own dSCDs and the impact of the use of different retrieval algorithms on the consistency among MAX-DOAS participants.
3. Corresponding discussions in the conclusions were extended.
4. Major results of the own dSCD comparison are mentioned in the abstract now

- The algorithms are assessed primarily with the root mean square difference. Authors focus primarily on this quantity, which is always positive, and definitively help to understand the comparisons, especially among instruments. However, I highly suggest to include a bias estimator to know the under-or overestimation with respect to the independent measurements. Figure 23 is key in the paper, and I highly suggest to include a similar figure but using bias in percent.

Response:
The "Bias" was introduced as an additional statistical parameter (see section 2.3.1) to capture systematic discrepancies between the individual evaluations (see also response to reviewer #2). It is simply defined as the weighted average of the difference between a pair of compared observations:

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

It appears as an additional parameter in the correlation analysis plots (Fig. 14, 17 and 20) and is discussed there. Further, the summarizing figure (Fig.23) was extended by a panel for the bias:



- I find very useful to include the three different type of algorithm approaches (OEM,

PAR, and ANA). However, a thorough analysis of what technique yields the best results is missing, especially in the abstract. According with the results, OEM seems to be most appropriate/reliable, but ANA approaches might be ideal for near-real time analysis. I would include a section with main finding regarding the comparison of these methods.

Response:
While we agree with the reviewer that it would be desirable to come to a conclusion which technique is "best" we feel that a quantitative statistical investigation on the results grouped by algorithm techniques is not very meaningful, because

1.) PAR and ANA approaches are heavily underrepresented compared to OEM.
2.) A single ANA and two PAR algorithms are not reasonably representative for the general technique. This becomes apparent for instance by looking at the two parameterised approaches which perform extremely different.

However, the advantages and disadvantages of the different techniques are qualitatively discussed in the conclusions (and have slightly been extended in the course of the revision).
Note that also among the authors there is not yet a consensus on the "best" approach, since this strongly depends on the assessment criteria. For the abstract we consider this topic as too complex to be discussed in an understandable and balanced way without going beyond the scope of the manuscript.

- For the groups using OEM, they use same dSCDs and main retrieval parameters are prescribed, still there are extremely large differences among the groups using OEM. A thorough analysis of the reason is missing.

Response:
Note that some of the reasons (e. g. in the case of the two HEIPRO participants) were identified, others not. As described in the paper, in detail OEM approaches can actually be implemented in very different ways. We agree that it would be extremely helpful to investigate the reasons for any of the deviations, however, we believe that this is not affordable at this point and out of the scope of a comparison paper, particularly of the given extent.

Additionally, if I understand correctly, the recommended altitude grid for all participants was from the surface to 4 km (20 layers of 200 m). This is quite unusual in transfer models, how is the atmosphere represented above 4 km?
If this is fact true I highly recommend having realistic information above 4 km.

Response:
We apologize for the misunderstanding. Some aspects here were not well communicated in the manuscript:

One must clearly distinguish between the "RTM grid" and the "retrieval grid". The RTM grid describes how the atmosphere is represented within the radiative transport forward model while the retrieval grid defines at what vertical resolution the actual inversion (e.g. the OEM formalism) is applied. In most retrieval algorithms, the RTM grid is inherently predefined by the developer and cannot be changed offhand (in particular in the case of look-up table approaches). In contrast to the retrieval grid, it typically features a higher resolution (25 m to 100 m layers close to the surface, increasing with altitude) and extends up to 40 to 90 km altitude. Radiosonde profiles of temperature, pressure and ozone were provided from 0 to 90 km altitude and implemented within the constraints of the RTM grid of the individual algorithms.

To make things clearer to the reader we changed the text:

From: *"Pressure, temperature, total air density, and O3 vertical profiles were averaged from O3 sonde measurements performed in De Bilt by KNMI during September months of the years 2013-2015. […] A fixed altitude grid was used for the retrieval, consisting of 20 layers between 0 and 4 km altitude, each with a height Δh = 200 m. The results of the parametrized approaches and OEM algorithms where the exact grid could not be directly implemented, were interpolated/averaged to this grid to simplify the comparison."*

To: *"Pressure, temperature, total air density, and O3 vertical profiles between 0 and 90 km altitude were averaged from O3 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 km altitude, each with a height of Δh = 200 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 transport simulations, the atmosphere was represented by finer (25 m to 100 m layers close to the surface, increasing with altitude) and farther extending (up to 40 to 90 km altitude) grid, inherently (and differently) defined by the individual retrieval algorithms."*

Furthermore, I am surprise that for the retrieval settings all participants use average values of pressure, temperature, and O_3 vertical profiles obtained in 2013-2015. However, the campaign was held in 2016. I believe pressure, temperature, water vapor, etc, might have an important effect in the forward model and foremost in the retrieval of aerosol extinction using O4. I do not understand why radiosondes (or even re-analysis) data obtained during the campaign are not used. If the goal is to validate profiles I highly suggest using the real atmospheric conditions during the campaign.

Response:
This comment is addressed in our response to the following comment.

- It is mentioned that "The ceilometer aerosol extinction profiles should be consulted for qualitative comparison only" and I fully agree due that many assumptions are used to calculate extinction from backscatter measurements. In this context, the aerosol extinction derived from the ceilometer cannot be used to validate the profiles. However, I do believe they offer you additional information that can be further used, especially for OEM. In the manuscript, a priori extinction profiles for both aerosol and trace gas retrievals were exponentially-decreasing and of course OEM will converge, i.e., it is an ill-posed problem. However, if you use the aerosol extinction profiles as an a priori at least you estimate a better profile shape and the OEM technique might give you a better result. I highly recommend to use the ceilometer extinction profiles as a priori profiles and compare with the exponentially decrease profile. Several questions might arise: do sensitivity increase at higher layers? do AKs change? is the partial AOD correction still the same?

Response:
We fully agree that the settings are not optimal and in particular for scientific studies (rather than methodological studies, as presented here), all available information should be used and the suggestions by the reviewer are exactly the way to go.

Yet, it must be considered:

1. The paper aims at the comparison and validation of MAX-DOAS profiles retrieved under typical measurement conditions. This includes using prior information as they are typically available for an arbitrary measurement location and season. Having daily radiosondes, ceilometer data and collocated sun photometer measurements at hand is not a very usual

scenario. In fact, most MAX-DOAS studies have to resort to climatologies for their prior assumptions.

2. Since the MAX-DOAS results are validated by the supported observations (at least qualitatively, in the case of the ceilometer profiles), they need to be kept independent, which is not the case if one observation serves as a priori for the other.

From this point of view, it is not obvious at which point to "stop" the adaption of prior information. Our settings are similarly carefully chosen as for other MAX-DOAS studies and therefore we think they are justified, as long as they are clearly communicated.

The reviewer's questions at the end of the comment can be answered qualitatively:

Do sensitivity increase at higher layers? do AKs change?
This depends on the a priori covariance. Since the uncertainty of ceilometer data is surely smaller than that of an exponential profile, the sensitivity and DOFs will decrease.

Is the partial AOD correction still the same?
No. Depending on the a priori covariance, aerosol profiles will remain close to the ceilometer profiles in particular at higher altitudes. Since the PAC is based on exactly these ceilometer profile, f_tau will be close to one and the PAC will not have any effect.

It is well-known that sensitivity needs to be considered when comparing different measurement techniques. However, after reading the manuscript it sounds like you introduce new findings, e.g., last short paragraph in the abstract. I do not think it is assumed that integrated extinction profiles from MAX-DOAS and the AOD from the sun photometer should be comparable. In my opinion, this is not a finding or result in this paper. I suggest to re-write your findings accordingly, e.g, include that after smoothing (applying the "AOD correction") comparisons yield better results.

Response:
We agree that it is not a "finding" or result that sensitivity needs to be considered. Generally, this is well known and applied. After reading again through former publications, we also found that the low sensitivity at higher altitudes was already suggested e.g. by Irie (2008) and Frieß (2016) to explain the discrepancies between sun photometers and MAX-DOAS observations, but it has not been proven. This information was added now in the beginning of Section 3.4.:

"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 5 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."

In any case the last paragraph in the abstract as submitted is pretentious and misleading. We therefore reformulated it in a similar manner:

"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 smoothing effects, namely biases arising from the reduced sensitivity of MAX-DOAS observations to higher altitudes and associated a priori assumptions."

Related statements in the main text were adapted correspondingly.

In fact, I think authors should describe that this correction (partial AOD correction) is related to the O4 scaling factor used in past studies (and here too for some groups). If I understand correctly, the "AOD correction" yields better results/comparisons because sensitivity in mainly in the lower troposphere, hence aerosol layers aloft are not captured with MAXDOAS. In this context, after reading Ortega et al. (2016) this reference is not pointed out but offers some insights and should be included.

Response:
Also to us a direct relation between the O4 scaling factor and the PAC seemed obvious in the beginning. However, after reading different publications on this issue (Wagner (2009), Clémer (2010), Ortega (2016) and Wagner (2019)) we believe that the relation is weak for several reasons:

1. The motivations are very different: the application of the PAC is necessary solely for mathematical reasons related to the concept of optimal estimation and prior constraints applied therein. In contrast, all publications listed above compare forward modelled O4 dSCDs (using an atmosphere derived from supporting observations to reproduce the real conditions as good as possible) to measured O4 dSCDs. They do not make use of optimal estimation or a priori profiles similar to those used in our study. Thus their findings are independent from any kind of PAC.
2. The PAC correction factors are dependent on the a priori profile and covariance. In principle, by changing the a priori constraints, any arbitrary correction factors can be generated. The agreement of the CINDI-2 PAC correction factors with typically applied scaling factors (≈0.8) must therefore be considered to be coincidence.
3. Not all discrepancies between MAX-DOAS and sun photometer are explained by the PAC. As shown in our study, biases remain (Figure 14) in the UV, that can indeed be removed by additionally applying a weaker (campaign averaged) O4 dSCD scaling factor of approx. 0.9 (Supplement, Figure S4). It is well possible, that stronger scaling is necessary for individual days.
4. Applying a scaling factor improves the agreement of modelled and measured O4 dSCDs (Supplement, Figure S5). However, we admit that the discrepancies might also be induced by a priori assumptions limiting the scope of the forward model.

This issue is discussed in the paper main text and also in the conclusions.

Regarding point 1, we added further explanations on P29L4:

"[…]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; Ortega et al., 2016; Wagner et al., 2019) directly compare forward modelled O4 dSCDs (using an atmosphere derived from supporting observations to reproduce the real conditions to best knowledge) to measured O4 dSCDs. They do not make use of optimal estimation or prior constraints similar to those used in our study. Thus their findings can be considered independent from any kind of PAC."

- Lastly, I do not agree that retrievals of NO2 in the UV and vis should give you same results, unless you proof homogeneity around the line of sight.

Response:
We agree. At least the potential inhomogeneity complicates the interpretation. We therefore removed the section according to the reviewer's suggestion below.

Specific Comments

P2, L1-6. This paragraph does not belong here, I suggest to move it to the introduction and expand the abstract based on major comments.

The paragraph was removed. The introduction already contains a very similar paragraph.

P2, L2. Change "boundary layer and the lower troposphere" with "lower troposphere"

The phrase was removed with the above paragraph. The introduction contains a similar statement, there it was corrected.

P2, L3. Change "radiation" with "absorption"

Is obsolete, since the corresponding paragraph was removed. A similar sentence in the introduction was corrected.

P2, L5. I would explicitly say that you retrieve aerosol extinction concentration for profiles.

We assume that the reviewer meant "aerosol extinction coefficient profiles"(?). Comment is obsolete since the line was removed. However, we adapted corresponding statements in the main text.

P2, L10. Include all the supporting observations and remove others in the parenthesis.

Done.

P2, L15. Do you mean magnitude instead of intensity?

Yes, changed.

P2, L15-20. Results are shown in root mean square, however, in order to have a more quantitative description please also include the bias in percentage, or the rmsd in percent. Otherwise, it is hard to interpret the magnitude of the differences.

Since many different RMSD values are given in the abstract (different species, different observations) we decided to simply add the average observed AOTs, VCDs and surface to simplify the interpretation of all RMSDs. As stated above the bias was introduced, but to obtain a concise abstract we decided to only show RMSDs which reflect both, systematic and random discrepancies at once.

P2, L21-23. It is well-known that different sensitivity needs to be considered when comparing different measurement techniques. I do not think it is assumed that integrated extinction profiles from MAX-DOAS and the AOD from the sun photometer should be comparable. In my opinion, this is not a finding or result in this paper. There is nothing new on this short paragraph. I suggest to remove this paragraph or re-write your findings accordingly, e.g, include that after smoothing (applying the AOD correction) comparisons yield better results due that similar air masses are compared.

See our answer in the major comments above.

P2, L26-28. Transport is missing in your description of chemical composition in the PBL.

We agree and changed the text from: *"Its chemical composition and aerosol load is determined by gas and particulate matter exchange with the surface and also driven by homogeneous and heterogeneous chemical reactions."*

To: *"Its chemical composition and aerosol load is driven by the exchange with the surface, transport processes and homogeneous and heterogeneous chemical reactions."*

P3, L5. I agree that MAX-DOAS is a well-established technique with information of absorption signature of trace gases. However, it is misleading because the whole point of these type of studies is that MAX-DOAS is NOT a well-established technique to measure accurately gas concentration.

We only partly agree. Intercoparison studies are still valuable and necessary, also for well established techniques. On the other hand, such a differentiation is probably too detailed for the first sentence on MAX-DOAS. We replaced "*well-established*" by "*widely used*", which is a weaker statement.

P3, L6, It is mentioned that MAX-DOAS infers information in the boundary layer and free troposphere. Please include some references for both cases.

Note that this sentence has been changed by addressing a comment above. Now we state that MAX-DOAS infers information "*on the lower troposphere*". Corresponding references are listed in the manuscript in the three lines directly above (P3, L3-5).

P3, L8. I would remove "from the top of the atmosphere (TOA) to the instrument"

Done.

P3, L10. Change "Detectable gases are nitrogen dioxide (NO2), formaldehyde (HCHO): …" with "Gases that have been analyzed in the UV and visible spectral range are nitrogen dioxide (NO2), formaldehyde (HCHO): …"

Done.

P3, L18. Change "radiative transport models" with "radiative transfer models".

Changed. Also in further occurrences.

P3, L19. Change "such" with "of"

We do not understand. "of" does not make sense here (grammatically). We replaced "*numerous such algorithms*" by "*numerous retrieval algorithms*" instead.

P3, L23. What do you mean by different conditions?... Weather conditions, pollution conditions?

The major differentiations made during the comparison are w.r.t. cloud conditions and whether flagging of profiles is allowed or not. However, it is not necessary to spell it out at this point of the manuscript. We therefore deleted the phrase "different conditions":

We changed: *"The main objective of this study is to assess their consistency with respect to different conditions and to review strengths and weaknesses […]"*

To: *"The main objective of this study is to assess their consistency and to review strengths and weaknesses […]"*

P3, L30. Again, add all supporting instruments and remove "others ". Otherwise, remove "others".

Done.

P5, L16. Mention shortly what other effects, otherwise remove this.

The comment refers to equation (1): $$\tau_\lambda(\alpha) = \log\left(\frac{I_{\lambda,TOA}}{I_\lambda(\alpha)}\right) = \sum_i \sigma_{i,\lambda} S_i(\alpha) + C$$

The variable "C" is a placeholder for a potentially long list of physical and instrumental effects (linear as well as non-linear), that are not of immediate relevance for the actual comparison study. Listing them here might not be very helpful. We think the "C" should still be mentioned to give consideration to them. As a compromise we add one prominent example in brackets.

We changed: "*C represents further terms accounting for other effects than trace gas absorption that will not be further discussed in this context*."

To: *"C represents terms accounting for other instrumental and physical effects than trace gas absorption (for instance scattering on molecules and aerosols)"*

P5, L27. I do not see see how Apituley et al fits in this study.

Thanks, we changed that to Apituley, 2020.

P5, L28 – P6, L9. As mentioned above, I see the value of using the "median dSCDs", but I strongly suggest to include in detail (and not in the supplement) the retrieval results using their own dSCDs. In fact, I recommend the "median dSCDs" to be included in the supplement if authors believe the manuscript will be lengthily.

As mentioned by the reviewer, discussing both types of data in detail in a single paper is problematic, so the focus should be on one of them. Whether the "own" or the "median dSCDs" are favoured depends on the aim of the paper. As our focus was on the comparison of the retrieval algorithms, the median dSCDs are the right choice. This is motivated in more detail in the manuscript P6, L1-6 (initially submitted version). However, as stated in our answer on the first major comment above, we added some additional information on the own dSCD results.

P6, L22. How is water vapor profile included in the forward model? is it important? Also, remove the dots after aerosol microphysical properties.

Most forward models allow to include water vapour. Therefore, we added it to the list. In the UV/Vis, there are a few H_2O absorption bands and the presence of H_2O changes the average Rayleigh scattering cross-section in the atmosphere but the total effect on the dSCDs (and thus the retrieved profiles) is very small. Assuming typical H_2O concentrations encountered during the CINDI-2 campaign, dSCD simulation results with and without H_2O differed by about 0.1 %. It was therefore considered negligible and was not prescribed in the retrieval settings.

We changed: *"(aerosol extinction, trace gas amounts, temperature, pressure, aerosol microphysical properties, ...)"*

To: *"(aerosol extinction, trace gas amounts, temperature, pressure, water vapour and aerosol microphysical properties)"*

P6, L25. What is p? Also, I'm surprise to see 4 DOF, for what gas? is there a referene?

p is implicitly defined here to be the DOFS. We made this clearer:
We changed: *"Typically only p = 2 to 4 degrees of freedom for signal (DOFS) […]"*
To: *"Typically only two to four degrees of freedom for signal (DOFS or p) […]"*
DOFS of 4 were actually achieved for NO2 Vis within this study for distinct profiles.

P7, L3. The short OEM description seems awkward. Remove "filling". In general, you have an ill-posed problem and the solution is constrained by an a priori state vector.

We revised the description. It is now:
"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."

P7, L7. It is mentioned that PAR require more memory, and the sentence sounds like this is a limitation. How much memory is needed for such a short campaign? Satellites use look up tables.

The campaign duration is irrelevant. The look up tables are calculated once over the parameter space of interest (realistic atmospheric/measurement scenarios) and can then be applied to any campaign dataset. For the PAR algorithms presented in this study, a look up table for ground-based aerosol and trace gas retrievals at multiple wavelengths requires about 1 GB of memory.

P7, L13. "The M3 algorithm by LMU appears as an additional algorithm in our study" looks awkward. What do you mean? Re-write this sentence. Why its description is included in the supplement?

We changed: "*The M^3 algorithm by LMU appears as an additional algorithm in our study*"
To: "*Besides the algorithms described therein, our study includes results from the M^3 (OEM) algorithm by LMU."*

We first included the description in the main text, however there it appeared out of place and rather distractive, this is why we moved it to the Supplements.

P7, L25. As mentioned in the general comments. I highly suggest using real atmospheric conditions instead of average PTW from other years.

See our answer to the corresponding general comments.

P7, L27. See my comment above regarding the altitude grid, it is not clear what was used above 4km.

See our answer to the corresponding comment above.

P7, L33. My understanding is that the AERONET angstrom exponent (440-675 nm) derived from a single day (14 Sep) is used to extrapolate to 360nm for all days during CINDI-2, is this correct? If this is correct, please explain why you use a single day and not coincident measurements. I expect the angstrom exponent changing unless you have similar aerosol composition.

Yes, this is correct. See our comment on the choice of prior information in the general comments above.

P8, L25. Remove the "…" in the sentence in parenthesis. Check many other sentences like this along the manuscript.

Done.

P9, L11. Change "true aerosol extinction" with "aerosol extinction". Many assumptions are carried out for the creation of extinction profiles and might not be the true aerosol extinction.

Done.

P9, L23. What mean error does the 0.03 RMSD represent?

We forgot the unit here (it's extinction coefficient in km^{-1}) and also over which altitude interval this value was calculated.

We changed: "*The average RMSD between scaled ceilometer and Raman lidar profiles is ≈ 0.03.*"

To: "*The average RMSD between scaled ceilometer and Raman lidar profiles up to 4 km altitude is ≈ 0.03 km^{-1}.*"

P9, L25. At the end of section 2.2.2 it is pointed out that "the ceilometer aerosol extinction profiles should be consulted for qualitative comparison only", which I fully agree since many assumptions are carried out to derive extinction profiles. In this case, the retrieval of extinction profiles cannot be fully validated during CINDI-2.

Yes, we agree with the reviewer's conclusion, this is why we stated that the aerosol extinction profiles should be consulted for qualitative comparison only. To emphasize that the focus is on AOTs, VCDs and surface concentrations, we added corresponding statements in abstract and conclusion:

"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."

P9, L25. It is mentioned that NO2 profiles from sondes and lidars were carried out sporadically, but include a description of how often. How many sondes were launched?

For the radiosondes this is given and referred to: a few lines further down, we reference Supplement S5.2, which includes a list with the details on each radiosonde flight.

For the Lidar we added a sentence: "*This resulted into 25 suitable Lidar profiles recorded on six different days during the campaign.*"

Note, that the exact timing of both observations can also be inferred from the comparison plots of the actual comparison (e.g. Fig. 16 and 19)

P12, L15. For the "different observations" do you mean MAX-DOAS and supporting measurements?, or different groups using MAX-DOAS?. Please clarify.

At this point "different observations" refers to any observation. This comprises multiple cases which are subsequently discussed in the same paragraph. The paragraph was revised in the course of the introduction of the "Bias" and should be clearer now.

P12, L18. IS xref,t measurement from a reference measurement?, i.e., collocated supporting observation?. Clarify.

It's either the MAX-DOAS median results or a supporting observation. This was clarified in the course of the revision of the paragraph.

P12. While the root mean square difference is useful, this is always positive. I highly recommend to include a bias to see the sign of bias with respect to collocated observations. Simply, use something like this: bias = median(max-doas-reference)/reference when comparing to collocated supporting observations.

As stated above, the "Bias" was introduced as an additional statistical quantity.

P13, L12. It is mentioned that UV and Vis dSCDs should be the same. I disagree, light path in the UV and Vis might be different. Hence, different dSCDs.

As suggested below, this comparison has been eliminated.

P14, Section 2.3.3. I believe you can quantify the spatial mismatch between sonde-MAXDOAS by using the sonde gps information. It might be interesting to see the actual spatial difference. Section 3.1.

We agree that this is useful: we added a table (S6 in the new manuscript version) with the average temporal and spatial mismatches between MAX-DOAS observations and all supporting observations in Supplement S7:

Table S6. Estimates for the average spatio-temporal mismatch of different supporting observations w.r.t. to the MAX-DOAS measurements. For the location of the MAX-DOAS observations the centers of mass of the horizontal sensitivity curves from section S6 were used. For the location of sun photometer and DS-DOAS observations, the center of the lines of sight towards the sun up to 2 km atitude were considered.

Observation	Spatial mismatch [km]	Temporal mismatch [min]
Sun photometer	13	8
Ceilometer	11	0
DS-DOAS	13	23
NO_2-Lidar	10	9
Radiosonde	6	13
LP-DOAS	10	6
In-situ in tower	11	0

We further refined our discussion in Supplement S7 according to these numbers and now also present a rough estimate of the impact of spatio-temporal variability on the comparison of NO2 surface concentrations in Sect. 2.3.3.:

"Table S6 summarizes the spatial and temporal mismatches between MAX-DOAS and supporting observations. Spatial mismatches are of the order of 10 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 Supplement S6) yielding an RMSD of $3.5x10^{10}$ molec cm^{-3}. This is indeed of similar magnitude as the average RMSD observed during the comparison (approx. $5x10^{10}$ molec cm^{-3})."

Discussion paper

P15, L12. "Figure 2 visualizes the average AVK matrices"… what do you mean by average AVK?. Are these averages of a single group using OE, or average of all groups?

It's the median over participants and the mean over time. This is described in the figure's caption but we also added it to the main text in brackets.

The text reads now: "*Figure 2 visualizes the average AVK matrices (median over participants and mean over time) […]*"

P15, L13. I agree with this "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" and I think some literature is missing, e.g., Friess et al (2006) showed that aerosol extinction above 3km can be retrieved using O4 dSCDs measured at different wavelengths. Ortega el al. (2016) showed that elevated aerosol layers modify O4 dSCDs, hence some sensitivity of aerosols aloft. In my opinion, this is a clear effect of an ill-posed issue, where an appropriate a priori information is important. In this case, I do not agree with authors claiming that there is not sensitivity of layers aloft, but it is difficult to retrieve layers aloft due to assumptions and less-ideal a priori information.

We reworded several statements on this issue throughout the manuscript to clarify that the low/no sensitivity to higher altitudes is not fundamental. E.g. on P15, L16, where we changed the text:
"For all species, the sensitivity is limited to about the lowest 1.5 km of the atmosphere."
To: *"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."*

We further added a paragraph to section 2.3.2. (on smoothing effects):

"It shall be pointed out however, that the sensitivity and spatial resolution is strongly affected by the exact approach that is chosen to solve the ill-posed inversion problem. Frieß (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. Also the sensitivity depends on the atmospheric state: the presence of clouds and aerosols at higher altitudes for instance change the radiative transport and can increase sensitivity particularly to the layers where they reside."

P15, L27. It is mentioned that "the presence of clouds can increase the sensitivity to higher layers due to multiple scattering and thus light path enhancement in the clouds". If clouds can enhance the sensitivity at higher altitudes, aerosols might have a similar effect, correct?

Yes, correct. This is now also mentioned in the text (see answer to the comment before).

P28, L3, I would add if a priori information is not reliable at the end of this sentence: "high-altitude abundances of trace gases and aerosol typically cannot be reliably detected by ground- based MAX-DOAS observations "

Note, that according to suggestions by reviewer 2 the wording was changed from:
"high-altitude abundances of trace gases and aerosol typically cannot be reliably detected […]"
To: "*high-altitude abundances of trace gases and aerosol typically cannot be reliably located and quantified […]"*

We only partly agree with the statement of the reviewer. It is only right if by "reliable a priori information" the reviewer means "the state of the atmosphere is known before the inversion". Otherwise, there will always be biases, also if a priori profile and covariance perfectly reflect the prior knowledge.

P28, L11. If I understand correctly, in addition to the description provided, the ratio from equation 11 provides you the fraction of the aerosol retrieved by OEM. So, a factor of 0.8 means that about 20% extinction should be aloft, is my interpretation correct? If so, I think this is a very important result and should be further explained. Furthermore, could this fraction be related with the correction factor?

Yes, the reviewer's interpretation appears to be correct. This should become clearer now since we embedded Supplement S2 (detailed results of the PAC factors) into the main text Section 3.4. Regarding the relation to the scaling factor please refer to our answer in the general comments.

P29, L3. It is mentioned that "a scaling of the measured O4 dSCDs prior to the retrieval with SF _ f_ might be used to at least partly account for the PAC for MAPA and probably other PAR and ANA algorithms (see Supplement S3), even though the physical reason for PAC and SF are different.", please explain further and provide the physical differences between PAC and SF. Would it be possible that past correction factors are used due that they miss aerosols aloft, which if I understand correctly might be in agreement with findings in Ortega et el. (2016)?.

We have to correct this statement regarding the "physical reason", as it is not well-founded. We replaced the sentence and extended the paragraph by a further explanation:

"[…] *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, 2009; Clémer, 2010; Ortega, 2016; Wagner, 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. They do not make use of optimal estimation or prior constraints similar to those used in our study. Thus their findings can be considered independent from any kind of PAC."*

Regarding the reviewers question **"Would it be possible that past correction factors are used due that they miss aerosols aloft":**
To our knowledge the typically observed disagreement between total AOT observations and MAX-DOAS integrated aerosol profiles has indeed been regarded as another evidence for the need of a scaling factor in some publications but it never was the primary argument.

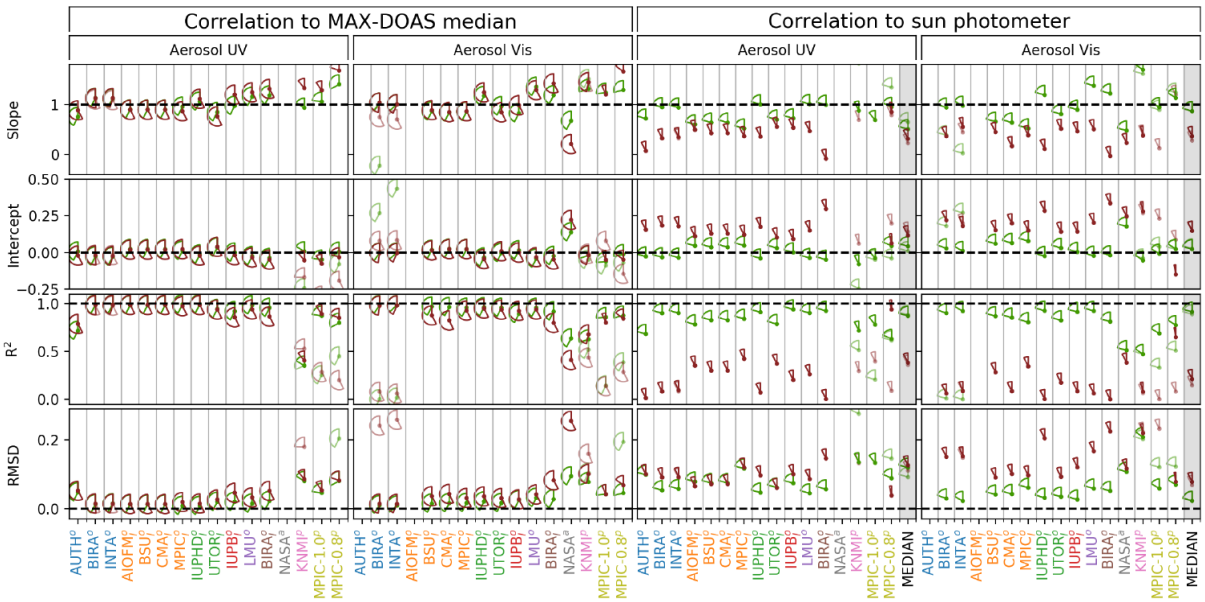
P29, L9. "underprivileged" sounds weird, please change it.

We changed: "*Keep in mind that the non-OEM approaches (NASA, KNMI and MPIC/ MAPA) are correlated against τ_s and might therefore be underprivileged*"
To: "*Keep in mind that the non-OEM approaches (NASA, KNMI and MPIC/ MAPA) are correlated against τ_s and are therefore expected to generally achieve worse agreement*"

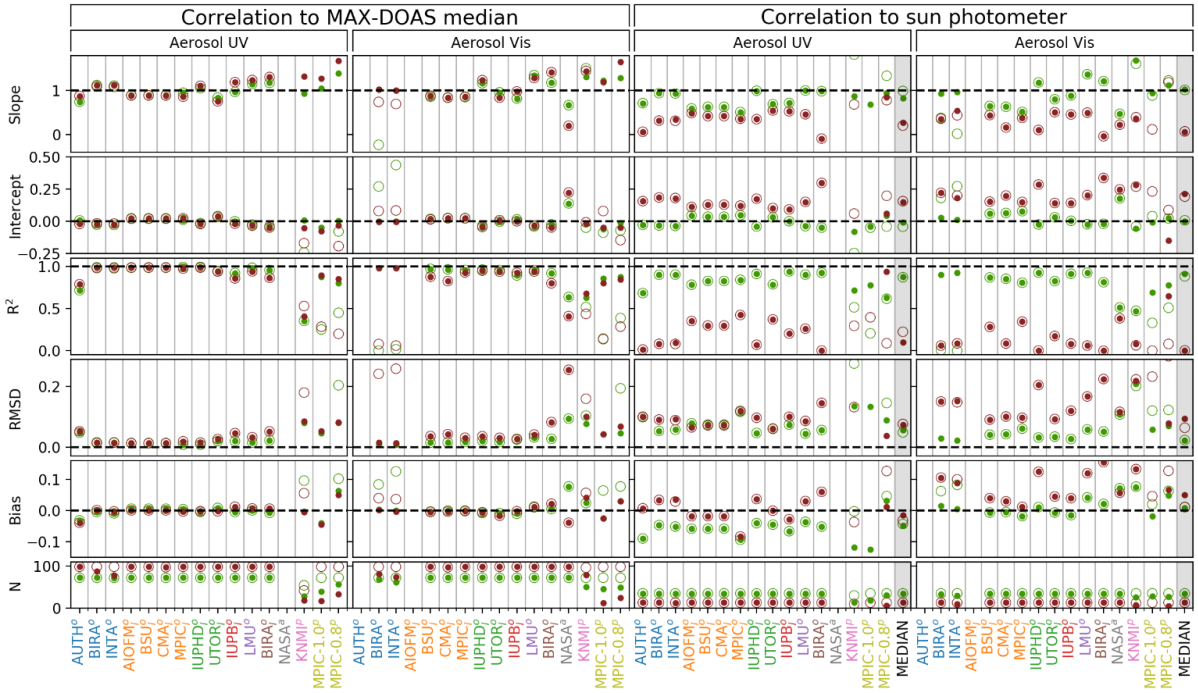
P30. figure 14. Please add bias in % (negative/positive) as mentioned above. Additionally, light vs opaque are not distinguishable, maybe using other colors might help? Furthermore, symbols on the two right column plots are not shown in the legend, maybe you meant to use the same symbols?

Figure 14 (and similar figures afterwards) were revised accordingly, also considering comments by reviewer 2.

Fig. 14 in the submitted version of our manuscript:



Now:



P30, L10. As suggested above, please include the bias in percent here, in addition to the rmsd.

Biases are discussed now. In this particular case:
"All participants except MPIC-0.8/ MAPA underestimate the AOT (Bias < 0.03) in the UV, despite the PAC has been applied for the OEM algorithms. 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 values are -0.06 in the UV and 0.02 in the Vis."

P31, L1-2. In the text, it would be handy to describe the group (as in Figure 14) and in parenthesis the approach/name) in order to avoid going to table 2 every time. For example, PRIAM is mentioned in line 2 but this is not in figure 14 and table 2 needs to be checked.

We changed this: we now use the "participant/ algorithm"–notation where applicable and otherwise "algorithm (list of participants)" notation.

P31, L7. KNMI/ MARK and NASA/ Realtime are mentioned as high rmsd, but I also see MPIC being high but not included in the text. So, all parameterization approaches show high rmsd.

We added MPIC/ MAPA to the list.

P31, L9-12. It seems like the correction factor improves the agreement, but further description is missing. According with your "partial AOT correction" this might be due that PAR approaches miss layers aloft?. I consider this an important finding but is not described.

This should be solved since we embedded Supplement S2 (detailed results of the PAC factors) into the main text Section 3.4.

P35, Section 3.7. I do not agree that NO2 Vis and UV should yield similar results, unless you show with independent measurements that there is homogeneity in the sensitivity range (vertical/horizontal). Rather than an "intrinsic consistency check" I would use this section to

actually assess inhomogeneity. On the other hand, the manuscript is long enough and I would consider removing this section.

Section and corresponding references were removed as suggested.

P37, Section 3.8. This section is important and deserves more description. A bunch of figures have been thrown in in Supplement S10 but not a complete description. In my opinion, this is a key section to show how reliable are the MAX-DOAS products, hence I also recommend a thorough description of the bias per participant, and not only rmsd.

See the answer on this issue in the general comments above.

P38, L11. Please include the approaches. Some people only read conclusions. I recommend to explicitly describe that lower tropospheric columns are assessed. I suggest to include the algorithm next to the group name, maybe in parenthesis. I suggest to include another figure, similar as Fig. 23, but for the bias in percent.

The whole conclusions were revised, also considering these suggestions.

Profiles are not really assessed, especially for trace gases.
Note that the agreement of profiles is assessed (however not discussed) in the Supplements (Fig. S21 to S25). In the main text we focus on those quantities that we have supporting observations for, to not further extend the manuscript.

Figure 23. It is difficult to track what algorithm is used for each group.
In the course of the revision of Fig. 23 (see also general comments) we added a list of the algorithm names in the corresponding colours.

P40, L20. It is mentioned that "O4 scaling and PAC were found to have similar impact on the MAX-DOAS AOT results." In my opinion, this is a major finding. It is shown that sensitivity needs to be considered when comparing two different remote sensing techniques, and here you have shown that the lower tropospheric column of extinction agrees well with Total column of AERONET when "corrected". This "PAC" is the same as the O4 scaling factor and by reading Ortega et al. (2016) might be due that aerosol layers aloft are normally neglected. I highly recommend to further describe this.

We only partly agree. Particularly, we disagree with the reviewer's statement "This "PAC" is the same as the O4 scaling factor". See our answer in the general comments.

Answers to anonymous referee 2

General information

First of all, we would like to gratefully acknowledge the efforts taken by the reviewers to read and revise this extensive manuscript. We are convinced that their comments helped to significantly improve the manuscript regarding comprehensibility and completeness, particularly in the conclusions.

Document formatting

- The reviewer's comments are reprinted here in bold face.
- Our answers are given in regular font
- Explicit changes made in the manuscript are in italic font
- Page-, Line-, Section-, etc. numbers apply for the initially submitted (unrevised) manuscript unless stated otherwise.

Summary on the changes

Major changes on the manuscript were made regarding abstract, conclusions and section 2.3.1 (on the description of the statistical approaches; most changes were made in the course of the introduction of the "Bias" as described below). Further, Section 3.7 (the comparison of NO2 UV and NO2 Vis results) was completely eliminated and Supplement S2 (on the partial AOT correction) has been embedded into Section 3.4 in the main text (on the comparison of AOTs).

Some comments required minor revisions throughout the manuscript, of which not all are explicitly mentioned here. For an overview on all the changes taken, please refer to the Latexdiff_Manuscript.pdf and Latexdiff_Supplements.pdf files.

Answers

Tirpitz et al. present a thorough assessment of MAX-DOAS profile retrieval algorithms using data collected during the CINDI-2 intercomparison exercise. The work is to this reviewer's knowledge the most comprehensive and up-to-date assessment of MAX-DOAS inversion using field data. As such, the work is worthy of publication.

However, the scale of the work presents certain challenges in understanding. Including the supplemental materials, the total work is 106 pages of text figures and references in length. As such it is likely that many readers will not consume it in its entirety. Several seemingly minor or technical conventions adopted for communication are at risk of creating misunderstanding if the work is read only in part.

Response:
We like to thank the reviewer for the commending words. Having addressed the reviewers' comments below and after revision particularly of abstract and conclusions, we are confident that this has improved in the new version of the manuscript.

Of critical importance, several possible reasons of discrepancies between MAX-DOAS and other techniques, and among MAX-DOAS inversions are identified and discussed at length yet the assessment of the relative relevance and importance of these is left unclear to the reader.

Response:
The study is meant to be a comparison, in the first instance quantifying the (in-)consistency of the different observations during CINDI-2. Further, likely reasons for the discrepancies were identified.

Of course it is highly desirable to even quantify all these effects, however, we believe that this is not affordable and out of the scope for a comparison paper, particularly of the given extent

Anyway we made corresponding efforts using available data and resources, but not all yielded simple quantitative results. Still we decided to publish them within the supplementary material, since they provide qualitative information which we hold to be of value.

Finally, we agree, that particularly the conclusions lacked quantitative results that are actually assessed during the study. In this regard we revised the conclusions considering the specific comments from both reviewers.

A concise summary of findings should be included in the abstract.

In this regard we also revised the abstract, considering the specific comments of both reviewers.

Specific major comments:

1) The authors make use of a number outside measurements (sometimes in combination) for the purposes of "validation". However, a statistical assessment of the validation is not transparent and digested. A summary of the form and source of discrepancies is distinctly lacking. The RMSD approach is adopted by the authors to capture both systemic differences and statistical noise, yet as the authors discuss RMSD sometimes reflects random variations and other times systemic differences. However, this discussion is scattered and not collected and summarized. Some systematic summary is needed. Comparisons to the validation products similar to Figs. 8 – 12 or 21 and 22 would suffice, although ideally the comparison would be more concise.

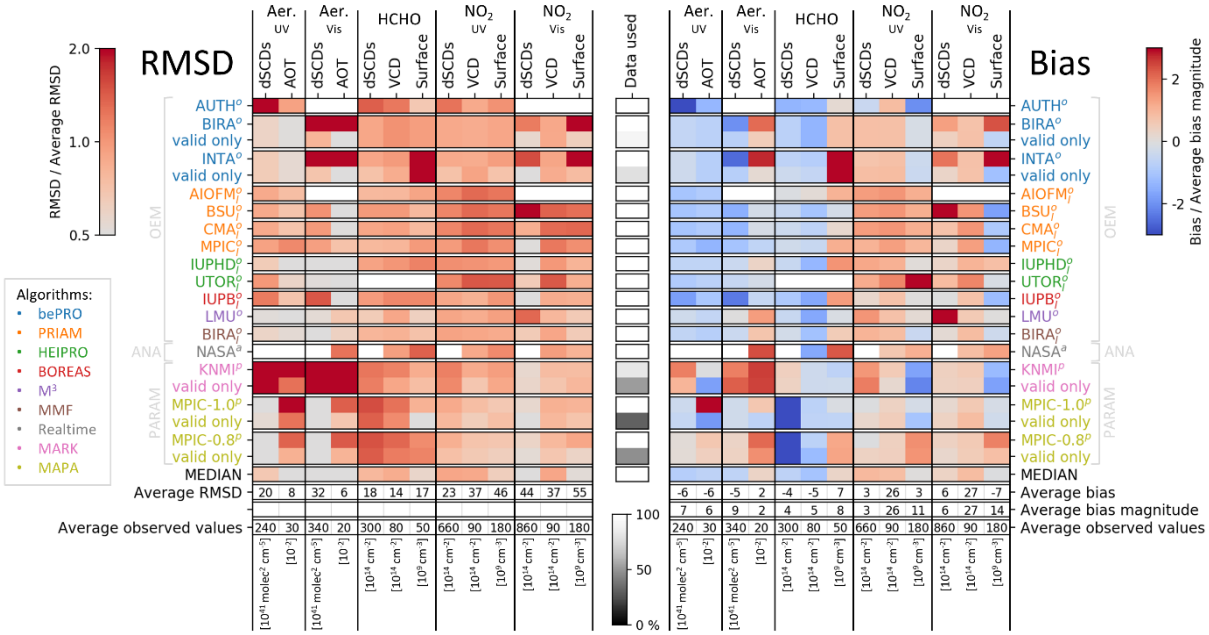
Response:
The conclusions were revised as stated above and according to the specific comments below (see also response to reviewer #1).

The "bias" was introduced as an additional statistical parameter (see section 2.3.1) to capture systematic discrepancies:

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

It appears now in the correlation analysis plots (Fig. 14, 17 and 20) and is discussed at relevant locations in the manuscript.

The new, summarizing figure (Fig.23) at the very end of the document was extended, amongst others by a panel for the bias:



Comparisons to the validation products similar to Figs. 8 – 12 or 21 and 22 exist and are included in the supplement. In the main text the regression results of these scatter plots are summarised in Figures 14, 17 and 20 for compactness. This way of visualisation was adapted from Frieß (2019) and Kreher (2020).

a. Supplement 5 gives some indication of the comparison of the differences between different measurement methods. Tables S4 and S5 give some indication of the relative magnitude of RMSD with the specified uncertainties (σ). However, it is not fully transparent which measurements contribute most to σ, nor whether the reported RMSD is primarily random or systematic. Systematic differences should be summarized, preferably the remaining residuals after correcting for systematic differences also.

Response:
We added the specified uncertainties of each observation in the tables (in brackets behind the corresponding labels). These values now also appear in the conclusions of the main to assess their contribution to the overall RMSD observed between MAX-DOAS and supporting observations.

We decided to not further extend the tables in the Supplement by bias or residual values, since these would then only be assessed w.r.t. to other supporting observations (not w.r.t. the truth). This is however not of major relevance for the main comparison, where particularly in the statistical analysis only single supporting observations are compared to the MAX-DOAS data. We hold it to be sufficient that the reader can draw the systematic and random discrepancies among the supporting observations qualitatively from the scatter plots in the figure above (Fig. S10).

The updated tables are now:

Table S4. Comparison of redundant measurements of the NO_2 surface concentration (in 10^{11} molec cm^{-3}). For each pair of observations, the observed scatter (RMS) is compared to the specified uncertainty (σ).

	Tower in-situ (0.56)		Radiosonde (0.50)		NO_2-Lidar (0.13)	
	RMSD	σ	RMSD	σ	RMSD	σ
LP-DOAS (0.06)	0.32	0.56	1.01	0.51	0.57	0.13
NO_2-Lidar (0.13)	0.72	0.57	0.40	0.52	-	-
Radiosonde (0.50)	0.99	0.78	-	-	-	-

Table S5. Comparison of redundant measurements of the NO_2 total columns (in 10^{16} molec cm^{-2}). For each pair of observation, the observed scatter (RMS) is compared to the specified uncertainty (σ).

	Radiosonde (0.44)		NO_2-Lidar (0.15)	
	RMSD	σ	RMSD	σ
DS-DOAS (0.23)	0.24	0.51	0.40	0.26
NO_2-Lidar (0.15)	0.34	0.48	-	-

b. In Sect. 3.8 and Supplement 10 instrument specific dSCDs are used for inversion rather than the median dSCDs. This most closely matches how the inversions would typically be applied. The authors show an impact on RMSD, including for some data products a decrease. However, it is unclear whether the error contribution from the dSCDs or from the inversion is greater or even whether they are similar in magnitude. Quantitative comparison presents several challenges, however, the authors should at least address this question.

Response:
We agree with the reviewer that this is important information. The most reliable way to determine it would be to evaluate the own dSCD datasets of all participants with the same algorithm (and ideally repeat this with each participating algorithm). However, this would be a large effort compared to the benefit. Note also that the result's general validity is limited: in the case of CINDI-2, the experience with the MAX-DOAS technique varied strongly among the participants. The quality of the own dSCDs might therefore not be representative for MAX-DOAS observations performed by experienced groups.

We therefore chose a simpler approach to obtain corresponding estimates. We explain it in the the following paragraph added to Section 3.8.:

"It is also of interest to explicitly estimate which fractions of the total observed discrepancies among the different MAX-DOAS profiling results 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 effects 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 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. The residual variance can be attributed to the choice and setup of the retrieval algorithm."

We also added a corresponding discussion in the conclusions.

2) The authors state that species more than ≈1 km above the MAX-DOAS detectors cannot be reliably detect, but then discuss at length the impacts of signals originating at these altitudes on

the retrievals. As such these signals are by demonstrably detected. Rather, the limitation the authors refer to is in determining the magnitude, shape, and location of the relevant signals. The language should be edited to reflect this.

Response:
Corresponding statements were adapted throughout the manuscript.

3) Related to points 1 and 2, some of the limitations of inversions are reported as fundamental, when, in fact, they are the result of design decisions. For instance that OEM retrievals tend toward the a priori is not surprising and is a reflection of the construction of the a priori as well as the covariance matrix. Similarly, that parameterization retrievals fail to capture cases which cannot be described by their limited set of parameters is not surprising either. Importantly, these examples point to specific improvements which should be made, namely a priori profiles and parameterizations need to be designed to better reflect reality. For OEM retrievals the specification of covariance must also be critically assessed. Statements to this effect are found in the supplement, however, they are fundamental to the findings and should be prominently featured in the main text.

Response:
Corresponding statements in the main text were adapted and extended to better describe the role of a priori profiles and covariance and to emphasize, that the limitations of inversions depend on their choice (see also specific comments below). Further, we moved Supplement 2 (describing the PAC results) to the main text, providing additional insight on these aspects.

4) The authors report root-mean-square differences, for aerosol optical thickness, trace-gas columns, aerosol extinction, and trace-gas concentrations as absolute errors. The relative magnitude of different errors are also compared as percentages. However, a comparison of root-mean-square differences with the relevant reported median/mean value is lacking. This makes the comparisons difficult to assess outside the particular community of experts.

Response:
Note, that all these information is included in the summarising Figure 23. However, we also added a corresponding sentence to the abstract as well as to the conclusions:

"These values compare to approximate average optical thicknesses of 0.3, trace gas vertical columns of $90x10^{14}$ molec cm^{-2} and trace gas surface concentrations of $11x10^{10}$ molec cm^{-3} observed over the campaign period."

5) The authors often use parentheses to communicate pairs of results with one value named followed by the second in parentheses followed later by the value of the first and the value of second in parentheses. While this can often be understood it sometimes conflicts with grammatical use of parentheses and in general creates confusion.

Response:
We revised corresponding passages.

Specific Comments

P2 L3 "different atmospheric parameters" is rather vague here, this work deals with "absorbers" and "scatterers" along the light path.

We appreciate the reviewer's comment, however it is now obsolete for the abstract, since reviewer 1 suggested to completely remove the paragraph. A similar sentence appears in the introduction. There, it was corrected.

P2 L15 "intensity" here can be misleading in the context of radiation measurements "magnitude" is unambiguous

Done

P2 L22 "… were found to not necessarily being comparable quantities," this is not grammatical, nor is it fully clear what the authors wish to communicate here. The authors compare these quantities and find they must use the PAC. The final paragraph of the abstract should be reworded and expanded, particularly to reflect point 2 above.

The whole paragraph was revised, also on request of reviewer 1. It now reads:

"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 smoothing effects, namely biases arising from the reduced sensitivity of MAX-DOAS observations to higher altitudes and associated a priori assumptions."

Related statements in the main text were adapted accordingly.

P3 L12 "oxygen collision complex" should instead be "oxygen collision induced absorption", a formal complex is unnecessary to explain the absorption and has not been demonstrated to exist in the atmosphere.

Done

P3 L15-16 consultation of the values reported in Kreher et al., suggests that the average full aperture is closer to 20 mrad than 10 mrad.

This is true regarding the instruments participating in the CINDI-2 campaign. Yet, for MAX-DOAS profiling applications typically a smaller FOV of <= 10 mrad is desired. As a compromise we wrote "*10-20 mrad*".

P3 L26 I assume that "Arnoud et al., 2019 in prep." here and elsewhere is the same work as Apituley et al., 2019 in prep. referred to in Kreher et al., this reference should be updated or eliminated.

We like to thank the reviewers for pointing this out and updated the reference to "*Apituley et al. 2020 in prep."*

P3 L32 Same as previous comment, Wang et al., 2019 in prep. is either no longer in preparation or is not from 2019. This should be updated

Meanwhile Wang et al. is under review at AMTD. The reference was updated accordingly.

P4 Fig1 The map on the right appears to be oriented with North on top, however, this should be marked for clarity. Notably, based on the position of the river in the photo on the left the orientation of the panels is rotated by ≈180° rotation of the map would improve clarity.

A mark for indicating north direction was added to the map.

P5 L10 see comment above, based on Kreher et al., the FOV is smaller than the elevation angle resolution, but hardly negligible.

Changed from "*the telescope's FOV is usually negligible compared to the elevation angle resolution*" to "*ideally the telescope's FOV is negligible compared to the elevation angle resolution*"

P5 Eq1 The use of λ to denote wavelength is not introduced here or previously

We changed the text from: "*The very initial data in the MAX-DOAS processing chain are spectra of scattered skylight $I_\lambda(\alpha)$ […]*"

To: "*The very initial data in the MAX-DOAS processing chain are intensities of scattered skylight $I_\lambda(\alpha)$ at different wavelengths λ […]"*

P5 Eq1 This equation is not valid unless the contributions $\sigma_{i,\lambda}*S_i(\alpha)$ are summed over the set of contributing absorbers indexed i.

We agree with the reviewer, the sum was inserted.

Instead of:

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

We now have:

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

P5 Eqs2-3 τ_λ in Eq 2 is not the same quantity as τ_λ in Eq 1 and this fact is critical to the validity of Eq 3. This should be reflected by a consistent system of symbols.

We changed "τ_λ" to "$\Delta\tau_\lambda$"

P6 L14 DSCDs are reported for five data products, however the UV and Vis retrievals of O4 and NO2 retrieve the same chemical species.

We made this clearer by changing the text from:

"DSCDs were provided for five species, namely O_4 UV, O_4 Vis, HCHO, NO_2 UV and NO_2 Vis, where "UV" and "Vis" indicate different DOAS spectral fitting ranges in the ultraviolet and the visible spectral region, respectively (see Table 1)"

To:

"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)".

P6 L24-25 Algorithmically the retrievals are minimizing a cost function as stated at the end of the sentence, this is what the "model parameters are optimized to obtain", "maximum agreement" is not strictly the same as "minimum difference" and should be substituted.

We changed the text from: "*To retrieve a profile from the measured dSCDs, the model parameters are optimized to obtain maximum agreement between the simulated and measured dSCDs by minimising a pre-defined cost function.*"

To: "*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.*"

P7 L2 The solutions obtained for the underconstrained problem are not unambiguous. In the case of OEM they are a maximum likelihood estimator predicated on the *a priori* information. Even if *a priori* information is perfect the obtained solution is not unambiguous simply the most likely. The authors should use a different word.

We changed the wording, see our answer on the comment below.

P7 L2-7 *a priori* information is more extensive than the *a priori* profile proper, it also includes the covariance matrix for OEM. This does more than "fill" the lack of information it also defines a portion of the cost function and forms the basis by which likelihood is assessed. This is critical background to understanding the path-dependent results the authors find and should be expanded upon.

The corresponding paragraph was revised, also considering the comments by reviewer #1. It now reads:

"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."

Also we extended some formulations throughout the manuscript, e.g. P13L29: *"At higher altitudes, OEM retrieval results are drawn towards the a priori profile (according to the definition of the cost-function, see Rodgers [2000])"*

For the very details of OEM the reader is encouraged to refer to the corresponding literature.

P7 L33 the aerosol profiles are "extrapolated" not "interpolated"

Done.

P8 L8-9 The definition of the *a priori* covariance as defined here is a predicate to the later findings and should be discussed as such in relevant locations.

Corresponding passages were revised. The importance of the choice of the a priori covariance is emphasized at relevant locations and the definition in P8 L8-9 is referenced.

P11 L18-20 If I understand correctly, this method of processing gives a large weight to the uppermost one or two measurements available as these measurements define a majority of the relevant layer. Can the authors comment or elaborate?

We agree with the reviewer. To make this point clearer we added a sentence very similar to the reviewer's comment: "*Note, that this approach gives a large weight to the uppermost measurements, as they are representative for the majority of the relevant layer.*"

P12 L8 temperature and pressure should be spelled out here.

Done.

P12 L9 Wagner et al., (2019) find effects of up to 7% on the modeled O4 profile when using a standard atmosphere. This could be a significant contributor or the retrieved RMSD, can the authors comment?

This is an aspect that we omitted so far. We did further investigation on this, with the results being summarised in the Supplementary material as follows:

S7 Impact of the choice of pressure and temperature profiles for the RTMs

Pressure (p) and temperature (T) profiles used for the RTMs within this study are averaged sonde measurements performed in De Bilt by KNMI during September months of the years 2013-2015 (see main text Sect. 2.1.3). To estimate the effect of this approximation on the results, IUPHD/ HEIPRO retrieved an additional set of profiles, using p and T information from radiosondes launched at KNMI (De Bilt) during the campaign. Between one and three sondes were launched every day except on 16 September. For each profile inversion, the temporally closest sonde observation was used. Table S7 shows the difference in RMSD and Bias magnitude between these results and the "standard" results of IUPHD/ HEIPRO (that used the prescribed averaged p and T profiles from years before) relative to the average RMSDs and average Bias magnitude for all participants.

The impact on the dSCD comparison is less than 5% for both, RMSDs and Bias magnitudes. For AOTs, VCDs and surface concentrations, significant improvement ($> 10\,\%$ in RMSD) is only observed for HCHO surface concentrations ($17\,\%$) that contrasts with a deterioration for UV AOTs by 13%. The average improvement in RMSD for AOTs, VCDs and surface concentrations is $3.2\,\%$. The overall consistency between MAX-DOAS and supporting observations can thus be considered to remain similar, despite larger changes in some Bias magnitudes are observed (up to $51\,\%$ improvement for NO_2 Vis surface concentrations and up to $20\,\%$ deterioration for UV AOTs).

Table S7. The differences in RMSDs and Bias magnitudes for the IUPHD/ HEIPRO results arising from using daily p and T profiles, relative to the average RMSDs and Bias magnitudes assessed within the main study. Values are given for the comparisons of modelled and measured dSCDs ("dSCDs") and the comparisons against the supporting observations of AOTs, VCDs and surface concentrations as described in the main text. Minus signs indicate improvement. Only clear sky conditions were considered.

	dSCDs		AOT/VCD		Surface	
	ΔRMSD [%]	ΔBias [%]	ΔRMSD [%]	ΔBias [%]	ΔRMSD [%]	ΔBias [%]
HCHO	2.7	3.5	6.8	10.5	-17.4	-22.0
NO_2 UV	-0.7	-1.1	-2.7	-2.6	-3.5	8.7
NO_2 Vis	-0.7	-3.3	-0.8	-1.0	-2.8	-50.9
Aerosol UV	-0.7	0.7	12.5	20.2	-	-
Aerosol Vis	-0.2	2.1	-8.7	-40.1	-	-

These findings are also briefly discussed in the conclusions now.

P12 L20-25 Is the least-squares regression a minimization of vertical distance or orthogonal distance?

The vertical distance is minimised. This information was added during the course of the revision of Sect. 2.3.1.: *"For the linear regression analysis, the vertical distance between the model and the data points is minimised […]"*

P12 Eq7 1/Np here should be in parentheses for clarity

Instead of adding parentheses we changed the formatting to achieve a similar effect.

We changed: $$\sigma_{arms,p} = 1/N_P \sum_p \sigma_{rms,p}$$

$$\sigma_{arms,p} = \frac{1}{N_P} \cdot \sum_p \sigma_{rms,p}$$

To:

P14 L24 replace "not given" with "inaccurate"

Done.

P15 L1-2 "Aij describes the sensitivity of the measured concentration in the ith layer to small changes in the real concentration in the jth layer.,"

Done.

P15 Eq11 The coefficient of 12 in this equation seems to be the result of summing over the lowest 12 layers, corresponding to 2.5 km. However, this is not stated.

The spread is calculated considering the cross sensitivity to each layer. The coefficient of 12 is a normalisation factor which is part of the original definition of the "spread" (see Rodgers, 2000, as cited in connection with Eq. 11 in the manuscript). Initially we thought it might be helpful to find some simple measure for the retrieval's spatial resolution and show it in the plots. However, as the spread does not provide any substantially new information to the reader and might rather be misleading than helpful (see also the reviewer 's comment on Fig. 2 below) we decided to completely remove it from the text and the plots.

P15 L16-18 The increase in information content reflects an increase in the differential light path specifically. While this follows from the longer light paths overall, it is the increased differential path which is the source of the information.

We replaced "*light path*" by "*differential light path*"

P16 Fig 2. The symmetric boxes illustrating are misleading. As the AVK traces demonstrate, the information content moves as well as being "smoothed". The boxes should be centered in a more rational way or else eliminated.

As explained above (comment on P15, Eq11), the boxes in the plots and corresponding paragraphs on the "spread" in the main text were eliminated.

P17 Table 2 Most groups are listed by city, however, Anhui is listed by province, should this not be Hefei?

We changed this to "Hefei". Further similar issues in the same table were also fixed:

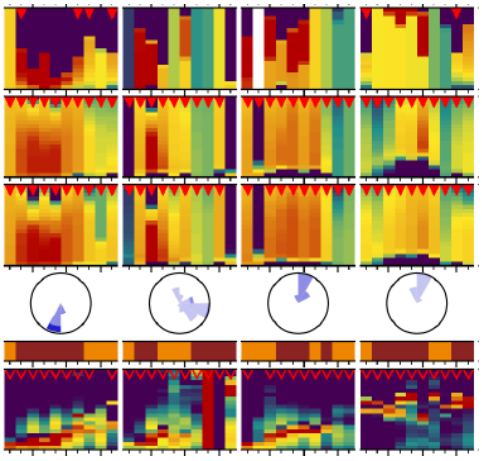
"Department of Physics, University of Toronto, Toronto, Canada" → "Department of Physics, University of Toronto, Canada"

"NASA-Goddard, Greenbelt, Maryland" → "NASA-Goddard, Greenbelt, United States"

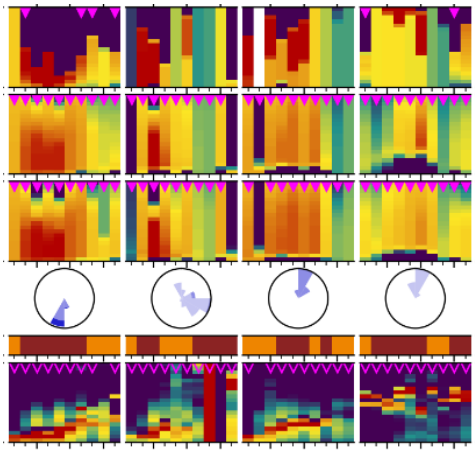
Figs. 3-7. The red triangles are not readily seen against the color scale.

We changed the colour of the triangles to pink, which is not ideal either but was the colour we consider best distinguishable from the colour scale in the background:

Submitted version of the mansucript:



Now:



Figs. 6-7 In the bottom row when only surface measurement are available these are almost imperceptible.

We agree. However, we do not see how to change this without introducing potentially confusing features. Please note, that the figures the reviewer refers to are meant to provide an at best complete overview of the available datasets for qualitative comparison and that data of this extent and inhomogeneity are challenging to visualise. Further note that the same data appears again in the following sections in more detailed plots which are easier to read. This is why we finally decided to leave them as they are.

P24 L6 what precisely do the authors mean by "update interval of the jacobians"?

In optimal estimation algorithms (where the model parameters are iteratively adapted), one of the computationally most expensive steps is to derive the jacobians of the simulated dSCDs w.r.t. the model parameters. Typically, inversion problems of the kind discussed in the manuscript are moderately linear and do not require a recalculation of these jacobians in each iteration to achieve convergence. This is used by some algorithms to save computing time. The impact of this "shortcut" on the final results depends on the atmospheric scenario, on the exact implementation and the settings defined by the user.

We replaced the text in brackets *"(e.g. number of iteration in the inversion, accuracy criteria for the RTMs, update interval of the jacobians, ...)"* by:

"*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."*

P24 L6-7 Are the larger discrepancies not simply a reflection of the greater DOFS?

This is well possible and also stated in just the following sentence: *"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"*.

P24 11-13 In this section while using the same set of dSCDs how can the authors speak to horizontal inhomogeneity? How would such an inhomogeneity be detected?

The idea was, that inhomogeneity leads to less stable solutions, making the algorithms more sensitive to differences in the inversion settings. But this might indeed be too far fetched to be mentioned here. We therefore removed the sentence: "*Horizontal inhomogeneities are an unlikely reason because the worse performance in the Vis was also apparent in the study by Frieß et al. (2019) with synthetic data, where horizontal gradients were non-existent*."

P24 L28 Can the authors clarify what they mean by "technical problems" do they think there was some error in the implementation of the protocol?

Yes this could have been the case. Or that improper/different retrieval settings were applied as it was the case for Heipro, where discrepancies between IUPHD and UTOR could be explained by different numbers of applied iteration steps. The paragraph was rearranged and revised. Amongst others we removed the statement with the "technical problems" and now "suspect similar reasons" as for the IUPHD <-> UTOR discrepancies.

Before:

"An example for large discrepancies between participants using the same algorithm is AUTH aerosol in the UV, where in contrast to other bePRO users oscillations seem to appear. We suspect this to originate from technical problems which could not yet been identified. The discrepancies between IUPHD and UTOR (both using HEIPRO) were found to mainly be caused by differences in the number of applied iteration steps in the Levenberg-Marquardt optimization scheme during aerosol retrieval. IUPHD (UTOR) applied 20 (5) iterations. The consequences are evident throughout the comparison."

Now:

"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 20 and 5 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."

Figs. 8-12 If there are uncertainties in these graphs as indicated by the legend for Fig 8, they cannot be seen.

We agree. We reduced the edge width of the markers to improve this. Still they are only visible when looking very closely at data points lying apart from the main point cloud. Anyway we decided to keep them as they at least give an impression of the uncertainties' order of magnitude.

P28 L3 As stated above, per the results presented signals aloft can be reliably detected, but not reliably located and/or quantified. Language should be edited to reflect this.

We changed: *"[…] cannot be reliably detected […]"*

To: *"[…] cannot be reliably located and quantified […]"*

Similar statements were adapted throughout the manuscript.

P28 L13-15 On first reading the finding that adjusting MAX-DOAS AOT by the ratio to the sun photometer improves the agreement seems obvious, even tautological. The actual processing as described in the supplement needs to be better reflected in the main text.

We agree that it is strange to emphasize the PAC all over the manuscript to finally show the results in the Supplement. Therefore, we embedded Supplement S2 into the main text Section 3.4.

P29 L3-4 The authors state "even though the physical reason for PAC and SF are different." This is surprising as it suggests that the authors posit a specific physical reason for SF which is not that for PAC, what is this?

We agree with the reviewer corrected this statement regarding the "physical reason", as it is not well-founded. We replace the sentence by:

"[…] even though the motivation for the application of the PAC and the SF are different."

The motivations are in fact very different: the application of the PAC is necessary solely for mathematical reasons related to the concept of optimal estimation and prior constraints applied therein. In contrast, the prominent publications motivating/discussing the application of an O4 scaling factor (Wagner (2009), Clémer (2010), Ortega (2016) and Wagner (2019)) forward modelled O4 dSCDs (using an atmosphere derived from supporting observations like Lidars) to measured O4 dSCDs. They do not make use of optimal estimation or a priori profiles similar to those used in our study. Thus their findings are independent from any kind of PAC.

We added a corresponding explanation to the same paragraph:

"[…] 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 (Wagner, 2009; Clémer, 2010; Ortega, 2016; Wagner, 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. They do not make use of optimal estimation or prior constraints similar to those used in our study. Thus their findings can be considered independent from any kind of PAC."

And to the paragraph above:

"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."

Fig. 13 and other Figs following same format. In the top row, why are the scatters plotted on an inverted axis? Cannot the scatter exceed one? Even quite significantly? Here and elsewhere the hashed and solid shading are not readily distinguishable.

We agree, that this was not a good solution. We inverted the axis back to the normal direction. Further we adapted the figure to make a distinction between hashed and solid areas unnecessary.

Example of the updated plot:

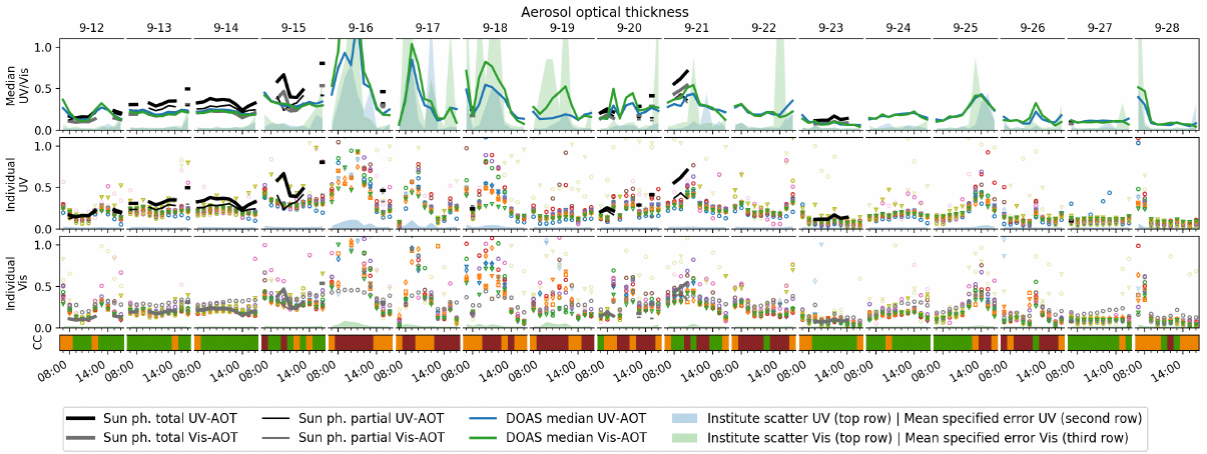
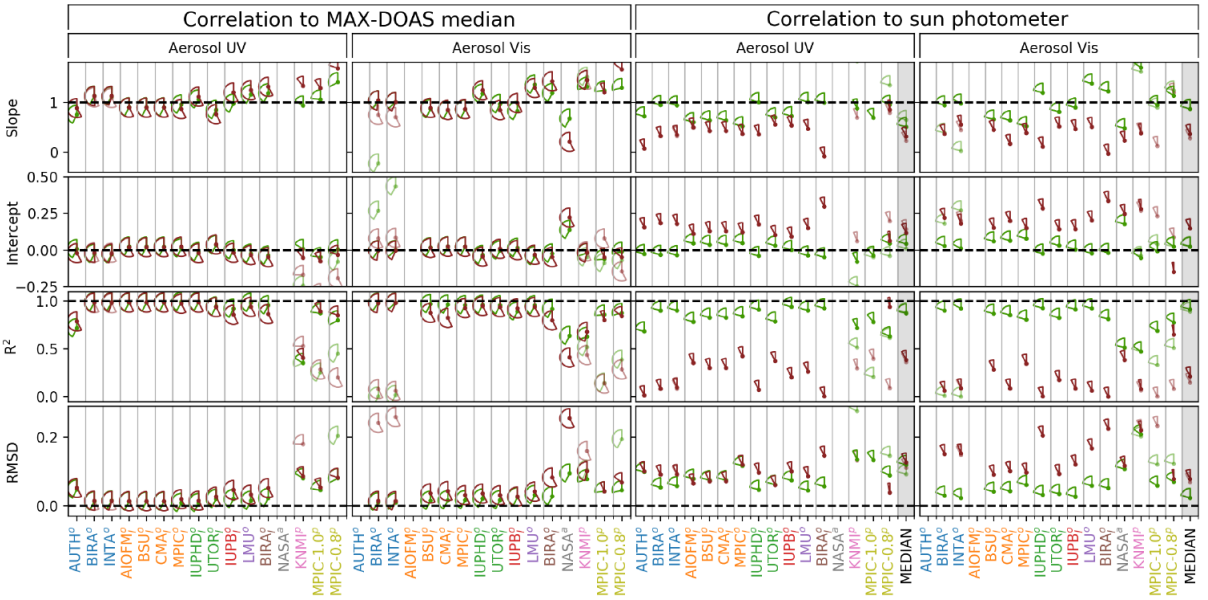


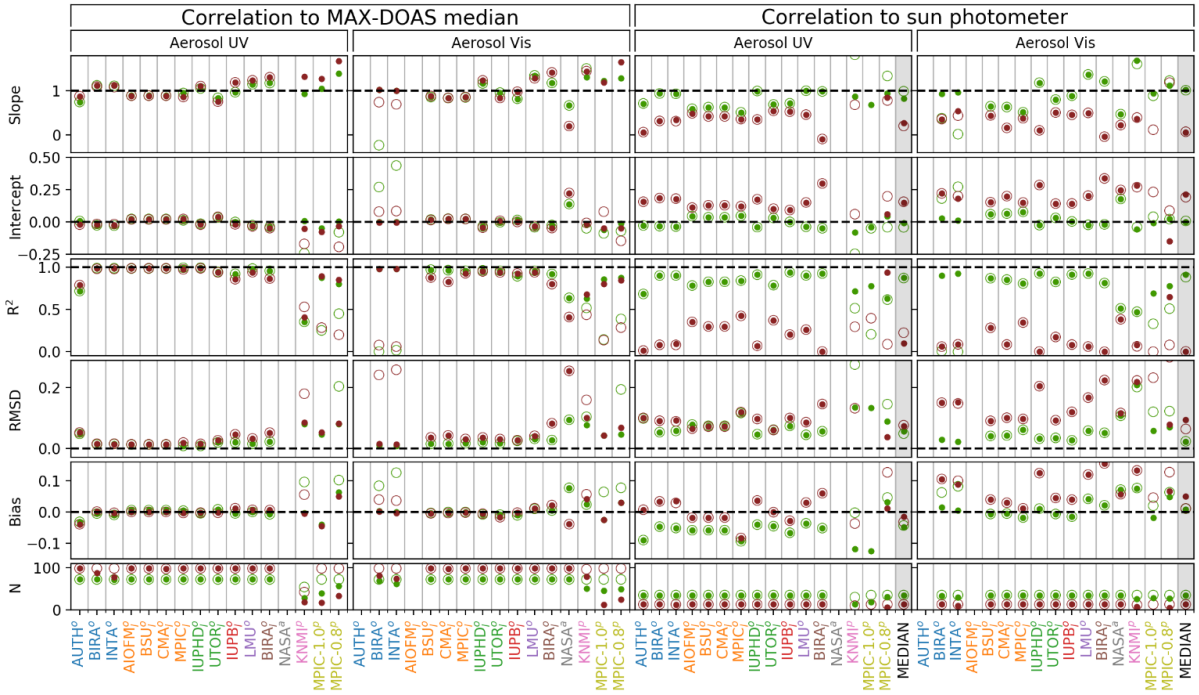
Fig. 14 and other Figs following same format. While I can appreciate what the authors are trying to communicate with the pie chart symbols, the clear and cloudy data are drawn from the same total and the symbols repeat within a given column. This should be simplified in some way.

We thank the reviewer for this suggestion, which makes the figures much easier and more comfortable to read. We discarded the pie chart symbol and added another thin row of plots indicating the number of used profiles for each of the columns.

Submitted:



Now:



P31 L9-12 This paragraph in particular demonstrates that aerosol aloft are detectable.

We partly agree. The detection of aerosol aloft is at least limited. However, as stated above, we revised text passages stating that aerosol aloft are undetectable.

P31 14 The first sentence should be reworded, the VCDs are compared to different standards or "assessed", but the NO2 VCDs are not compared to the HCHO VCDs

We changed: *"This section compares the VCDs of HCHO and NO_2."*

To: *"This section assesses the consistency of the VCDs for each of the trace gases HCHO and NO_2".*

Fig. 15 where is the outlier referred to on P31 L21?

By "outlier" we refer to a radiosonde profile here, these are not shown in Fig. 15. In the case of this "outlying" profile, the NO_2 concentrations were close to the radiosonde detection limit and instrumental offsets made it unsuitable for the corresponding study, which was to show whether a correction similar to the PAC might be necessary also for NO_2 VCDs. However, "outlier" is probably not the right word to use here.

To make things clearer we changed the text: "*Ignoring an outlier on 09-27 07:00:00, where NO_2 concentration was close to the radiosonde detection limit, […]"*

To: "*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), [...]*"

P33 L13-14 the LP-DOAS data are described as "very accurate, representative, and complete" while these are likely well supported assessments, such strong statements should be demonstrated or else backed up by a citation.

This statement is already justified in Section 2.2.5, where the LP-DOAS setup at CINDI-2 is introduced. We added a cross reference to this section.

"*Very accurate*" is supported by multiple references there: Pöhler et al., 2010; Merten et al., 2011; Nasse et al., 2019. We added Pikelnaya et al., 2007 to further support this statement.

"Representative", since its light path covers the lowest MAX-DOAS retrieval layer fully and exclusively.

"Complete" since it provides a near-continuous dataset over the campaign period.

Fig 19. Sondes are not listed in the legend. Here and elsewhere the color of the lidar and sondes is very challenging to distinguish.

We like to thank the reviewer for pointing out this omission, we added radiosondes to the legend. Further, we brightened the orange color and darkened the red color which are used to visualize NO_2-Lidar and radiosonde data throughout the paper.

P34 L3 The language here should be more precise. The surface concentration does reflect the ability of MAX-DOAS retrieval to isolate the surface layer specifically. However, the isolation and resolution of the surface layer does not imply in and of itself the resolution of the vertical profile above it.

We agree that this could be misleading.

We changed: *"[…] the surface concentration comparison also reflects the MAX-DOAS' ability to actually resolve vertical profiles, as it requires an isolation of the surface layer from the layers above."*

To: *"[…] 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."*

P35 L5-7 How the consistency of the surface concentrations point to a problem in the direct sun data? Is it not equally possible that the MAX-DOAS VCD apart from the lowermost layer are flawed?

Yes, we agree with the reviewer. We changed the text: *"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 direct-sun data."*

To: *"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"*

P35 L10-11 I believe this final sentence refers to the comparisons in Tables S4 and S5, however, that is not clear in the text.

The sentence refers to Fig.18 (HCHO time series) and Fig. 19 (NO2 time series), where in the top row the scatter among the participants and in the two lower rows the specified uncertainties of the MAX-DOAS observations are indicated by the faint areas.

To clarify this point, we changed the text: "*Again the scatter to the MAX-DOAS median even for clear-sky conditions are similar or larger than the specified errors (factors of about 1, 2 and 3 for HCHO, NO_2 UV, NO_2 Vis, respectively)."*

To: *"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 NO2 Vis, see Fig. 19 and Fig. 20)"*

Further, we added a sentence to the caption of Fig. 19: "*Note, that the mean specified uncertainties in the two lower rows of the figure are very small and thus barely visible."*

P36 L1-4 Can this thinking be made more quantitative by reference to the fτ for the Vis and UV products?

This point became obsolete, since the whole section was removed as suggested by reviewer 1.

In the supplement:

P2 L18 the shift to lower altitudes is a simple reflection of the construction of the covariance. This is hinted at on L21, but should be spelt out. As constructed the retrieval does not have uncertainty into which to place the information at higher altitudes, but the information is present in the measurements and is placed at an altitude which is accessible within the constraints of the prescribed covariance.

This comment explains the issue accurately and concise. We adopted the reviewer's wording:

We changed: "However, a part of the high-altitude aerosol appears to be shifted to lower altitudes here by the retrieval."

To: "However, 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."

P4 L12-14 Clear-sky O4 dSCD are not the largest possible, if there is small but non-zero aerosol scattering concentrated at altitudes below the median altitude of photon scattering for a relevant geometry this leads to brightening. Hence why aerosol can appear as increased albedo for satellites.

The reviewer is correct here, our statement is wrong. Note, however, that the sentence refers to low aerosol clear sky scenarios, where this assumption is nearly fulfilled.

We therefore changed the text: "*Finally, Wagner et al. (2009) reported, that under low aerosol conditions, measured dSCDs sometimes even exceed dSCDs modelled within an aerosol free atmosphere, where O4 dSCDs are expected to be the largest possible (regarding clear-sky scenarios only).*"

To: "Also, *Wagner et al. (2009) reported that, under low aerosol conditions, measured dSCDs sometimes even significantly exceed dSCDs modelled within an aerosol free atmosphere, where O4 dSCDs are close to the largest possible (regarding clear-sky scenarios only).*"

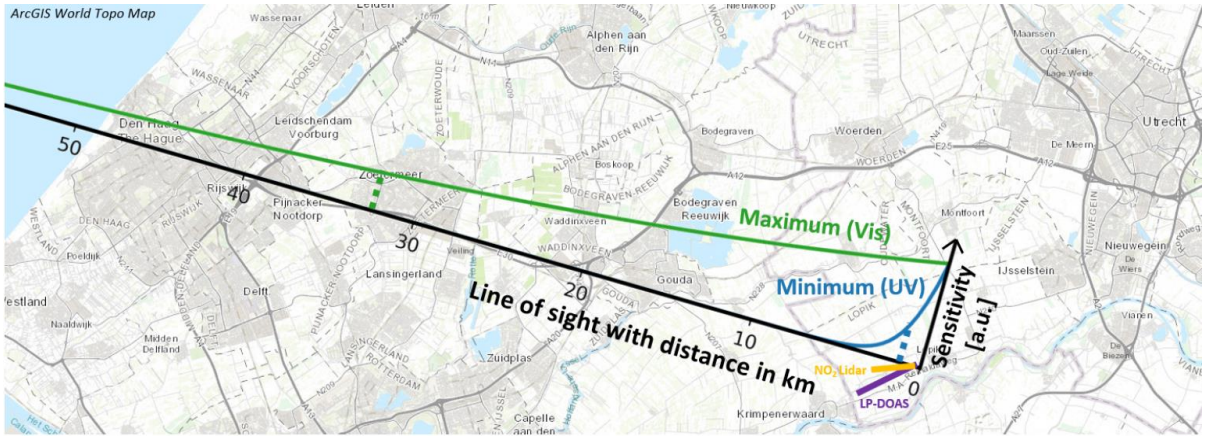
Fig S11 The color scheme makes this figure very difficult to read.

This problem was solved by changing the colors for radiosondes and NO2-Lidar throughout the paper.

Fig S12 The distance scale in this figure seems somewhat misleading in light of Fig. S13. The provided exponential curves appear to imply a radical difference in ranging between the Vis and UV, whereas Fig. S13 makes clear that changes in atmospheric conditions are responsible for most of the difference.

We changed the figure by showing the average, minimum and maximum sensitivity range for UV and Vis, respectively:

Submitted version of the manuscript:



Now:

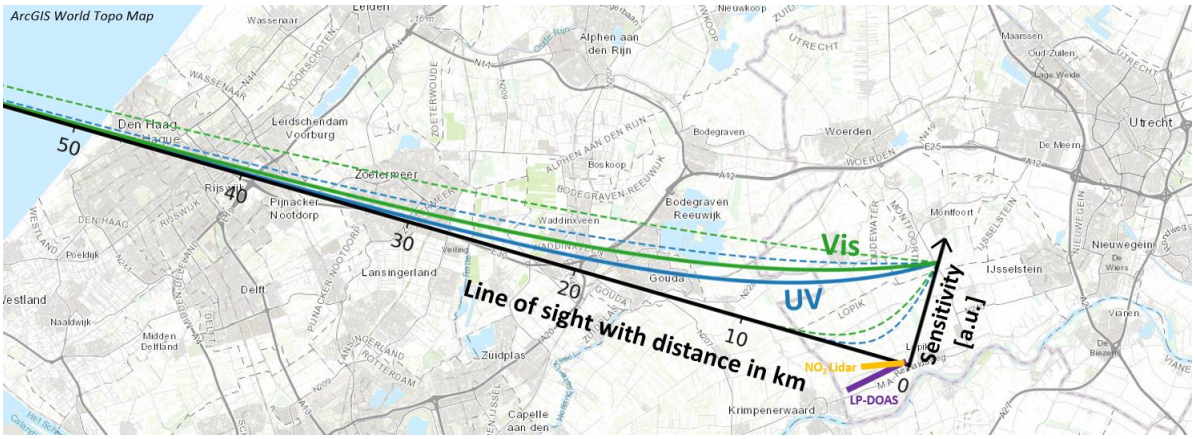


Fig. S34 If I understand this figure correctly virtually all data are within two standard deviations, is this not as expected. P33 L6-7 seems to imply something unexpected.

The word “indeed” is misleading here. Further, a short conclusion on the actual meaning of this study is missing.

We changed the text: “*Figure S34 shows histograms of the calculated differences. An estimate of the impact of smoothing on the retrieval results is actually provided by the OEM retrievals themselves as the "smoothing error". The specified smoothing errors are also indicated in Fig. S34 and indeed slightly larger than the standard deviation observed in in this test.”*

To: “*Figure S34 shows histograms of the calculated differences. The standard deviation is about 5x10^{9} molec. cm^{-3} which is only about 10 % of the total average RMSD between MAX-DOAS and LP-DOAS observations. An estimate of the impact of smoothing on the retrieval results is actually provided by the OEM retrievals themselves as the "smoothing error". The specified smoothing errors are also indicated in Fig. S34 and are similar to the standard deviation observed in in this test, meaning that for the surface layer they are well representative for the real impact of smoothing.”*

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. ~~Multi-AXis Differential Optical Absorption Spectroscopy (MAX-DOAS) is a well-established ground-based measurement technique for the detection of aerosols and trace gases particularly in the boundary layer and the lower troposphere: ultraviolet- and visible radiation spectra of skylight are analysed to obtain information on different atmospheric parameters, integrated over the light path from space to the instrument. An appropriate set of spectra recorded under different viewing geometries ("Multi-Axis") allows retrieval of tropospheric aerosol and trace gas vertical distributions by applying numerical inversion methods.~~

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 ~~others~~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~~. The profiles~~ , 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 ~~intensity~~magnitude. Under clear sky conditions, the root-mean-square differences ~~of aerosol optical thicknesses,~~ (RMSDs) among the results of individual participants vary between $(0.01 - 0.1)$ for AOTs, $(1.5 - 15) \times 10^{14}$ molec cm^{-2} for trace gas (NO_2, HCHO) ~~vertical columns and surface concentrations among the results of individual participants vary between $0.01 - 0.1$, $(1.5 - 15) \times 10^{14}$ molec cm^{-2} and~~ VCDs and $(0.3 - 8) \times 10^{10}$ molec cm^{-3} ~~, respectively.~~ for trace gas surface concentrations. These values compare to approximate average optical thicknesses of 0.3, trace gas vertical columns of 90×10^{14} molec cm^{-2} and trace gas surface concentrations of 11×10^{10} 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, ~~these values increase to~~ the RMSDs increase to $(0.02 - 0.2)$ against AOTs from the sun photometer, $(11 - 55) \times 10^{14}$ molec cm^{-2} against trace gas VCDs from direct-sun DOAS observations and $(0.8 - 9) \times 10^{10}$ molec cm^{-3} ~~. It is likely that a large part of this increase is caused by imperfect~~ 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 ~~overlap of the different observations~~ mismatch among the observations and simplified assumptions particularly on aerosol optical properties made for the MAX-DOAS retrieval.

~~In contrast to what is often assumed, the~~ 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 ~~extinction profiles and the sun photometer total aerosol optical thickness were found to not necessarily being comparable quantities, unless information on the real aerosol vertical distribution is available to account for the low~~ 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 smoothing effects, namely biases arising from the reduced sensitivity of MAX-DOAS observations ~~at higher altitudes .~~ to higher altitudes and associated *a priori* assumptions.

Copyright statement. TEXT

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 ~~determined by gas and particulate matter~~ driven by the exchange with the surface~~and also driven by~~ , 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 ~~well-established~~ widely used ground-based measurement technique for the detection of aerosols and trace gases particularly in the ~~PBL and the lower free~~ lower troposphere: ultraviolet (UV)- and visible (Vis) ~~radiation~~ absorption spectra of skylight are analysed to obtain information on different atmospheric ~~parameters, integrated along~~ absorbers and scatterers, integrated over the light path (in fact a superposition of a multitude of light paths)~~from the top of the atmosphere (TOA) to the instrument~~. 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). ~~Detectable gases~~ 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 ~~complex~~ 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 ~~concentration~~ 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 ~~transport~~ transfer models. These profile retrieval algorithms are the subject of this comparison study.

Today, there are numerous ~~such~~ 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 ~~with respect to different conditions~~ 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, $51.9676\,^\circ$ N, $4.9295\,^\circ$ 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 ~~others~~meteorological instruments) for validation. This study compares MAX-DOAS profiles of NO_2 ~~, HCHO and aerosol extinction (~~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. ~~The results are compared with each other and validated against CINDI-2 supporting observations.~~ 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.

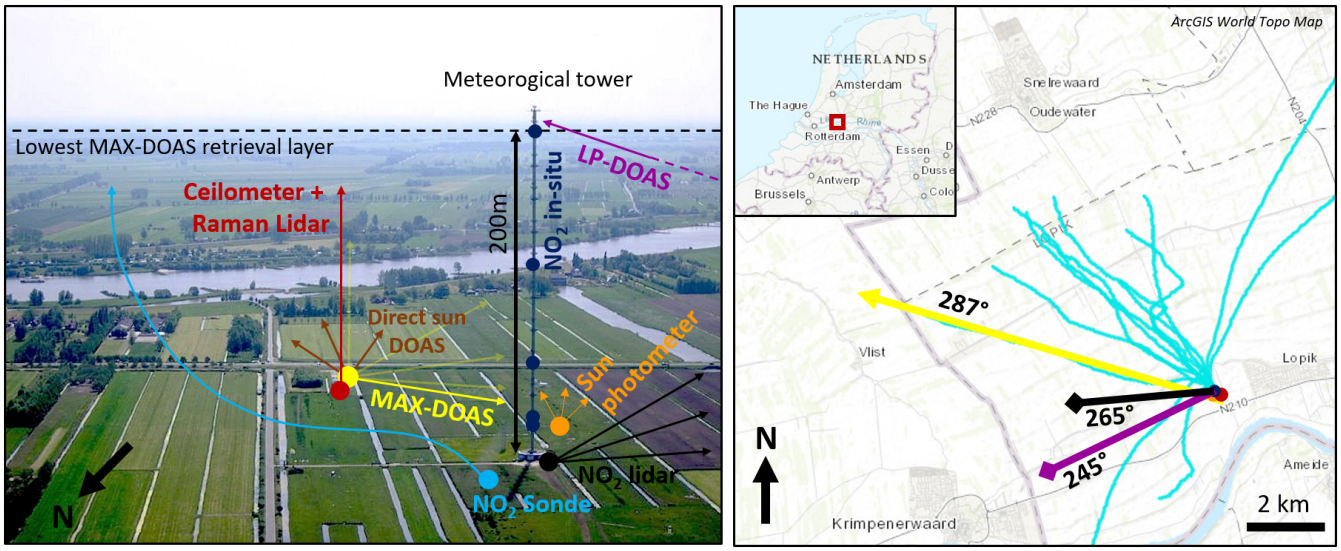


Figure 1. Left: Image of the CESAR site with position and approximate viewing directions of the MAX-DOAS instruments and supporting observations of relevance for this study. Right: Map (Esri et al., 2018) with instrument locations, viewing geometries and sonde flight paths indicated.

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 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 ~~spectra~~ 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 nm) recorded under different viewing elevation angles α (ideally the telescope's FOV is ~~usually~~ 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

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

$S_i(\alpha)$ is the slant column density (SCD), which is the trace gas concentration integrated along l. C represents ~~further~~ 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

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

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

$$\Delta S(\alpha) = S(\alpha) - S(\alpha = 90°) \quad (3)$$

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., 2009)~~ (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~~ 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 ~~five~~ three chemical species, namely O_4~~UV, O_4 Vis, HCHO,~~ , NO_2 ~~UV and~~ and HCHO. O_4 and NO_2 ~~Vis, where "UV" and "Vis" indicate different DOAS~~ were each provided for two different spectral fitting ranges~~in the ultraviolet~~ , in the ultra-violet (UV) and the visible (Vis) spectral region, ~~respectively~~ 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.

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	336.5 - 359
NO_2 Vis	Number concentration [$molec\,cm^{-3}$]	NO_2 Vis	425 - 490
HCHO	Number concentration [$molec\,cm^{-3}$]	HCHO	336.5 - 359

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, ~~aerosol microphysical properties, ...~~water vapour and aerosol properties) attributed to each layer. This model atmosphere is then used to simulate MAX-DOAS dSCDs under consideration of the viewing geometries. To retrieve a profile from the measured dSCDs, the model parameters are optimized to ~~obtain maximum agreement~~ minimise the difference between the simulated and measured dSCDs ~~by minimising~~ based on a pre-defined cost function. ~~Typically only $p = 2$ to 4~~

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 ~~layers are underconstrained and *a priori*~~ independent retrieved parameters are ill-posed and prior information has to be assimilated to ~~obtain unambiguous solutions~~achieve convergence. For OEM algorithms, this is provided in the form of an *a priori* profile ~~(Rodgers, 2000) "filling" the lack of information which is most prominent at higher altitudes (see Sect. 3.1). Parametrized approaches achieve this~~ 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 ~~typically~~ 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). ~~The~~ Besides the algorithms described therein, our study includes results from the M^3 algorithm by LMU~~appears as an additional algorithm in our study~~. 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 ~~(Wagner et al., 2019; Ortega et al., 2016, and references~~ (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 ~~if smoothing effects, in particular the low sensitivity of MAX-DOAS observations to higher altitudes, are taken into account (see Sect. 2.3.2 and~~ (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 km altitude were averaged from O_3 sonde measurements performed in De Bilt by KNMI during September months of the years 2013-2015. ~~The surface albedo was fixed to 0.06, according to ?.~~ A fixed altitude grid was used for the ~~retrieval~~inversion, consisting of ~~20 layers between 0 and 4 km~~ 20 layers between 0 and 4 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 ~~be directly implemented~~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 m) layers close to the surface, increasing with altitude) and farther extending (up to 40 to ~~this grid to simplify the comparison~~90 km altitude) grids, inherently defined by the

individual retrieval algorithms. Surface and instruments' altitudes were fixed to ~~0 m~~0 m, which is close to the real conditions: the CESAR site and most of the surrounding area lie at ~~0.7~~ 0.7 metres b.s.l., whereas the instruments were installed at ~~0 to 6 m~~ 0 to 6 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 nm were interpolated to 343 nm~~ 360 nm were extrapolated to 343 nm using the mean Ångström exponent for the ~~440-675 nm~~ $440 - 675\,\mathrm{nm}$ 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 nm~~ 440 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 m, respectively.

2.1.4 Requested dataset

All participants were requested to submit the following results of their retrieval: ~~(1~~i) Profiles and profile errors, optionally with errors separated into contributions from propagated measurement noise and smoothing effects. ~~(2~~ii) Modelled dSCDs as calculated by the RTM for the retrieved atmospheric state. ~~(3~~iii) Averaging Kernels (AVKs) for assessment of information content and vertical resolution (only available for OEM approaches). ~~(4~~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 ~~profiles~~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 ≈ 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, ~~an enhanced uncertainty of 0.04 is assumed due to temporal and spectral inter-/extrapolation~~the proposed accuracy of 0.02 was enhanced by the variability between them (typically on the order of 0.008).

2.2.2 Aerosol ~~extinction~~ profiles

Information on the ~~true aerosol extinction (AE) profiles~~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 ~~(in the following referred to by "extinction profiles") by~~ 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 ~~is ≈ 0.03~~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 ~~extinction~~ 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 ~~(DS-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 Pöhler et al., 2010; Merten et al., 2011; Nasse et al., 2019)~~(LP-DOAS, see Pikelnaya et al., 2007; Pöhler et al., 2010; Mer . 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, ~~T/P – profiles, ...~~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 ~~correlation analysis (weighted least-squares regression) and~~ 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 ~~RMSDs as in contrast to correlation~~ 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 ~~, RMSD is representative for both, statistical and systematic deviations. For~~ are provided and consulted for a more differentiated view.

Consider two time series of length N_T: the retrieval result $x_{p,t}$ ~~and~~ of a participant p at time t and some reference observation $x_{ref,t}$ (~~each consisting of N_T data points, t and p indicating time and participant, respectively~~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 ~~given by~~ defined as:

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

~~For both, RMSD calculation and least square regression, contributing data points are weighted by the reciprocal of the quadratic sum of their uncertainties:~~ 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

~~w~~Bias:
$$\sigma_{bias,p} = \frac{1}{\sigma_{p,t}^2 + \sigma_{ref,t}^2} \frac{1}{N_T} \cdot \frac{1}{\sum_t w_t} \cdot \sum_t w_t \left(x_{p,t} - x_{ref,t} \right) \tag{7}$$

Sometimes the term "average RMSD" ("average Bias") is used, which refers to the average over the RMSD (Bias) values of the individual participants~~, hence~~

Average RMSD:
$$\sigma_{arms,p} = 1/N_P \sum_p \sigma_{rms,p}$$

~~with N_P being the number of included 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~~, hence:~~

Relative RMSD:
$$\sigma_{rrms,p} = \frac{N_T \sigma_{rms,p}}{\sum_t x_{ref,t}}$$

~~The~~ . 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~~is assessed by comparing the results of individual participants with the~~ , 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.

~~The~~ 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.

~~Inherent quality indicators for~~ An inherent quality indicator for the retrieval algorithms are the consistency of modelled and measured dSCDs~~and the consistency of NO_2 results retrieved in different wavelength ranges~~. 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. ~~The consistency of retrieval results of NO_2 in the UV and the Vis spectral ranges is another indicator for an algorithm's reliability since they should ideally yield the same results.~~

In a few cases (e.g. ~~Section 3.2~~Sect. 3.2, where full profiles are compared) the scatter among several participants p (of number N_P) and ~~potentially~~ 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 ~~$\bar{x}_{t,h}$~~ $\bar{x}_{h,t}$ 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 ~~or RMSD~~ 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 ~~$\approx 1\,\mathrm{km}$~~ $\approx 2\,\mathrm{km}$ typically cannot be reliably ~~detected~~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. ~~Hence,~~ 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 $3.5 \times 10^{10}\,\mathrm{molec\,cm^{-3}}$. This is indeed of similar magnitude as the average RMSD observed during the comparison (approx. $5 \times 10^{10}\,\mathrm{molec\,cm^{-3}}$). It shall further be noted, that under strong spatial variability the horizontal homogeneity assumed by the retrieval forward models is ~~not given. It was not possible to derive a reliable quantitative estimate of the impact on the comparison, but investigations on the NO_2 surface concentration in Supplement S6 and investigations by **?** indicate, that it significantly contributes to the residual RMSD observed between different observations.~~ 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 $\mathbf{A}$). ~~Each of its elements~~ 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. ~~A reasonably defined characteristic~~ The length of this decay (note ~~,~~that i can be converted to the corresponding altitude by multiplication with the retrieval layer thickness Δh) ~~can serve as a measure~~ is an indicator for the vertical resolution of the retrieval. ~~Here, the so called "spread" $s(i)$ was chosen as the characteristic length, as defined according to equation 3.23 in Rodgers (2000):~~

$$s(i) = 12 \cdot \Delta h \cdot \frac{\sum_j (i-j)^2 A_{ij}}{\left(\sum_j A_{ij}\right)^2}$$

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.

~~For all species~~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.

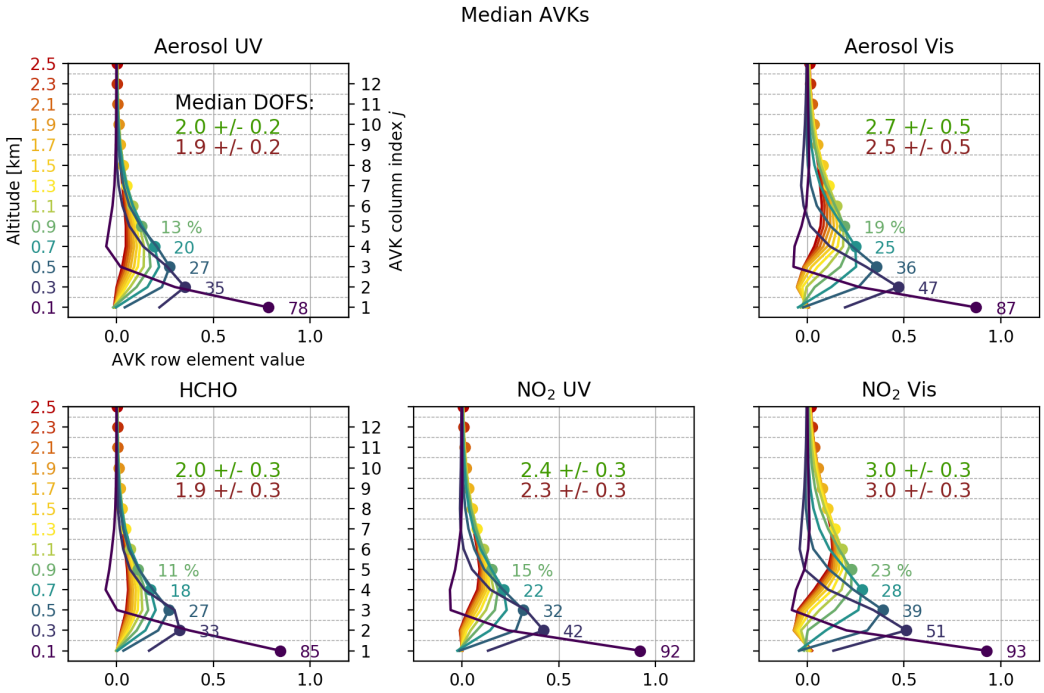


Figure 2. Mean AVKs for the retrieved species (median over participants, mean over time). Their meaning is described in detail in the text. Each altitude and corresponding AVK line A_i are associated with a colour, which is defined by the colour of the corresponding altitude-axis label. The dots mark 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).~~The vertical bars indicate the vertical resolution (the "spread", defined according to Eq. **??**) for the five lowest layers.~~

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, ~~Anhui~~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, ~~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)			
MMF	OEM[l]/ VLIDORT	Friedrich et al. (2019)	Royal Belgian Institute for Space Aeronomy, Brussels, Belgium	BIRA	
Realtime	ANA[a]/ -		NASA-Goddard, Greenbelt, ~~Maryland~~ 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 ~~transport~~ 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	$9 \cdot 10^{15}$ molec cm^{-2}
NO_2 Vis	460	$9 \cdot 10^{15}$ molec cm^{-2}
HCHO	343	$8 \cdot 10^{15}$ 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
~~UV vs. Vis retrieval N.A.[c]~~ **??** 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), ~~red~~ 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 ~~red~~ 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.

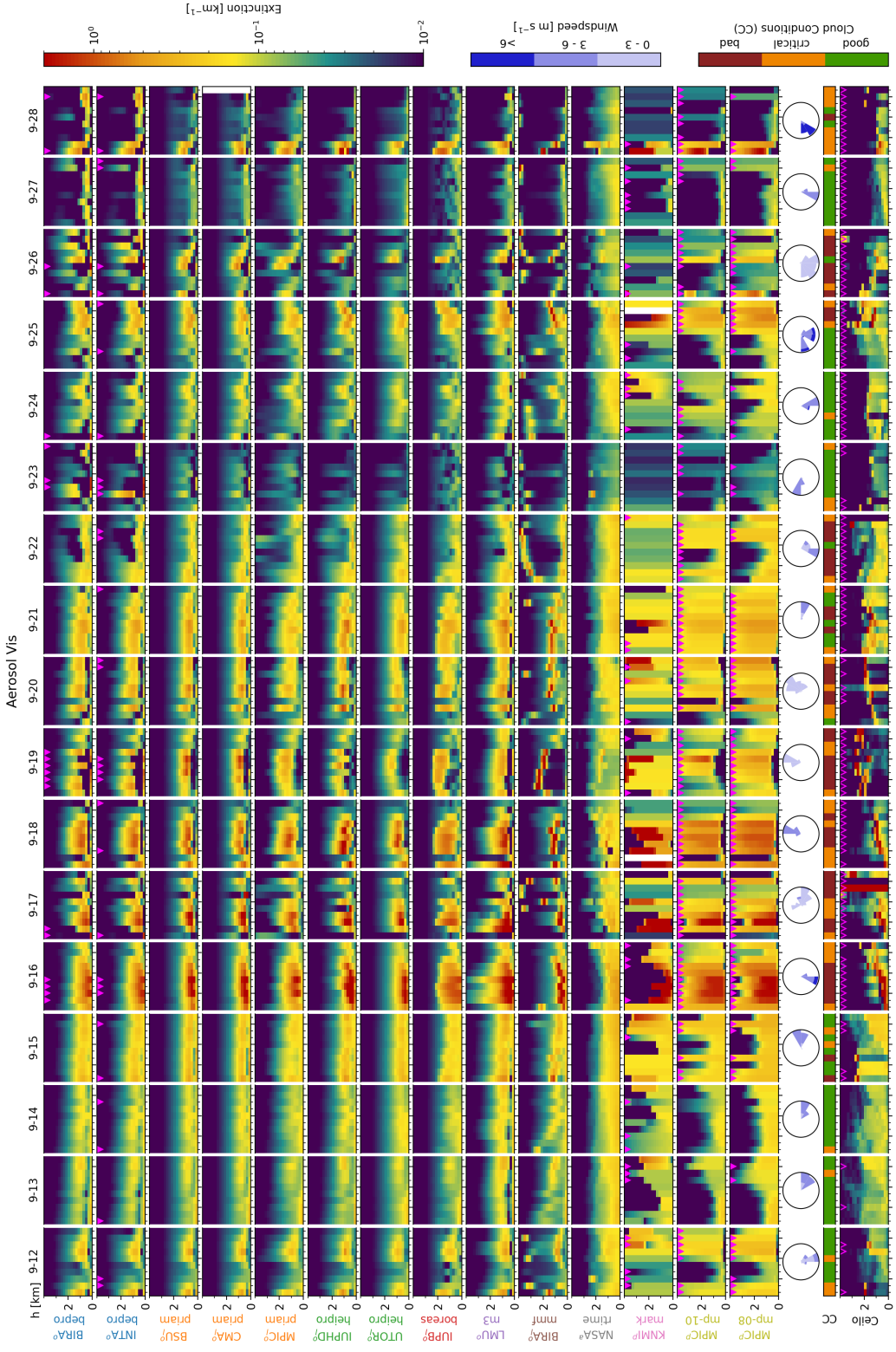


Figure 4. Aerosol Vis extinction profiles. Caption of Fig. 3 applies.

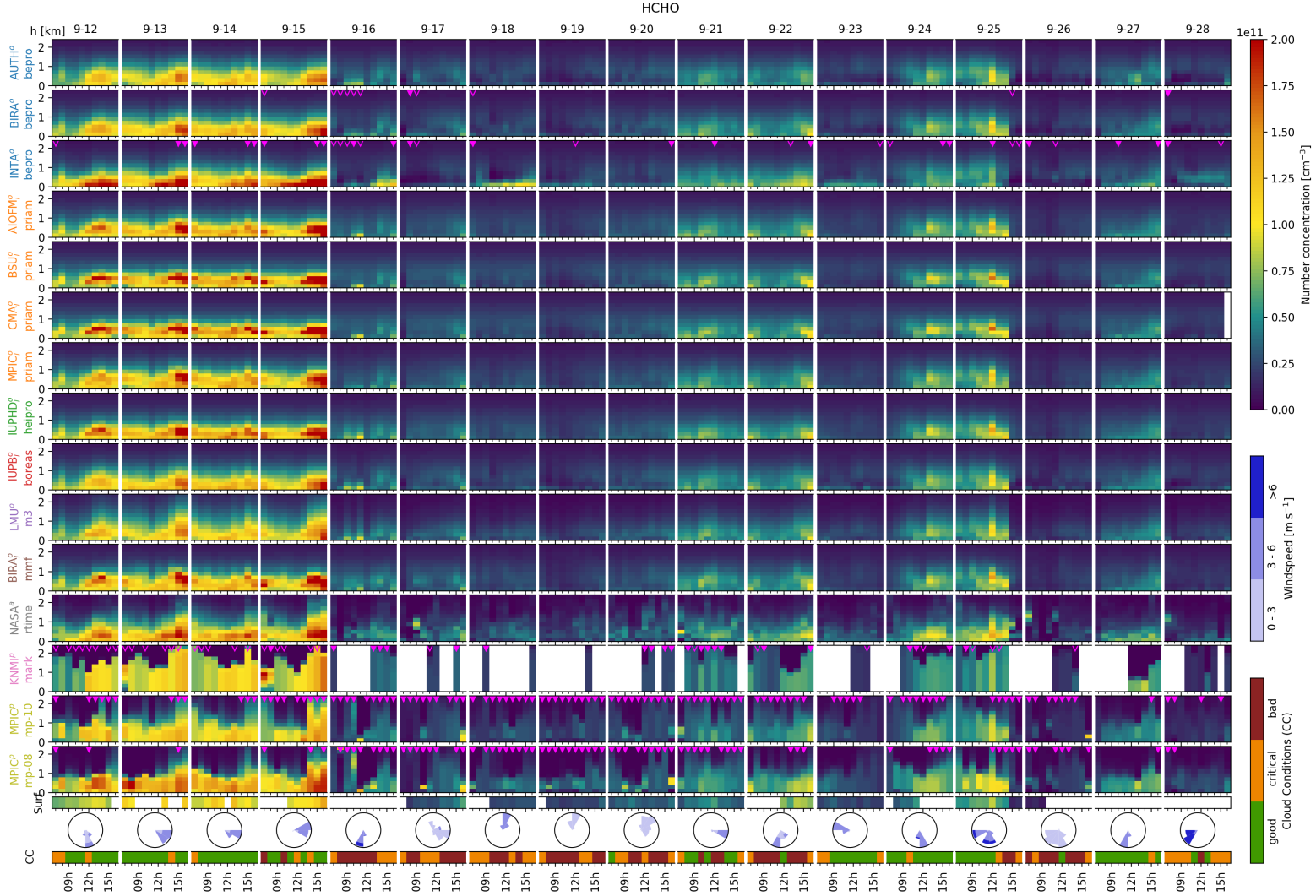


Figure 5. HCHO concentration profiles. The plot is similar to Fig. 3. Open ~~red~~ pink triangles at the top of the MAX-DOAS profiles indicate that the underlying aerosol retrieval failed, whereas the trace gas profile retrieval itself was considered successful. The "Surf"-row shows LP-DOAS surface concentrations.

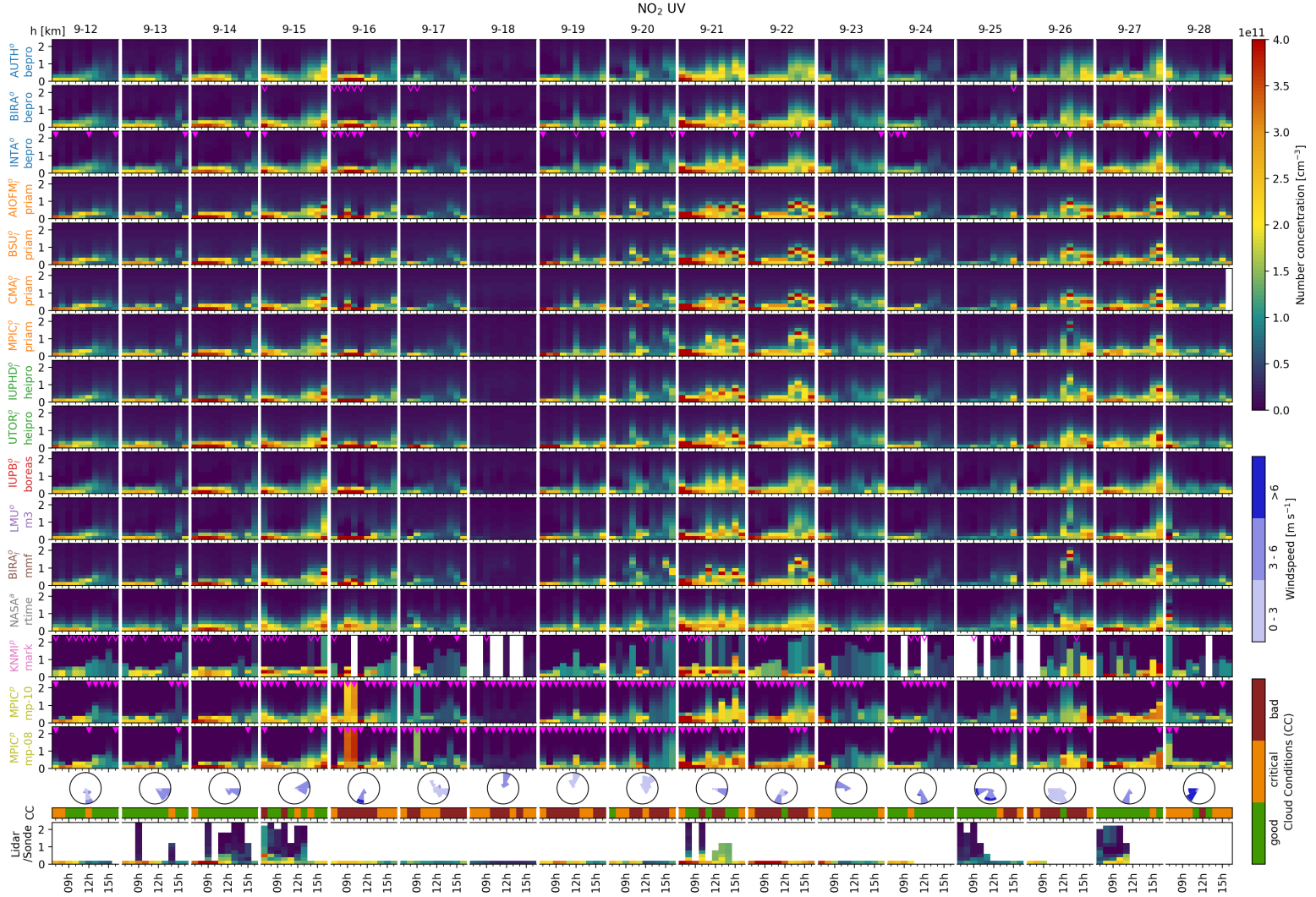


Figure 6. NO_2 UV concentration profiles. The lowest row shows a combined dataset of NO_2 lidar, radiosonde, LP-DOAS and tower in-situ data. Redundant surface concentration measurements were averaged.

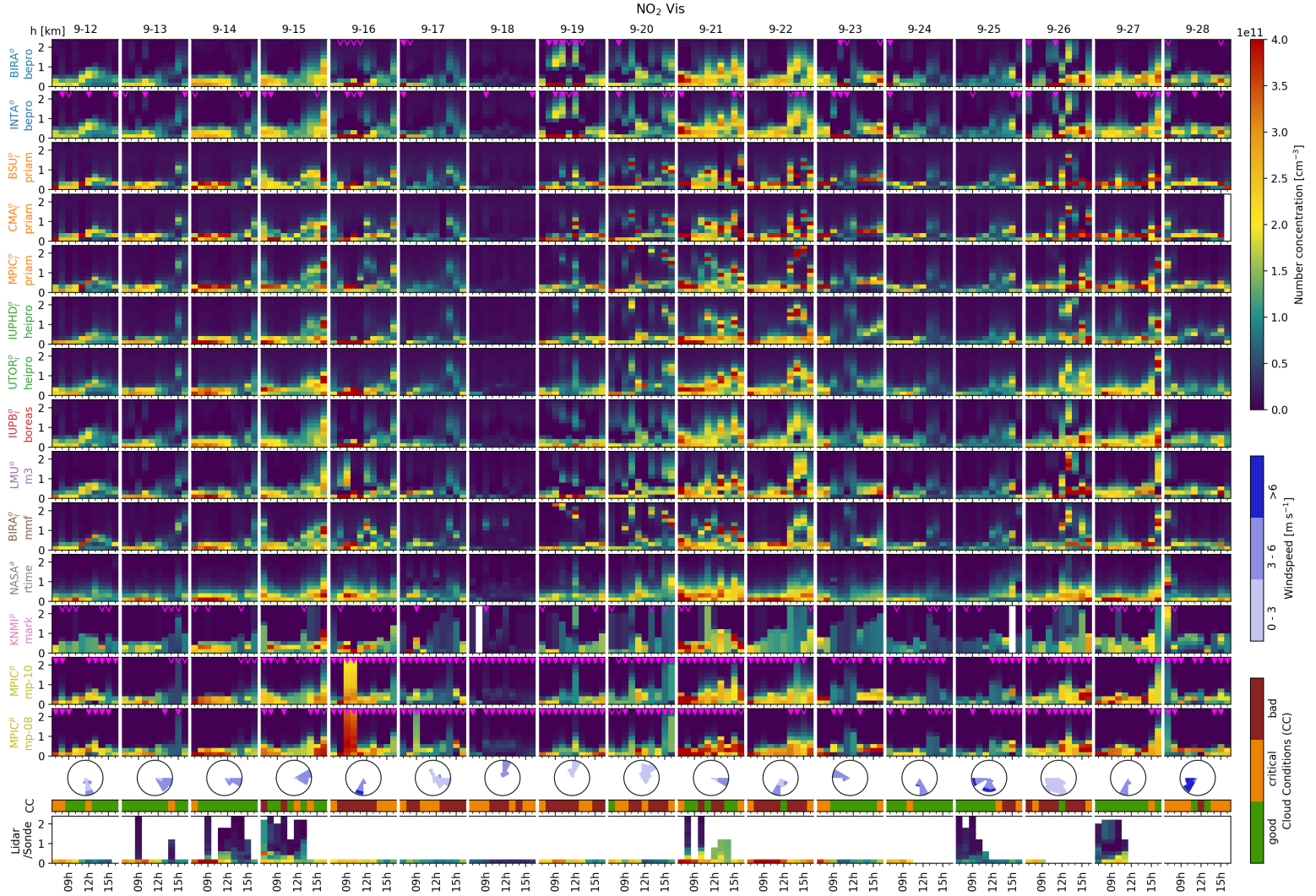


Figure 7. NO_2 Vis concentration profiles. The lowest row shows a combined dataset of NO_2 lidar, radiosonde, LP-DOAS and tower in-situ data. Redundant measurements were averaged here.

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 $3.5\times10^{-2}\,\mathrm{km}^{-1}$ ~~, $4.0\times10^{-2}\,\mathrm{km}^{-1}$, $1.2\times10^{10}\,\mathrm{molec\,cm}^{-3}$, $2.4\times10^{10}\,\mathrm{molec\,cm}^{-3}$ and $4.4\times10^{10}\,\mathrm{molec\,cm}^{-3}$~~ for Aerosol UV, $4.0\times10^{-2}\,\mathrm{km}^{-1}$ for Aerosol Vis, ~~HCHO,~~ $1.2\times10^{10}\,\mathrm{molec\,cm}^{-3}$ for HCHO, $2.4\times10^{10}\,\mathrm{molec\,cm}^{-3}$ for NO_2 UV and $4.4\times10^{10}\,\mathrm{molec\,cm}^{-3}$ NO_2 Vis~~,~~ ~~respectively~~. 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~~(e.g. number of iteration~~. The latter are for instance the accuracy criteria for the RTMs, the number of iterations in the inversion, ~~accuracy criteria for the RTMs, update interval of the jacobians, ...). Larger~~ 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 (~~Earth curvature, treatment of~~ like the Earth curvature or the treatment of the instrument field of view~~, ...~~) gain influence. ~~Horizontal inhomogeneities are an unlikely reason because~~ Note that the worse performance in the Vis was also apparent in the study by Frieß et al. (2019) with synthetic data~~, where horizontal gradients were non-existent~~. 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 $4\,\mathrm{km}$. ~~An example for large discrepancies between participants using the same algorithm is AUTH aerosol in the UV, where in contrast to other bePRO~~

~~users oscillations seem to appear. We suspect this to originate from technical problems which could not yet been identified. The discrepancies between IUPHD and UTOR (both using HEIPRO) were found to mainly be caused by differences in the number of applied iteration steps in the Levenberg-Marquardt optimization scheme during aerosol retrieval. IUPHD (UTOR) applied 20 (5) iterations. The consequences are evident throughout the comparison. Compared to the parametrized approaches, OEMs and the Realtime algorithm~~ 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~~, AUTH, BSU, KNMI~~ / 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

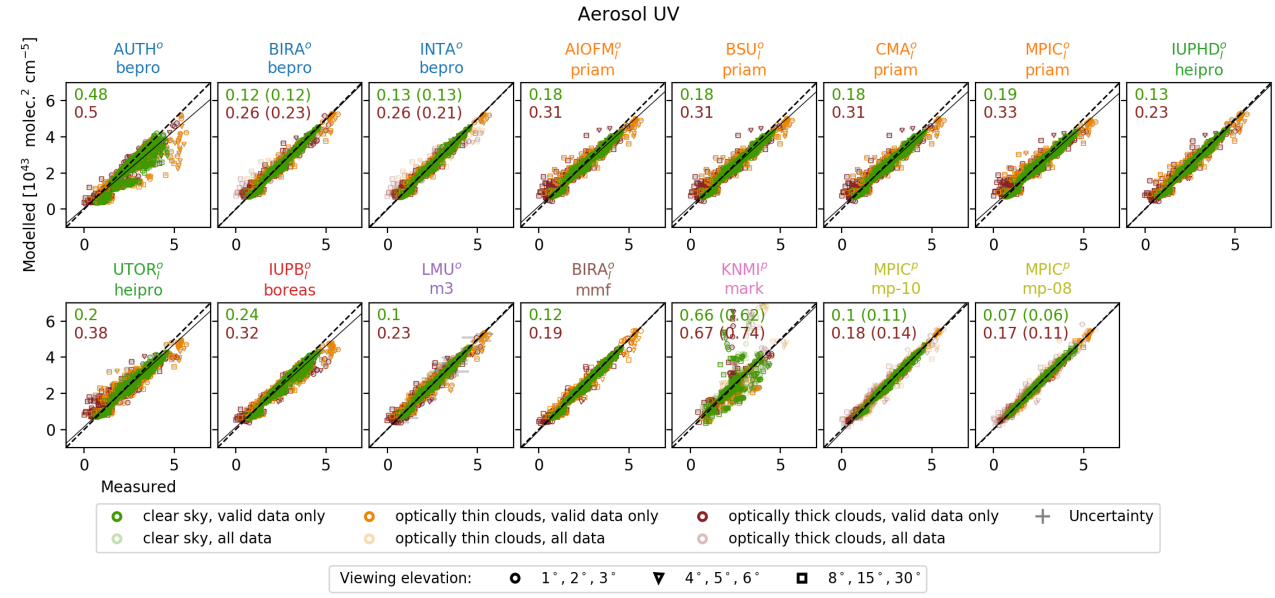


Figure 8. O_4 UV dSCD correlation. Marker colours and marker shapes indicate the cloud conditions and viewing elevation angles, respectively, as indicated in the legend. Numbers represent the measurement-error-weighted RMSD between measured and modelled dSCDs in units of 10^{43} molec^2 cm^{-5} for clear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

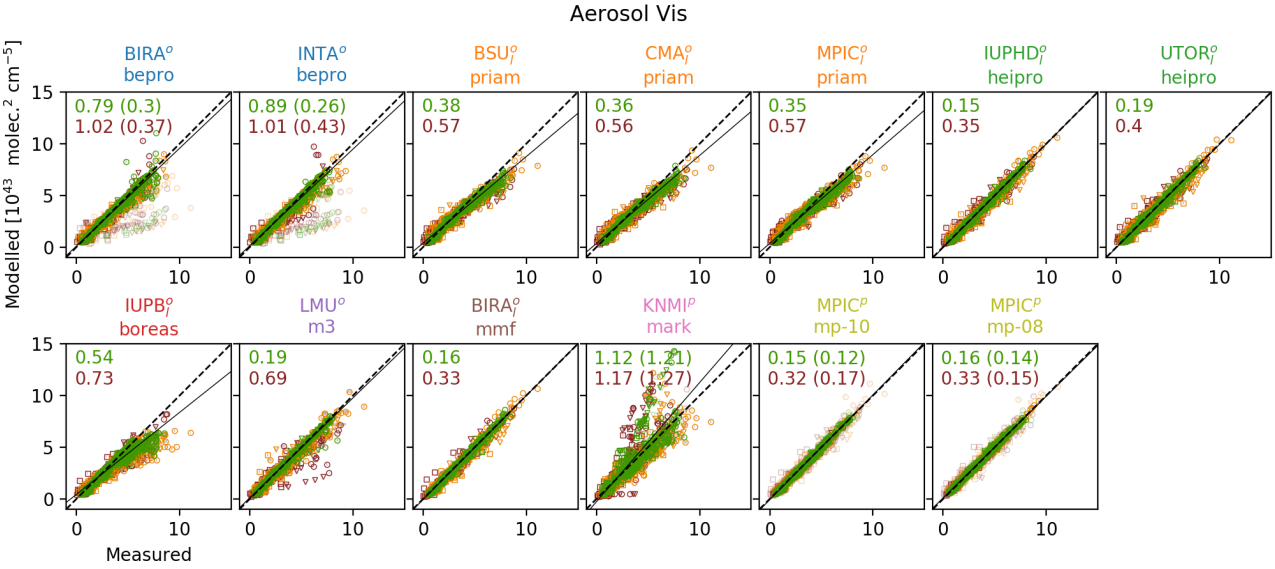


Figure 9. O_4 Vis dSCD correlation. Legends and description of Fig. 8 apply.

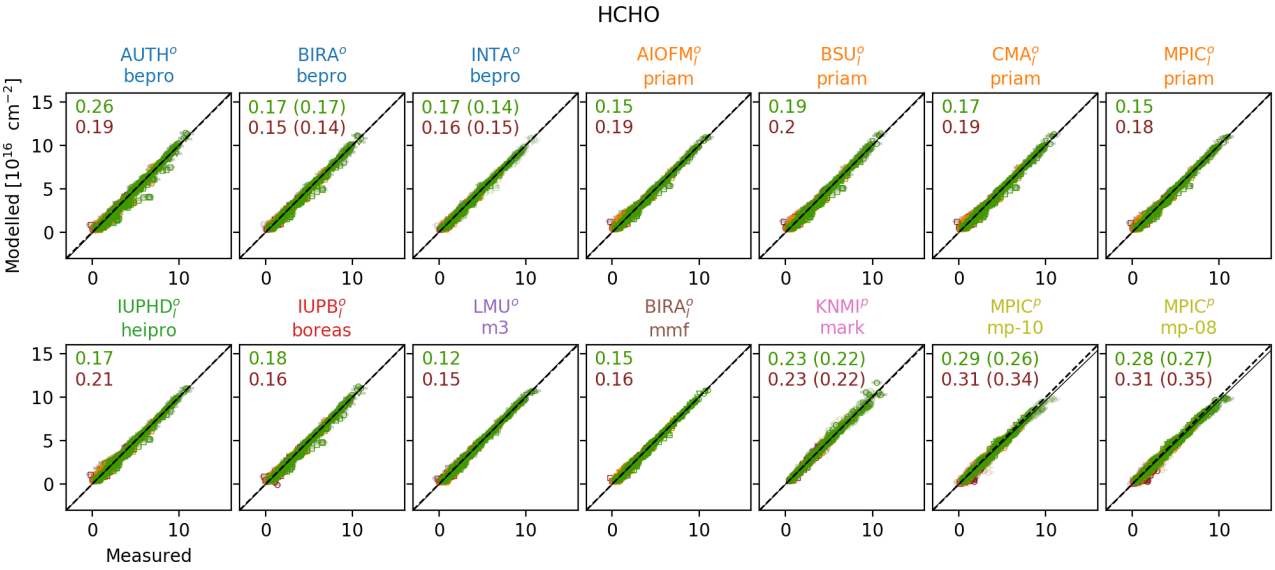


Figure 10. HCHO dSCD correlation. RMSD between measured and modelled dSCDs in units of 10^{16} molec cm^{-2}. Legends and description of Fig. 8 apply.

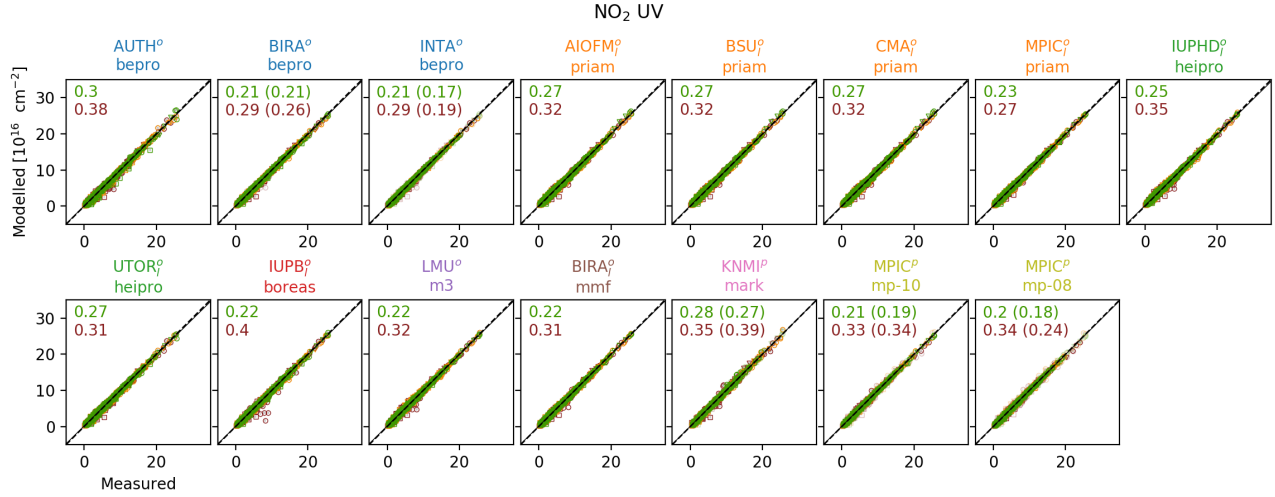


Figure 11. NO_2 UV dSCD correlation. RMSD between measured and modelled dSCDs in units of 10^{16} molec cm^{-2}. Legends and description of Fig. 8 apply.

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

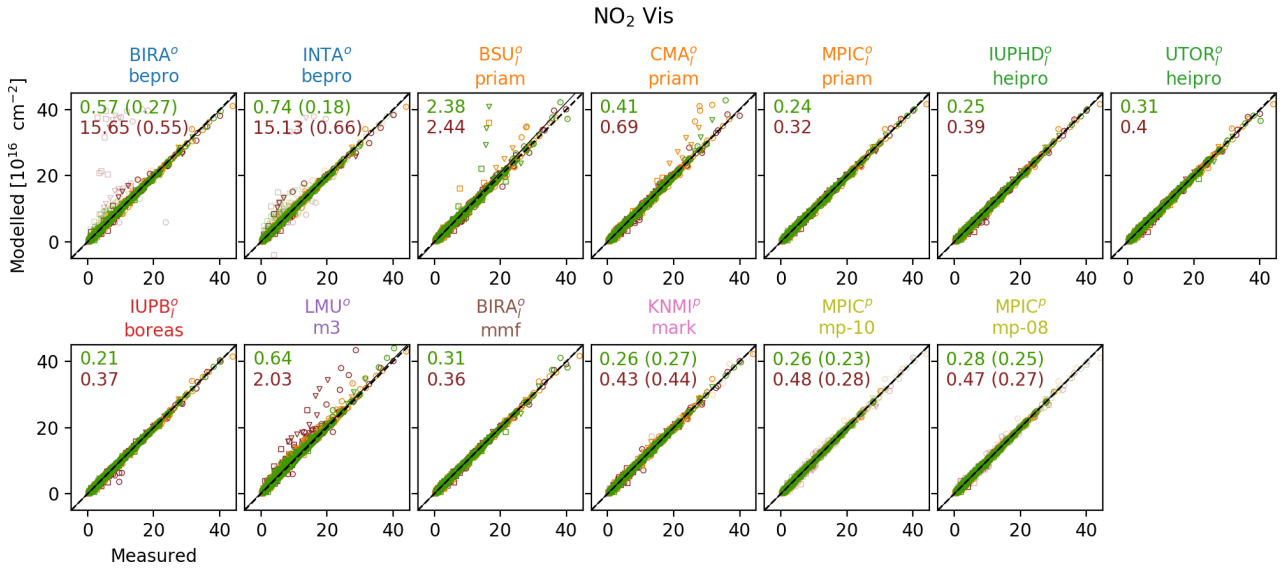


Figure 12. NO_2 Vis dSCD correlation. RMSD between measured and modelled dSCDs in units of 10^{16} molec cm^{-2}. Legends and description of Fig. 8 apply.

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 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).

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. ~~As discussed in Sect. 2.3.2 these two quantities are not necessarily comparable. As shown in Sect. 3.1 the~~ 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 ~~decreases rapidly with altitude~~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 ~~correpsonding~~ 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 ~~detected~~ located and quantified by ground-based

MAX-DOAS observations. ~~Thus they can only~~ 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 ~~vertically integrated extinction profiles~~ and sun photometer ~~AOTs τ_s~~ 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". ~~Average values over the whole campaign~~ 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 ~~f_τ are 0.81 ± 0.16 for~~ Aerosol UV and ~~0.90 ± 0.13 for~~ Aerosol Vis (~~using the median AVKs of all OEM retrievals).~~ $\boldsymbol{x}_{UV}$ and $\boldsymbol{x}_{Vis}$, blue and green), respectively. In this particular case it is expected that a large fraction of the aerosol above 1 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~~(see Supplement ?? for details)~~. In the following, this ~~correction~~ 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 ~~in this section~~ 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 ~~physical reason for PAC and SF are different.~~ 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; 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

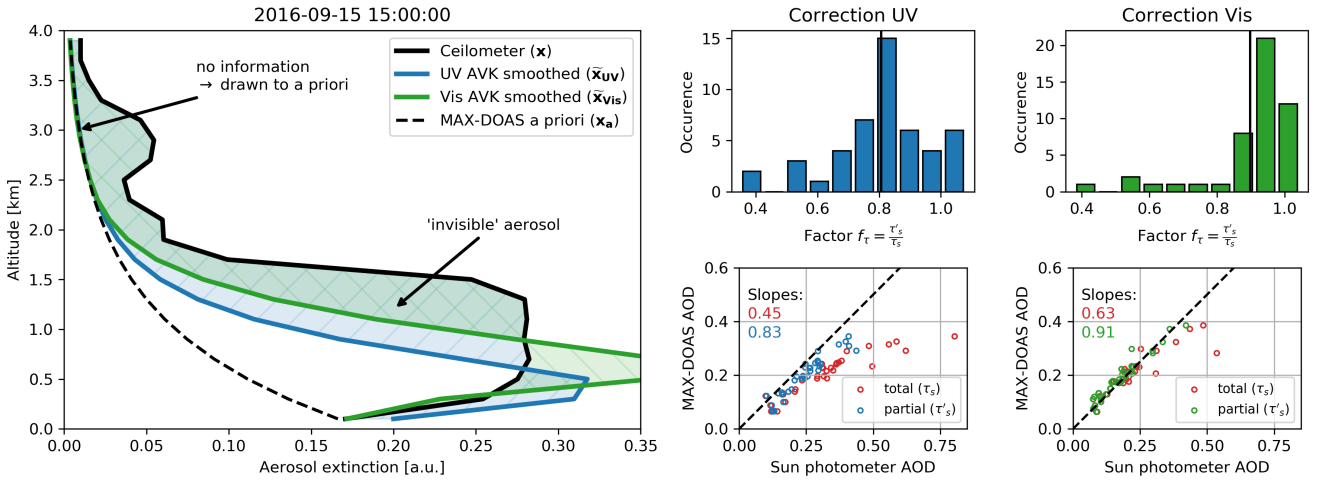


Figure 13. Left panel: example for the smoothing of a ceilometer backscatter profile $\boldsymbol{x}$ (according to Eq. (9)) with particularly heavy aerosol load at high-altitudes retrieved in the UV and Vis, respectively. Right panel: distribution and impact of the correction factor $f_\tau = \tau'_s/\tau_s$ 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 $\tau_s = f_\tau \cdot \tau'_s$.

best knowledge) to measured O_4 dSCDs. They do not make use of optimal estimation or prior constraints similar to those used in our study. Thus their findings can be considered independent from any kind of PAC.

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~~, KNMI / Realtime, KNMI/ MARK~~ and MPIC/ MAPA) are correlated against τ_s and ~~might therefore be underprivileged~~are therefore expected to generally achieve worse agreement. For correlations of OEM algorithms against τ_s please refer to Supplement S8.3. Correlation parameters~~and RMSD~~ , 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 ~~) for Aerosol UV (Vis)~~in the Vis. In the presence of clouds they increase by about 30 % ~~(~~and 80 %~~), ~~, respectively, which is to ~~a large part caused by~~ 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.

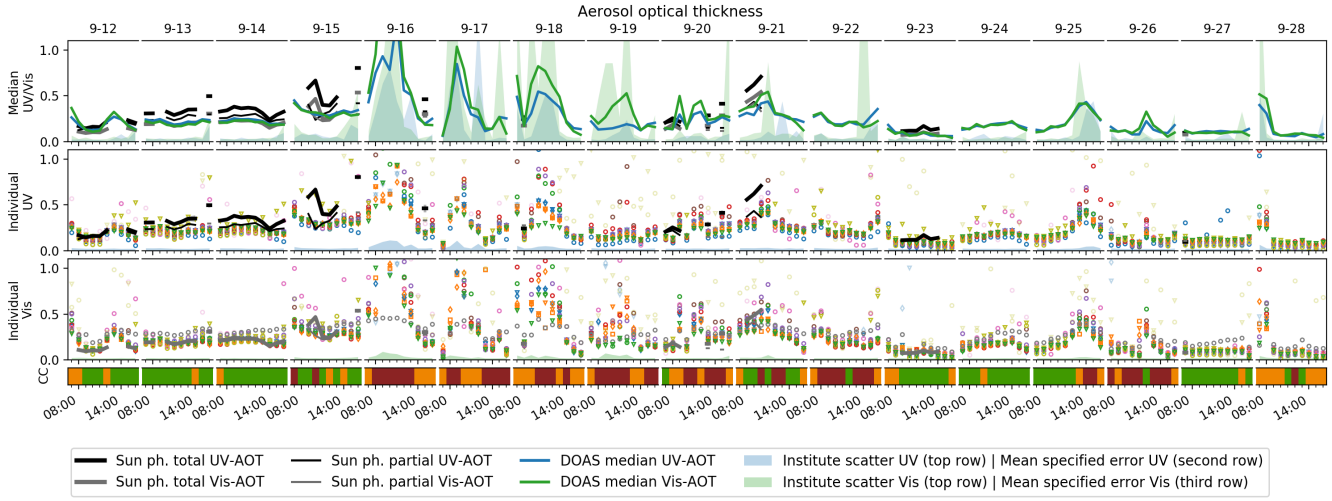


Figure 14. MAX-DOAS retrieved AOTs in comparison to sun photometer data. Symbol and symbol colours are chosen according to Table 2. ~~Open~~ Transparent symbols indicate data flagged as invalid. Top row: MAX-DOAS median results vs. the available supporting observations, according to the legend below the plot. The "institute scatter" ~~hatched~~ areas ~~(sharing the AOT's y-axis scaling but starting at the top of the plot)~~ 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.

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 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). ~~Good performance~~ Best performance in the UV is observed for ~~bePRO (except AUTH),~~ 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 ~~), M^3 and MMF with RMSDs around 0.05 (0.03) . For other OEM algorithms, larger underestimations of the partial AOT (0.5 < slope < 1) are observed~~ and UTOR) and the BOREAS (IUPB) algorithm. For all participants except MPIC-0.8/ MAPA, negative Biases < -0.03 in the UV ~~, which are most evident in the case of PRIAM (≈ 0.5). Interestingly, the AVKs at higher layers derived from PRIAM are systematically higher than most other algorithms(see Sect. S8.1), which reduces the impact of the PAC and results in a larger partial AOD τ_s' than for most other datasets. Therefore, the lower slopes of PRIAM might rather be owed to its assessment of information content than to the retrieval algorithm itself. For Aerosol Vis bePRO~~ remain, even though the PAC has been applied for the OEM algorithms. The

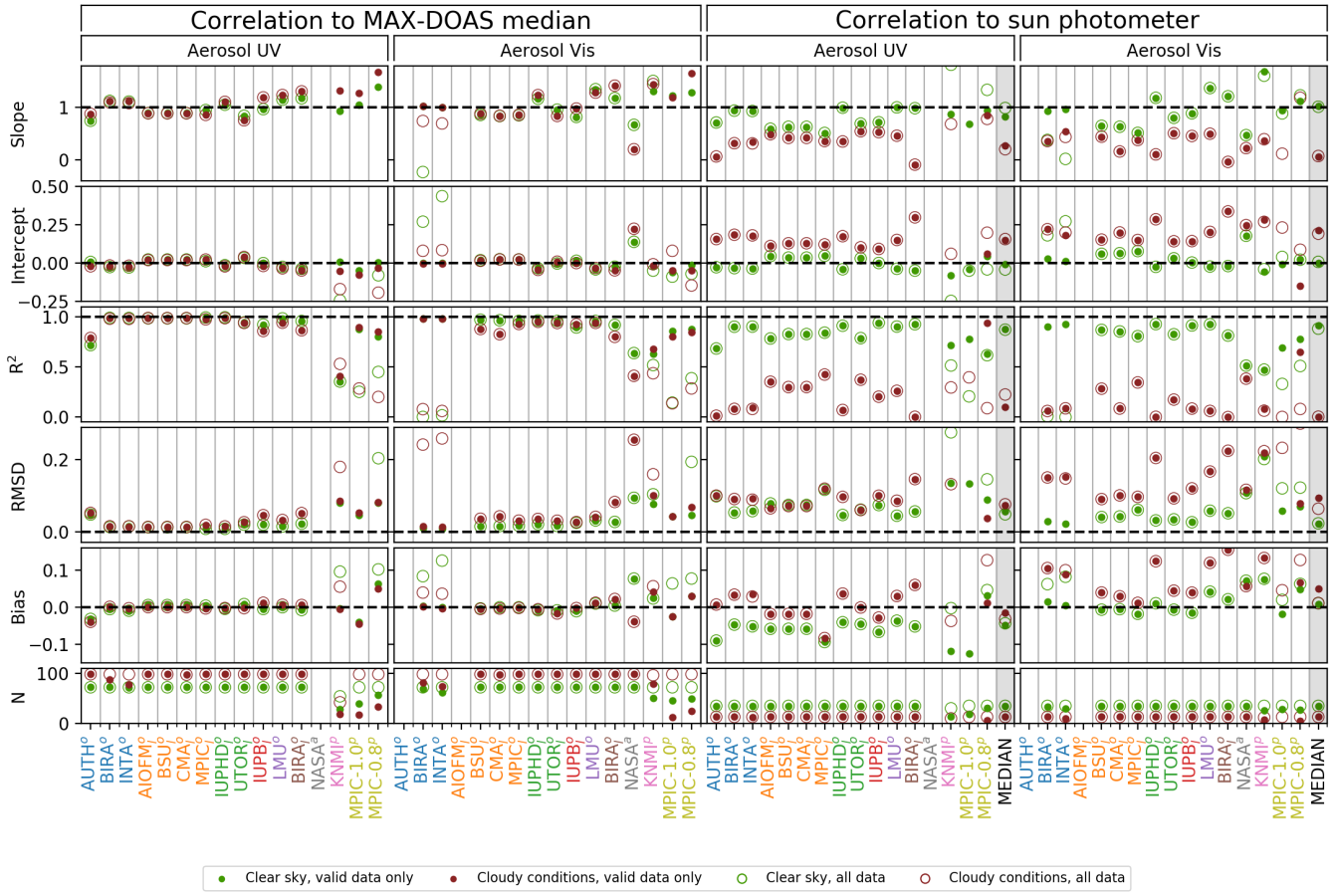


Figure 15. Correlation statistics for AOTs. The two left columns give an impression on the agreement among the institutes, as they show the correlation of the individual participant's retrieved AOT (ordinate of the underlying correlation plot) against the median (abscissa). The two right columns show 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. ~~Light symbols~~ Hollow circles represent values for all submitted data, ~~opaque symbols~~ the dots only consider data points flagged as valid. ~~The pies indicate, which fraction of~~ N is the ~~total~~ number of profiles ~~(170)~~ which contributed to the respective ~~value~~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.

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 conver-

gence problems during inversion in the Vis (see Sect. 3.3) but the affected results are reliably flagged. KNMI/ MARK~~and~~, 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 ~~SF $\approx f_\tau$~~ an O_4 dSCD scaling factor of $SF \approx f_\tau$ have very similar impact on the AOT correlation. Consequently, the application of ~~an O_4 dSCD scaling factor 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 ~~slope > 1 and intercept > 0.~~ a Bias of about 0.05.

3.5 Trace gas vertical column densities

This section ~~compares the VCDs of~~ 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~~(NO_2 and HCHO)~~, 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 ~~(in particular the low sensitivity of MAX-DOAS observations for higher altitudes)~~ 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 ~~an outlier~~ 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 ~~and 1.03 ± 0.03~~ in the UV and 1.03 ± 0.03 in the Vis are obtained, ~~respectively,~~ 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.

Under clear sky conditions, average RMSD values against the MAX-DOAS median are 5×10^{14} molec cm^{-2} for HCHO and 7×10^{14} molec cm^{-2} ~~) for HCHO (~~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 1.4×10^{15} molec cm^{-2}. Note however that the two observations are not fully independent, as for the direct-sun data, the residual HCHO amount in

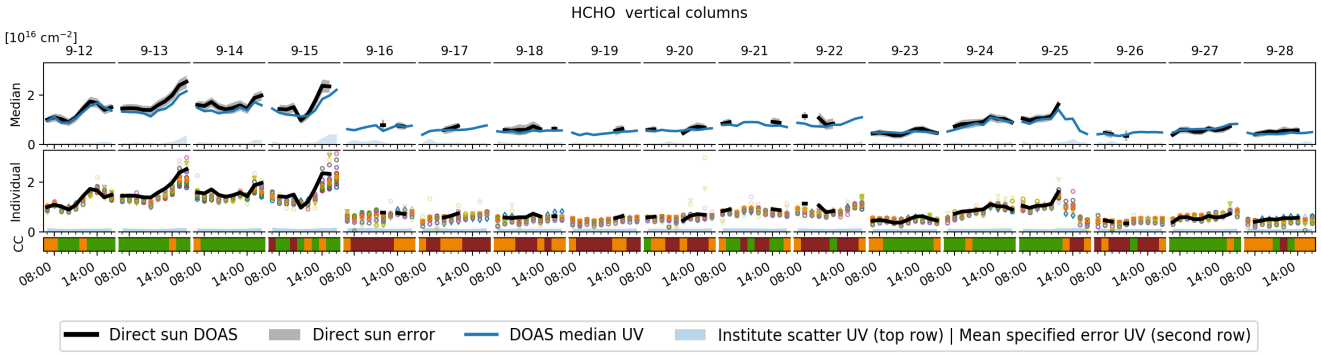


Figure 16. Comparison of MAX-DOAS retrieved HCHO VCDs vs. direct-sun DOAS. Basic descriptions of Fig. 14 apply.

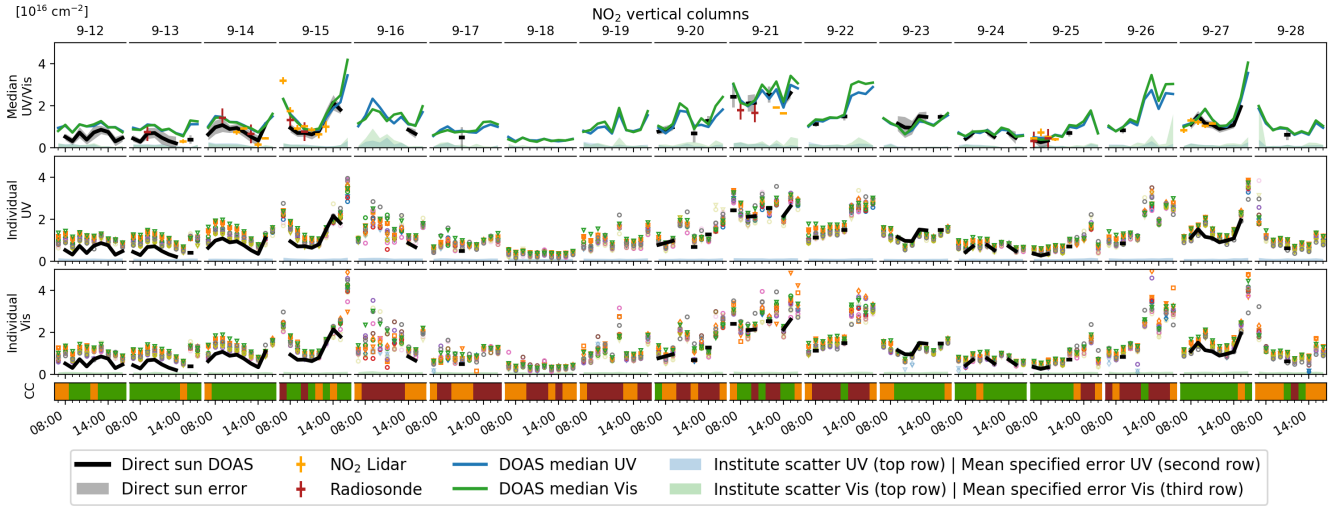


Figure 17. Comparison of MAX-DOAS retrieved NO_2 VCDs vs. direct-sun DOAS, NO_2 lidar and radiosonde. Basic descriptions of Fig. 14 apply.

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 3.7×10^{15} molec cm^{-2} (3.8×10^{15} 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 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. ~~A candidate are the different sampling volumes: while radiosonde, lidar and direct-sun DOAS typically sample air at maximum distances of a few kilometres to the site, the MAX-DOAS instruments have a much larger horizontal sensitivity range~~

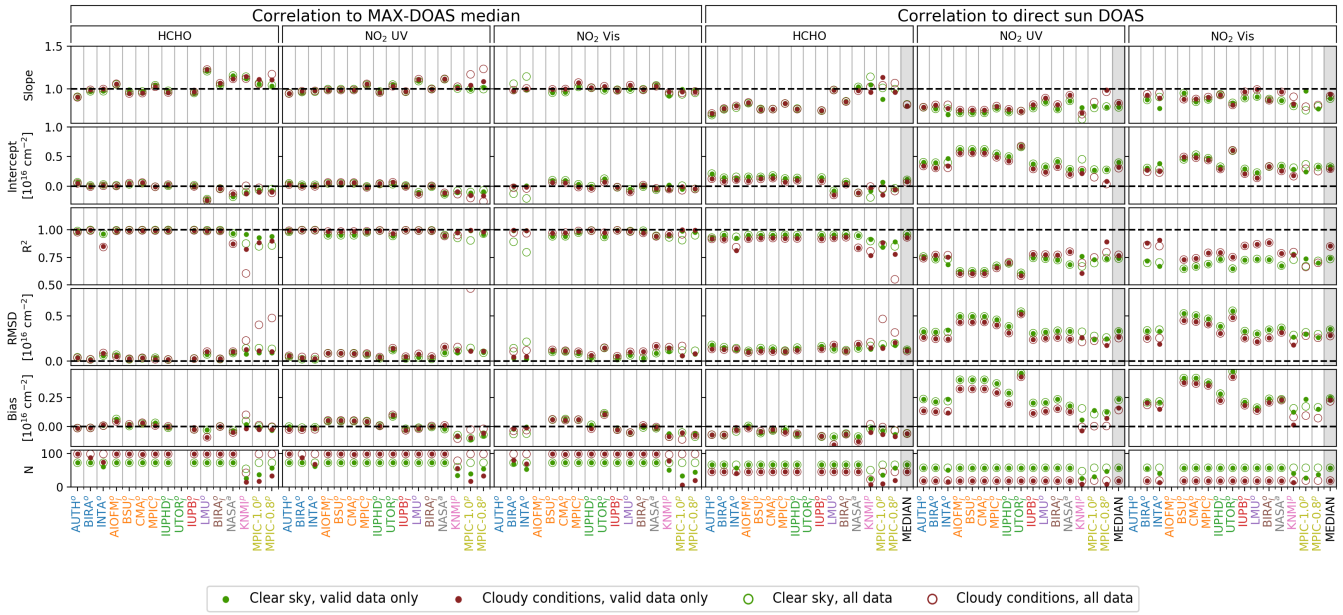


Figure 18. Correlation statistics of trace gas VCDs. The plot is similar to Fig. 15. In the underlying correlation plots, ordinates are MAX-DOAS VCDs of individual participants and abscissas are the MAX-DOAS median and direct-sun VCDs, respectively. The correlation plots are shown in Supplement S8.3.

~~(see Supplement S5), even extending to The Hague on some days, which is $> 40\,\mathrm{km}$ away. Indeed the agreement improves with decreasing visibility~~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 $1.3 \times 10^{15}\,\mathrm{molec\,cm^{-2}}$ ~~, $1.3 \times 10^{15}\,\mathrm{molec\,cm^{-2}}$ and $1.2 \times 10^{15}\,\mathrm{molec\,cm^{-2}}$~~ 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~~, respectively~~. 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 ~~(SNR) of the measured~~ 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 ~~provides~~

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 5.5×10^9 molec cm^{-3} (4×10^9 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.

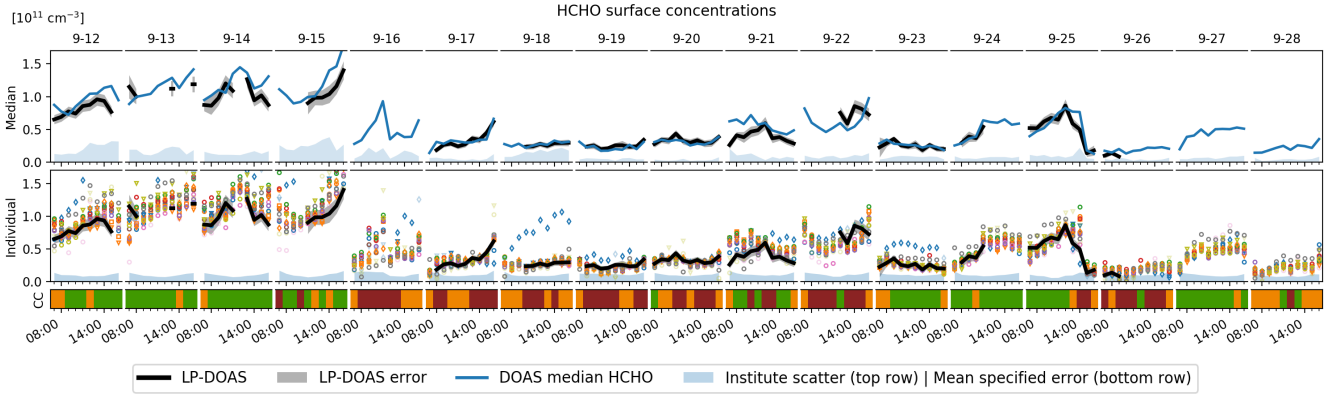


Figure 19. Comparison of MAX-DOAS retrieved HCHO surface concentrations. Basic descriptions of Fig. 14 apply. Note that the mean specified uncertainties in the two lower rows of the figure are very small and thus barely visible.

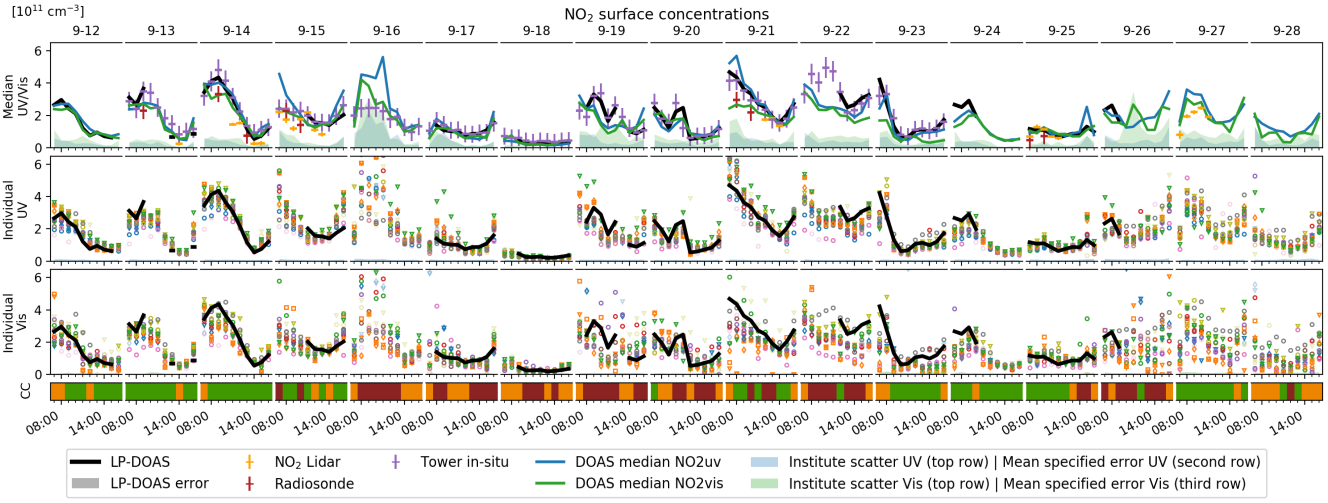


Figure 20. Comparison of MAX-DOAS retrieved NO_2 surface concentrations. Basic descriptions of Fig. 14 apply. Note that the mean specified uncertainties in the two lower rows of the figure are very small and thus barely visible.

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 ~~also~~ requires an isolation

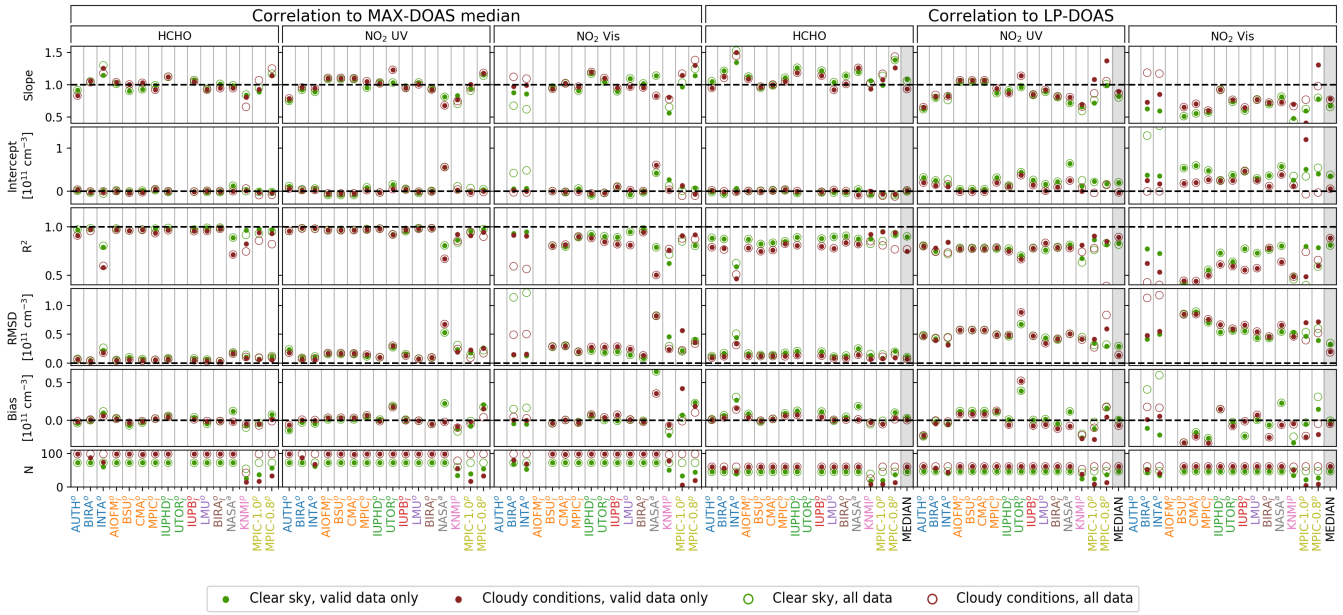


Figure 21. Correlation statistics of trace gas surface concentrations. The plot is similar to Fig. 15. In the underlying correlation plots, ordinates are MAX-DOAS surface concentrations of individual participants and abscissas are the MAX-DOAS median and direct-sun VCDs, respectively. The correlation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

of the surface layer from the layers above and therefore reflects the MAX-DOAS' ability to actually resolve vertical profiles ~~,~~ ~~as it requires an isolation of the surfacelayer from the layers above~~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 8.8×10^9 molec cm^{-3} ~~, 1.8×10^{10} molec cm^{-3} and 2.7×10^{10} molec cm^{-3}~~ for HCHO, 1.8×10^{10} molec cm^{-3} for NO_2 UV and 2.7×10^{10} molec cm^{-3} for NO_2 Vis~~, respectively~~. For the comparison to the LP-DOAS, ~~they~~ these values increase to 1.8×10^{10} molec cm^{-3}, 4.7×10^{10} molec cm^{-3} and 5.6×10^{10} 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 ~~these~~ the results: the average RMSD to the median increases by ~~15, 26~~ 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~~ 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. Againthe RMSDs to the MAX-DOAS median even for clear-sky conditions are , 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 1, 2 and 3 one for HCHO, two for NO_2 UV , NO_2 Vis, respectively).

3.7 NO_2 UV-Vis comparison

Another intrinsic consistency check for the algorithms, besides the comparison of modelled and measured dSCDs in Sect. 3.3, is the comparison of the and three for NO_2 retrieval results in the two different spectral ranges (UV and Vis). These should ideally yield equal results at least when assuming a horizontally homogeneous atmosphere. Figures **??** and **??** show the correlation of VCDs and surface concentrations. Vis, see Fig. 19 and Fig. 20).

Correlation of MAX-DOAS retrieved NO_2 VCDs in the UV and the Vis spectral ranges. Marker colours and transparency indicate the cloud conditions and flagging, respectively, according to the legend.

Correlation of MAX-DOAS retrieved NO_2 surface concentrations in the UV and the Vis spectral ranges. Legends and description of Fig. **??** apply.

For the VCDs, the average RMSD is $1.5 \times 10^{15}\,\mathrm{cm}^{-2}$, which increase by 70 % in the presence of clouds. For clear sky conditions very good agreement (less than 10 % relative RMSD) is observed for MAPA, M^3, MARK, NASA/ Realtime and bePRO/ INTA. There is a tendency for Vis VCDs to be larger than UV VCDs (by 6 % regarding the campaign averages)which might be caused by the different sensitivity in particular in the retrieval layers between 200 m and 1 km altitude. The extended horizontal viewing distance is an unlikely reason since in contrast to the VCDs, surface concentrations in the Vis are smaller than in the UV.

For the surface concentrations, the results are very different for the individual algorithms and participants: The average RMSD is $6.0 \times 10^{10}\,\mathrm{cm}^{-3}$ and increases by 25 % in the presence of clouds. Best agreement with 10 % < relative RMSD < 20 % are achieved by MAPA, Heipro/IUP-HD, MMF and NASA. Both bePRO users show a similar pattern with systematically smaller values in the Vis retrieval. bePRO suffers from a few strong outliers (even exceeding the plotting range), which are however in most cases removed by flagging. For PRIAM, there is large scatter for all the participants. For HEIPRO, there are large discrepancies between the two participants: while IUPHD achieves very good results here, UTOR shows large scatter (approx. factor of 4) similar to PRIAM users which is once more likely to be explained by the different number of applied iteration steps during the aerosol inversion. The remaining algorithms perform reasonable (relative RMSD < 30 %), apart from few outliers that usually occur under cloudy conditions. Particularly good correlation for both, VCDs and surface concentrations, are only achieved by NASA/ Realtime and MPIC/ MAPA .

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). ~~The complete results are shown in Supplement S10~~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. ~~Increase~~ Relative increase in average RMSD (first value) and average Bias magnitude (values in brackets) when participants retrieve profiles from their own dSCDs instead ~~from~~ 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 ~~decreases~~decrease, indicating that the median dSCDs suffer from systematic ~~biases~~errors. Under clear sky conditions, low impact ($\leq 10\,\%$) was found for Aerosol UV AOTs ~~under clear sky conditions~~ 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 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. 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 ~~aim to assess performance and reliability of individual algorithms but also of the MAX-DOAS profiling technique in general~~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.

~~Summary of RMSDs from the comparisons in Sect. 3 for clear-sky conditions. The RMSD values of AOT, VCD and surface concentration are calculated with respect to the corresponding supporting observations. Average RMSD values define the colour scale of each column (see colourbar on the top right). White spaces indicate no data. Average observed values (bottom row) are rounded campaign averages of the supporting observations. The column on the far right 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").~~

Figure 22 shows an overview of RMSD and Bias values for the ~~inherent quality indicators (correlations~~ correlation between measured and modelled dSCDs ~~as well as between NO_2 UV and Vis results) and~~ and the comparisons to ~~available supporting observations(AOT, VCD and surface concentration)~~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 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 ~~any lack of information is filled~~ parameters of few or no information are constrained by *a priori* ~~knowledge~~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 ~~intensity (see also Supplement S8.2)~~magnitude. Under clear-sky condi-

tions, the RMSDs between individual participants and the MAX-DOAS median results range between $(0.01-0.1)$ for AOTs, $(1.5-15)\times10^{14}$ molec cm^{-2} for trace gas VCDs and $(0.3-8)\times10^{10}$ molec cm^{-3} for trace gas surface concentrations~~range between $0.01-0.1$, $(1.5-15)\times10^{14}$ molec cm^{-2} and $(0.3-8)\times10^{10}$ molec cm^{-3}, respectively.~~. These values compare to approximate average AOTs of 0.3, trace gas VCDs of 90×10^{14} molec cm^{-2} and trace gas surface concentrations of 11×10^{10} 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. 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 ~~, these values increase to~~ (see Fig. 22) RMSDs increase to $(0.02-0.2)$ against AOTs from the sun photometer, $(11-55)\times10^{14}$ molec cm^{-2} against trace gas VCDs from the direct-sun DOAS and $(0.8-9)\times10^{10}$ molec cm^{-3} ~~, most likely due to (systematic)errors and~~ 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, 19×10^{14} molec cm^{-2} and 0.74×10^{10} 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~~of all observations .~~. For NO_2 surface concentrations the RMSD resulting from this could roughly be estimated to be around 3×10^{10} 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 ~~in the UV (Vis) the~~ 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 (~~Section~~ Sect. 2.3.2). In particular ~~, the low sensitivity at higher altitudes has the effect that~~ it was shown for the first time, that formerly observed systematic discrepancies between MAX-DOAS integrated aerosol ~~extinction~~ profiles and sun photometer ~~total AOTs are not necessarily comparable quantities (Section 3.4 and Supplement **??**). Such comparisons can lead to doubtful conclusions if no additional information on the real aerosol distribution is available to perform the necessary corrections.~~ 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). 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 ~~$SF = 1.0$~~ 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) 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 ~~3~~ 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) ~~might be~~ are too optimistic as a measure for the MAX-DOAS retrieval accuracy and have to be regarded with care. ~~The discrepancies between the results of the participants using the same algorithm indicate that the retrieval settings that were not prescribed within this study (e.g. number of applied iteration steps in the optimisation process, RTM accuracy options, ...) leave a lot of room for variations. However, technical reasons cannot be fully excluded as the source of the discrepancies. 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.~~

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. ~~A particularly large impact~~ 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 ($65\,\%$).

~~Finally, investigations on the spatio-temporal variability (see Supplement S6) indicate that a significant fraction of the RMSD observed between~~ Further, we estimated what fractions of the observed discrepancies among the MAX-DOAS ~~and supporting observations is caused by imperfect spatio-temporal overlap. Thus for future campaigns~~ 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.

For future campaign and comparison exercises, fixed model parameters (particularly aerosol optical properties) and prior constraints might be chosen more carefully. 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 observations from unmanned aerial vehicles, to reduce the spatio-temporal mismatch between different observations.

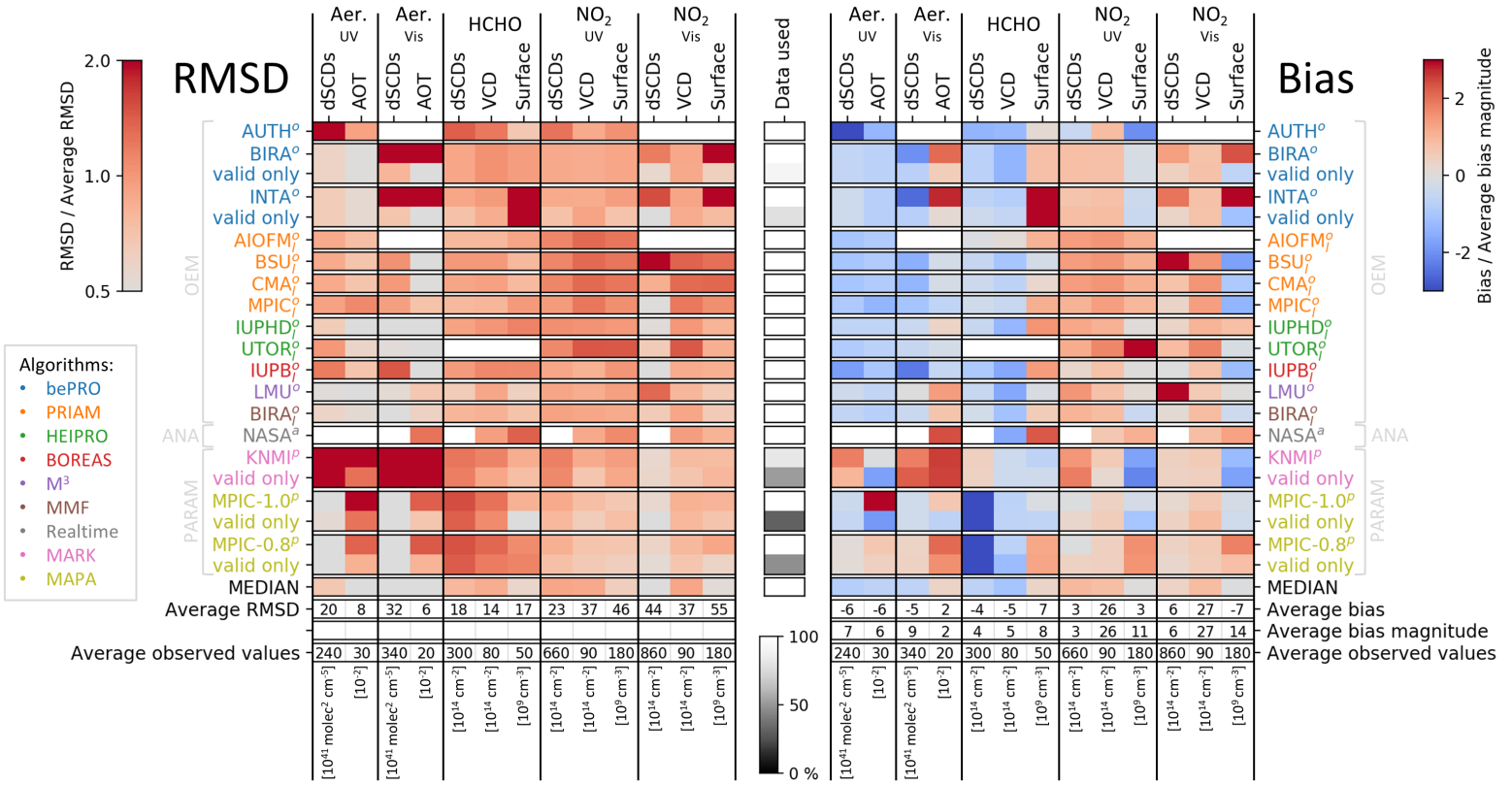


Figure 22. Summary of the comparisons in Sect. 3 for clear-sky conditions. Left panel shows RMSD, right panel shows Bias. Average values of RMSD (Bias) define the colour scale of each column of the left (right) panel as indicated by the color bars on the top left (top right) of the figure. Values of 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.

~~Arnoud~~

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., ~~2019~~ 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.

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.

Koelemeijer, R., De Haan, J., and Stammes, P.: A database of spectral surface reflectivity in the range 335–772 nm derived from 5.5 years of GOME observations, Journal of Geophysical Research: Atmospheres, 108, 2003.

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.

Peters, E., Ostendorf, M., Bösch, T., Seyler, A., Schönhardt, A., Schreier,

Pikelnaya, O., Hurlock, S. F., Henzing, J. C., Trick, S., Wittrock, F., Richter, A., Vrekoussis, M., and Burrowsand Stutz, J. P.: Full-azimuthal imaging-DOAS observations of NO_2 and O_4 during CINDI-2, Atmospheric Measurement Techniques Discussions, 2019, 1–30, , 2019.

: 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., ~~Borovski~~Apituley, A., Bais, A., Beirle, S., Benavent, N., Borovski, A., Bruchkouski, I., Chan, K. L., Donner, S., Drosoglou, T., Finkenzeller, H., Friedrich, M. M., ~~Finkenzeller, H., Friess,~~ Frieß, U., ~~Garcia~~Garcia-Nieto, D., ~~Gómez~~Gómez-Martín, L., ~~Hilboll,~~

A., Hendrick, F., Bruchkouski, IHilboll, A., Jin, J., Kumar, V., Kreher, K., Johnston, P., Koenig, T., Liu, C., Liu, H., Lodi, L K., Kreher, K., Kumar, V., Kyuberis, A., Lampel, J., López, A. SLiu, C., Liu, H., Ma, J., Postylyakov, O., Polyansky, O. L., Postylyakov, O., Querel, R., Roozendael, M. v., Volkamer, R. M., Saiz-Lopez, A., Schmitt, S., Tian, X., Tirpitz, J.-L., Tian, X., Van Roozendael, M., Volkamer, R., Wang, Z., Xing, CZ., Xie, P., Xing, C., Xu, J., Yela, M., Zhang, C., and Wagner, T.: Inter-comparison of MAX-DOASInter-comparison of MAX-DOAS measurements of tropospheric HONO HONO slant column densities and vertical profiles during the CINDI-2 CampaignCINDI-2 Campaign, Atmospheric Measurement Techniques Discussions, 2020, 2019 in prep. 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.

Supplement: 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)

S1 M^3 algorithm description

The M^3 algorithm developed at the Ludwig-Maximilians-University in Munich is an OEM algorithm with Newton Gauss optimization. The radiative transfer model LibRadTran (Mayer and Kylling, 2005) serves as forward model for the retrieval. LibRadTran provides several radiative transfer equation solvers which can handle both pseudo spherical and full spherical geometry. The Jacobian is calculated numerically using the finite difference method, while the box air mass factors for trace gas profile retrievals are calculated using the Monte Carlo module of LibRadTran (MYSTIC). The M^3 retrieval used in the CINDI-2 campaign is a modified version of the algorithm described in detail in Chan et al. (2019) with the iterative optimization of the *a priori* profile disabled. Aerosol extinction profiles are retrieved in the linear space.

S2 ~~Partial AOT correction~~

~~As described in the main text Sect. 3.4, ground-based MAX-DOAS observations have limited sensitivity to aerosol at higher altitudes and thus the profile is strongly biased by *a priori* assumptions here. In the case of OEM algorithms, this effect can be accounted for by applying what in the following is referred to by "partial AOT correction" (PAC). This Sect. demonstrates the impact of the PAC on the comparison between MAX-DOAS and sun photometer data. The left panel of Fig. **??** shows an example of an extreme case during the campaign from September 15^{th}, 15:00h. Shown are a ceilometer backscatter profile (, black) and the same profile smoothed by the MAX-DOAS median OEM averaging kernels for Aerosol UV and Aerosol Vis (, blue and green), respectively. In this particular case it is expected that a large fraction of the aerosol above 1 km altitude will not be detected by 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. However, a part of the high-altitude aerosol appears to be shifted to lower altitudes here by the retrieval. The right panels in Fig. **??** show information for the UV and the Vis retrieval (2nd and 3rd columns) over the whole campaign. The mean values are $f_\tau = 0.81 \pm 0.16$ (0.9 ± 0.13) for the UV (Vis) AOTs. It shall be pointed out that for OEM algorithms the necessity for the PAC can be reduced by using improved *a priori* profiles and covariances (e.g. from climatologies and/or model data). Left panel: example for the smoothing of a ceilometer backscatter profile $\mathscr{x}$ (according to Eq. (9) in the main text) 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. On the top, the distributions of f_τ are shown 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 A, other dots represent the partial AOT $\tau_s = f_\tau \cdot \tau_s'$.~~

S2 O_4 scaling factor

By some groups, the O_4 scaling factor SF is applied to the measured dSCDs before the profile inversion. Initial motivation for its application are previous MAX-DOAS retrieval studies (e.g Wagner et al., 2009; Clémer et al., 2010) which report on a significant mismatch between measured and simulated dSCDs and/ or between MAX-DOAS integrated aerosol extinction and simultaneously measured sun photometer AOT that could not yet be explained (Wagner et al., 2019; Ortega et al., 2016). Meanwhile a series of studies use an SF, empirically determined to values between 0.75 and 0.9.

As described in Sect. 2.1.3, in this study no scaling of O_4 measured dSCDs was applied, except for MPIC-mp0.8 (MAPA algorithm with $SF = 0.8$). For CINDI-2, a SF of 0.8 was observed to enhance the number of valid profiles retrieved by MAPA and to significantly improve the agreement to the sun photometer total AOT in particularly in the UV (Beirle et al. (2019) and within this study). However, as mentioned before, for MAPA (as a parametrized approach without *a priori* profile and AVKs), a PAC as described in Sect. 3.4, cannot be correctly applied and thus deviations to the sun photometer are expected. To further investigate the impact and necessity of the SF for CINDI-2 retrievals, also HEIPRO (as an OEM retrieval) was run with different SFs. In Fig. S1 the impacts of the SF and the PAC on the agreement between MAX-DOAS profiling results (of HEIPRO and MAPA) and the sun photometer are directly compared. Application of $SF = 0.8$ or the PAC, respectively, lead to a very similar improvement in the agreement (regarding RMS), while the application of both together results in a clear overcompensation. This suggests that the PAC and $SF = 0.8$ are equivalent to a large extent and that in the case of MAPA the SF is a way to at least partly account for high-altitude aerosol when it comes to retrieving total AOTs. On the other hand a closer look reveals that if only the PAC is applied, a systematic negative offset of ≈ -0.04 remains in the correlation (for both algorithms and also other participants, compare to main text Sect. 3.4). Indeed, the top row of Fig. S3 shows that for HEIPRO the best RMSD is observed for $SF = 0.92 \pm 0.02$ (UV). Regarding Aerosol Vis (Figure S2), the impact of high aerosol is smaller due to the enhanced vertical sensitivity range (see AVKs in main text Sect. 3.1), such that applying the same scaling factor $SF = 0.8$ as for the UV (without PAC) should already lead to an overcompensation. Indeed, this is observed for HEIPRO as well as MAPA. The bottom row of Fig. S3 shows that, in contrast to the UV spectral range, the best agreement between sun photometer and MAX-DOAS in the Vis is observed for an $SF > 1$.

The second indicator for the need of a scaling factor is a significant mismatch between modelled and measured dSCDs. Figure S4 shows, that for HEIPRO the application of an $SF < 1$ indeed improves the agreement between measured and modelled O_4 UV dSCDs by up to $35\,\%$ in RMSD under clear-sky conditions. Modelled dSCDs are systematically lower than the measured dSCDs in particular for higher elevation angles. ~~Finally~~Also, Wagner et al. (2009) reported ~~, that~~that, under low aerosol conditions, measured dSCDs sometimes even significantly exceed dSCDs modelled within an aerosol free atmosphere, where O_4 dSCDs are ~~expected to be~~ close to the largest possible (regarding clear-sky scenarios only). Therefore, the median measured dSCDs during CINDI-2 at low aerosol load ($\tau_s < 0.1$) were compared to a set of aerosol free modelled dSCDs, showing that this did not happen during CINDI-2 except for the two highest elevation angles $\alpha = 15°$ and $30°$.

Finally, as already stated by Beirle et al. (2019) applying $SF = 0.8$ to MAPA leads to an increased number of valid profiles (see Sect. S3), which again indicates that scaling brings the RTM closer to reality.

The conclusions drawn are as follows: Even without O_4 dSCD scaling, reasonable results and agreement with supporting observations are achieved (if a PAC is applied). In general, a scaling factor of 0.8 seems to be too small but might at least partly be used to account for high-altitude aerosol for algorithms, that cannot quantify their sensitivity or the assimilated *a priori* information. However, there are indications that a less extreme scaling ($0.8 < SF < 1.0$) might in general improve the retrieval. Finally, we think that for this study the prescribed $SF = 1.0$ 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 similar to Wagner et al. (2019) are necessary.

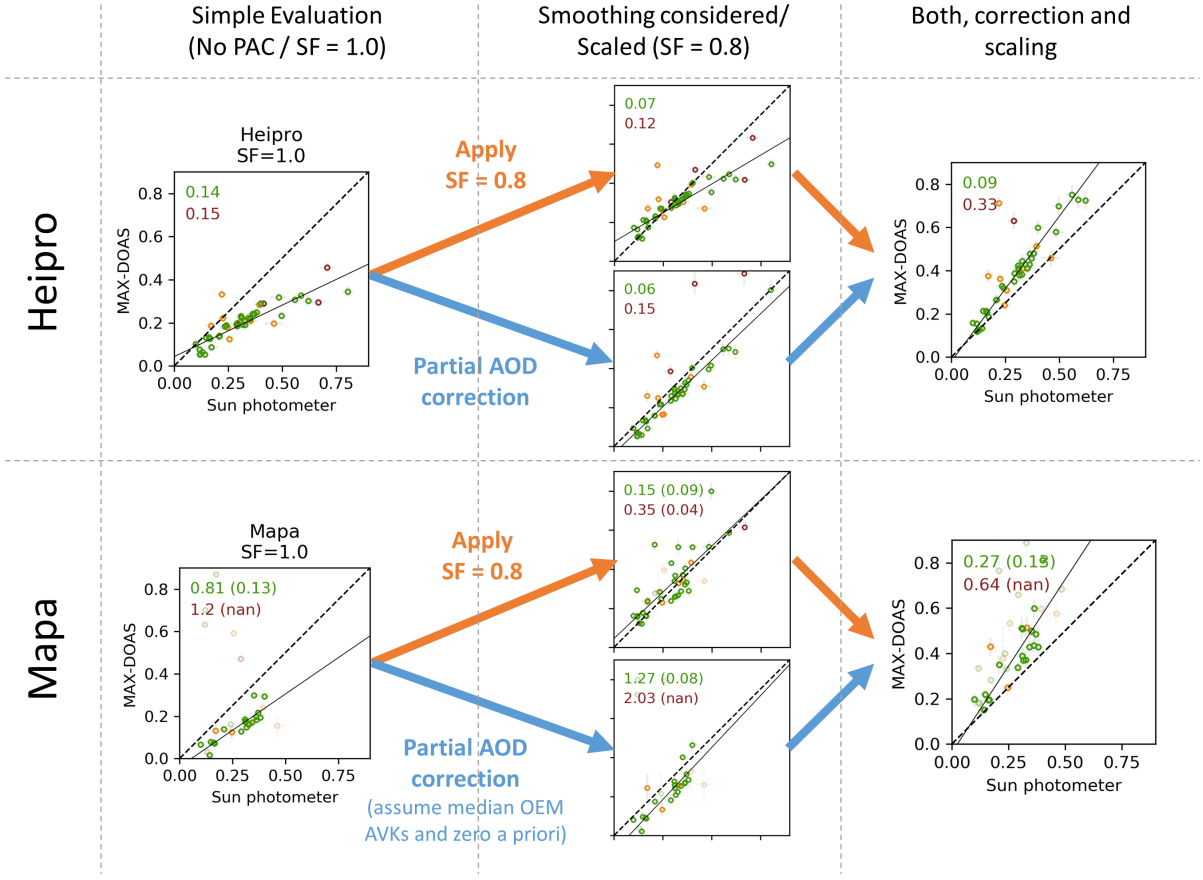


Figure S1. The impact of $SF = 0.8$ and PAC on the agreement between sun photometer and MAX-DOAS AOT (Aerosol UV) in the case of HEIPRO (OEM approach, top row) and MAPA (parametrized approach, bottom row) in a direct comparison. Axes limits and labels of the plots on the left apply for all plots in the figure. Left column: A standard retrieval with $SF = 1$ yields a clear underestimation of the sun photometer AOT for both algorithms. Middle column: Applying the PAC or $SF = 0.8$ leads to a significant improvement. Right column: Applying both leads to overcompensation. Note, that the PAC for MAPA incorporates vague assumptions (median AVKs from all OEM algorithms and $\boldsymbol{x_a} = \mathbf{0}$) and is therefore less meaningful.

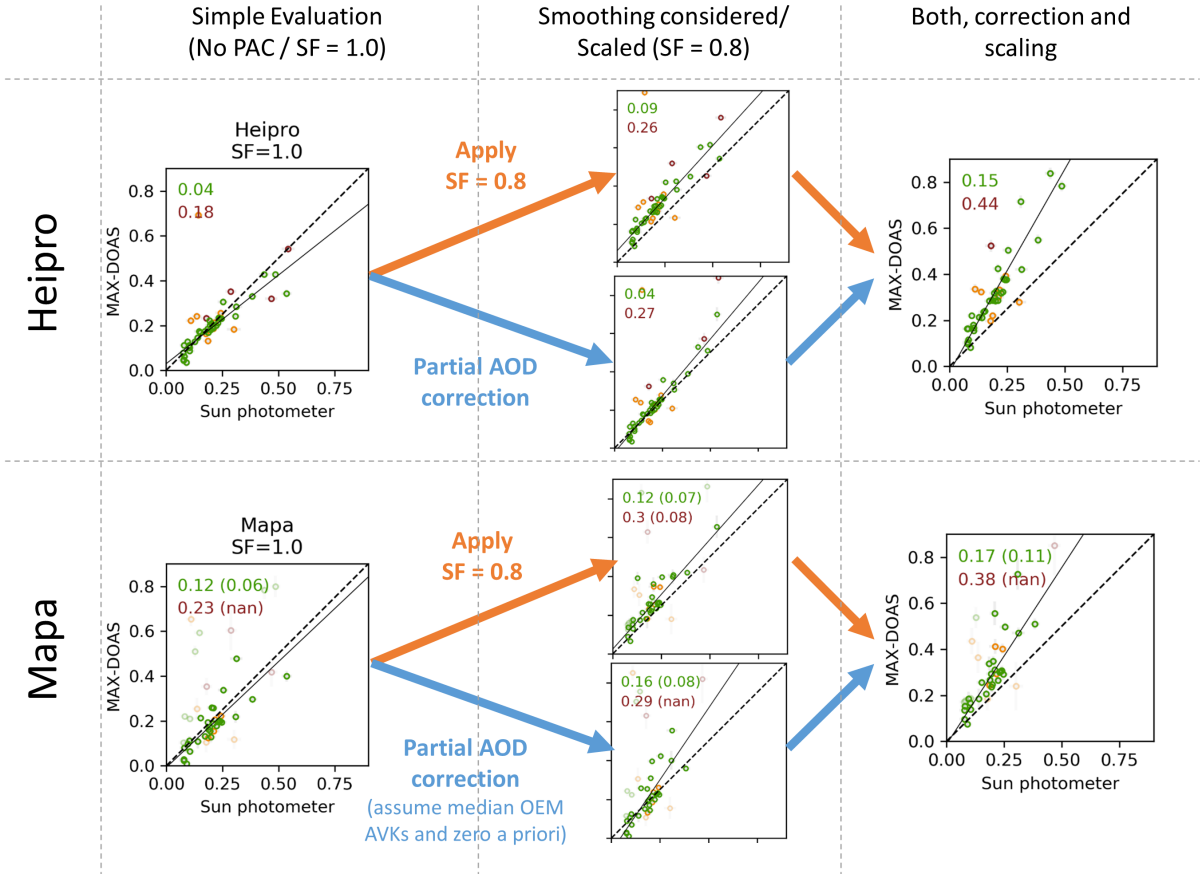


Figure S2. The impact of SF and PAC on Aerosol Vis results. Description of Fig. S1 applies. Due to the extended vertical sensitivity range of the MAX-DOAS observations (compare to main text Sect. 3.1), the effect of high-altitude aerosol is less significant. While this is accounted for in the PAC, the application of the same $SF = 0.8$ as for Aerosol UV leads to an overcompensation here already without additionally applying the PAC.

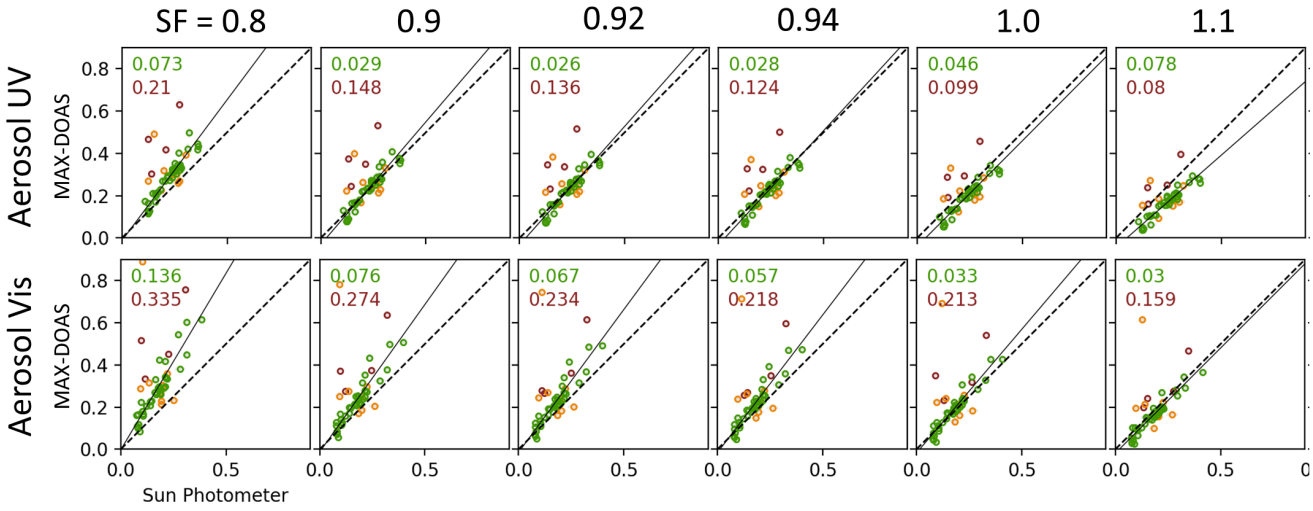


Figure S3. Correlation between the sun photometer AOT (PAC applied) and the HEIPRO algorithms for different SF for Aerosol UV (top row) and Aerosol Vis (bottom row). Symbol colours differentiate between clear-sky (green), and cloudy conditions (orange, red). Numbers in the upper left corners show the corresponding RMSD values. Interestingly, the best agreement for the two spectral ranges is found for different scaling factors, namely for $SF = 0.92 \pm 0.02$ (UV) and $SF > 1 \pm 0.02$ (Vis). The thin line depicts the linear fit for clear-sky data only.

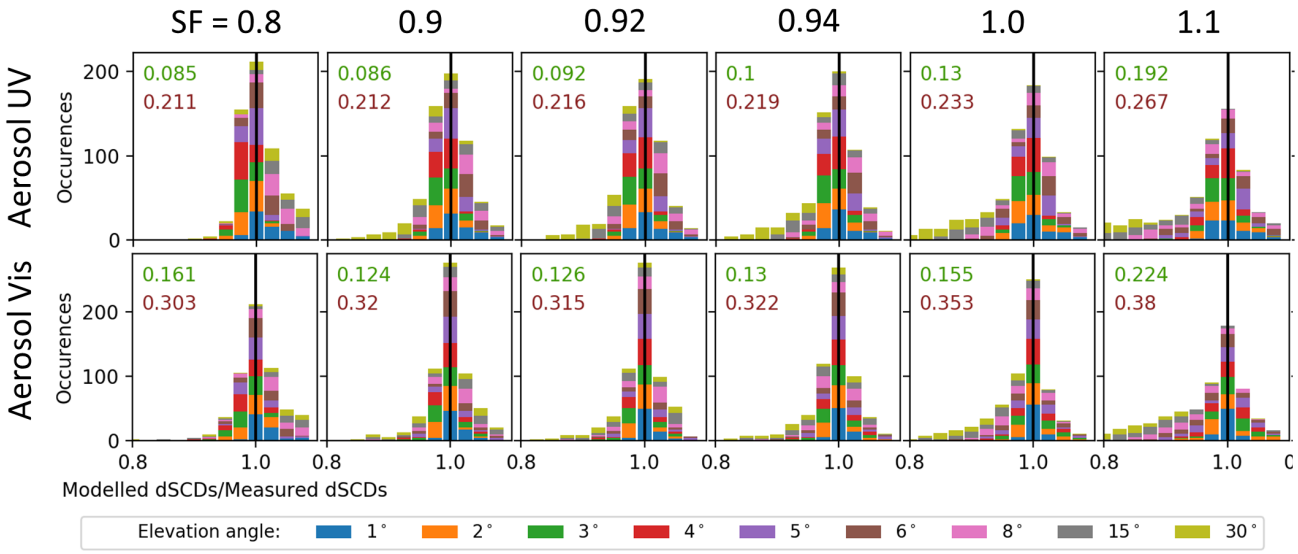


Figure S4. Agreement between modelled and measured dSCDs (depicted as histograms of the ratio) for HEIPRO with different SF for Aerosol UV (top row) and Aerosol Vis (bottom row). Here, only clear-sky and valid data are considered. Numbers in the upper left corners show the corresponding RMSD values. There is a clear tendency, that the agreement improves for $SF < 1$.

S3 Details on profile flagging

Participants were allowed to submit flags, giving them the opportunity to mark profiles as invalid. Four participants submitted flags based on the following criteria:

- BIRA/ bePRO: Profiles are considered valid if the retrieved degrees of freedom are > 1 and if the difference between measured and modelled dSCDs is smaller than $30\,\%$.
- INTA: Profiles are valid if DOFS > 1 and if the RMSD between measured and simulated dSCDs is smaller than 1.5 times the daily averaged RMS.
- KNMI: Profiles are invalid if the spread in the ensemble of solutions for AOT or the NO_2 or HCHO tropospheric column is larger than $15\,\%$ of the retrieved value, or if there are less than 5 out of the ensemble of 20 retrievals for which a solution was found.
- MAPA: Flagging is based on a row of criteria (e.g. the agreement of modelled and measured dSCDs, implausible results ~~,~~ or the consistency within the ensemble of possible solutions~~, ...~~) with carefully chosen thresholds. A detailed description can be found in Beirle et al. (2019).

Table S1 shows the statistics regarding the number of submitted profiles and the fraction of invalid profiles for all participants and species.

Table S1. Submission and flagging statistics.

		HCHO		NO_2 UV		NO_2 Vis		Aerosol UV		Aerosol Vis	
		Total	Valid [%]	Total	Valid [%]	Total	Valid [%]	Total	Valid [%]	Total	Valid [%]
bePRO	AUTH^o	170	100	170	100	-	-	170	100	-	-
	BIRA^o	170	93	170	93	170	87	170	93	170	88
	INTA^o	170	78	170	75	170	71	170	87	170	80
PRIAM	AIOFM^o_l	170	100	170	100	-	-	170	100	-	-
	BSU^o_l	170	100	170	100	170	100	170	100	170	100
	CMA^o_l	169	100	169	100	169	100	169	100	169	100
	MPIC^o_l	170	100	170	100	170	100	170	100	170	100
HEPRO	IUPHD^o_l	170	100	170	100	170	100	170	100	170	100
	UTOR^o_l	-	-	170	100	170	100	170	100	170	100
BOREAS	IUPB^o_l	170	100	170	100	170	100	170	100	170	100
M^3	LMU^o	170	100	170	100	170	100	170	100	170	100
MMF	BIRA^o_l	170	100	170	100	170	100	170	100	170	100
Realtime	NASA^a	170	100	170	100	170	100	-	-	170	100
MARK	KNMI^p	107	38	152	61	168	76	107	43	168	77
MAPA	MPIC-1.0^p	170	31	170	32	170	22	170	32	170	33
	MPIC-0.8^p	170	52	170	51	170	37	170	52	170	43

S4 Further details on supporting observations

S4.1 Aerosol extinction profiles

The available raw data from the ceilometer are attenuated backscatter coefficient profiles $\tilde{\beta}(h)$. For altitudes below 180 m, data are invalid due to insufficient overlap between sending and receiving telescope's FOVs of the instrument, thus $\tilde{\beta}(h < 180\,\mathrm{m})$ was set to $\tilde{\beta}(h = 180\,\mathrm{m})$. For profiles with simultaneously available sun photometer AOTs $\tau_s(\lambda)$, attenuated backscatter profiles $\tilde{\beta}(h)$ were converted to approximate backscatter profiles $\beta(h)$, applying

$$\beta(h) = \tilde{\beta}(h) \cdot \exp\left[2 \int_0^h \frac{\tau_s(\lambda = 1064\,\mathrm{nm})}{\int \tilde{\beta}(h')dh'} dh\right] \tag{1}$$

Extinction coefficient profiles $\alpha_\lambda(h)$ at the MAX-DOAS retrieval wavelengths were then obtained by scaling of $\beta(h)$ with the sun photometer AOT according to:

$$\alpha_\lambda(h) = \frac{\tau_s(\lambda)}{\int \beta(h')dh'} \cdot \beta(h) \tag{2}$$

Integrands with no specified limits in Eq. (1) and (2) indicate integration over the entire available profile. Values for $\tau_s(\lambda)$ at the desired wavelengths were derived according to Eq. (4) in the main text. In case of missing sun photometer data, for instance

due to clouds, the MAX-DOAS retrieved AOT was used instead of τ_s. In this case no attenuation correction (Eq. (1)) could be applied and the integration in Eq. (2) was performed over an averaging kernel smoothed profile (see main text Sect. 2.3.2), to take into account the blindness of MAX-DOAS instruments for higher aerosol layers.

The resulting extinction coefficient profiles at 360 nm could partly be validated with Raman lidar observations 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). Since for the Raman lidar there is not sufficient telescope FOV overlap for altitudes < 1 km to retrieve reliable extinction profiles this comparison is limited to the altitude range between 1 and 4 km. The average RMSD between scaled ceilometer and Raman lidar profiles is ≈ 0.03. Table S2 summarizes the instrument's properties. Figure S5 shows the available Raman lidar profiles in comparison to the ceilometer derived profiles scaled with the respective sun photometer and MAX-DOAS AOT respectively.

Table S2. Properties of the two lidar instruments.

Instrument	Raman lidar	Ceilometer
Data product	Aerosol extinction profile	Elastic backscatter profile
Operational wavelength	355 nm	1064 nm
Altitude range	1 to 10 km	0.2 to 15 km
Vertical resolution	7.5 m	10 m
Temporal resolution	30 s	12 s
Data coverage	5 profiles between 13.9. and 15.9.	Whole campaign

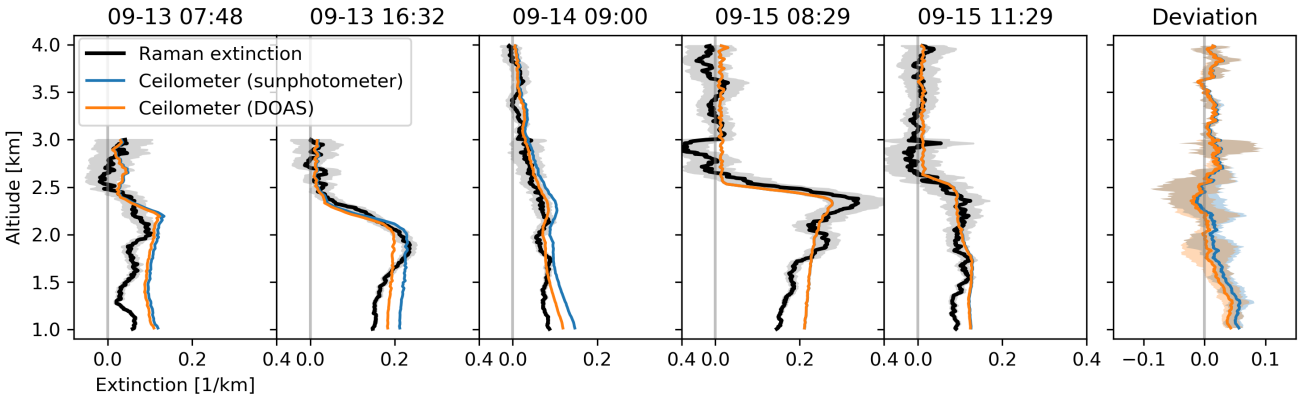


Figure S5. Comparison of the five available aerosol extinction profiles obtained from Raman lidar measurements (in black, with uncertainties indicated by the grey areas) with AOT scaled ceilometer backscatter profiles. Scaling was performed with AOTs from sun photometer (blue) and DOAS data (orange). In the last plot, lines show the mean deviation, whereas the borders of the couloured areas mark maximum and minimum deviation.

S4.2 Radiosonde flights

An overview of the radiosonde flights is given in Fig. S6. Example profiles are shown in the course of a comparison between lidar and radiosonde observations in Supplement S4.5. Another preprocessing step of the sonde profiles shall be mentioned: data quality was affected by calibrational drifting of the sensor, as it was optimized for low weight and cost rather than performance. Even though a calibration against the CE-DOAS was performed directly before each launch, most profiles showed a clear instrumental offset of about $(1-2)\times10^{10}\,\mathrm{molec\,cm^{-3}}$ in the free troposphere. The offset was subtracted and the profiles were subsequently rescaled to their initial surface concentration.

Table S3. Overview over the radiosonde sampling flights shown in this study.

Launch date	Flight time[a] [min]	Travel distance[a] [km]	Wind direction
9-13 08:42	10	7	SE
9-14 09:03	12	5	SE
9-14 13:06	14	4	SE
9-15 08:04	10	8	E
9-15 10:25	11	8	SE
9-21 07:57	12	10	SE
9-21 10:14	15	5	SE
9-25 06:59	17	7	S
9-25 09:29	12	18	S

[a] Only considering trajectory through the lowest 4 km of the atmosphere.

S4.3 NO_2 lidar

The NO_2 lidar provides profiles consisting of a series of altitude intervals or "boxes" with constant gas concentration between a lower and an upper altitude limit. The conversion to profiles on the MAX-DOAS 200 m grid is demonstrated in Fig. S7. First, the boxes were converted to a continuous profile by linearly interpolating over box overlaps or gaps, which was then averaged down to the 200 m MAX-DOAS retrieval grid resolution.

S4.4 Long path DOAS

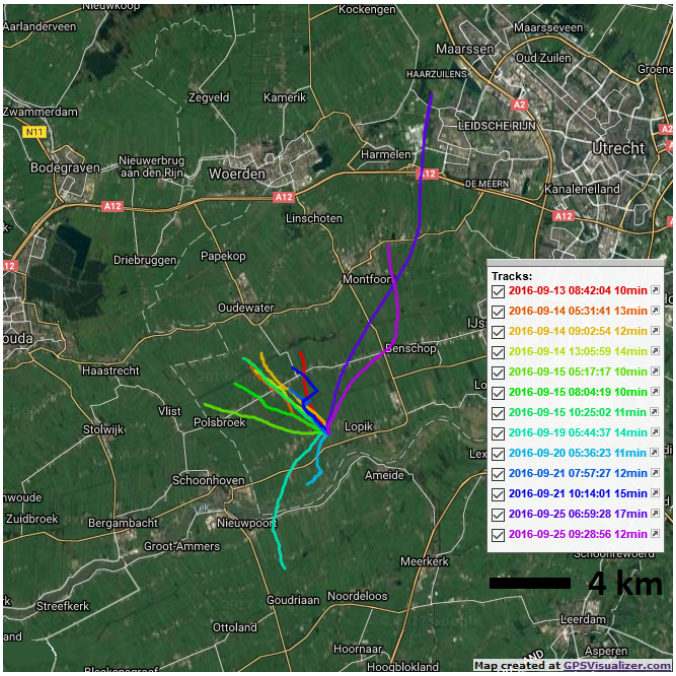


Figure S6. Sonde flight paths in the course of the campaign. Only data of the sonde ascent through the lowest 4 km of the atmosphere are shown. Some flights shown here are not included in the comparison as they were launched before 6:30 h. The "minute" values in the legend labels represent the flight time.

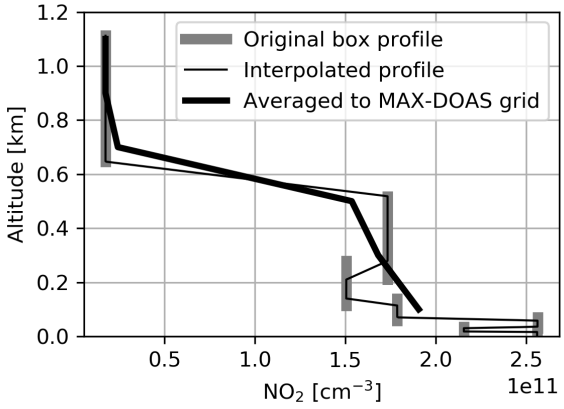


Figure S7. Regridding of an example NO_2 lidar box profile (27 September 2016, 11:00) to the MAX-DOAS 200 m vertical resolution. In a first step, gaps and overlaps within the box profile are linearly interpolated. The resulting profile is then averaged within the MAX-DOAS retrieval layers.

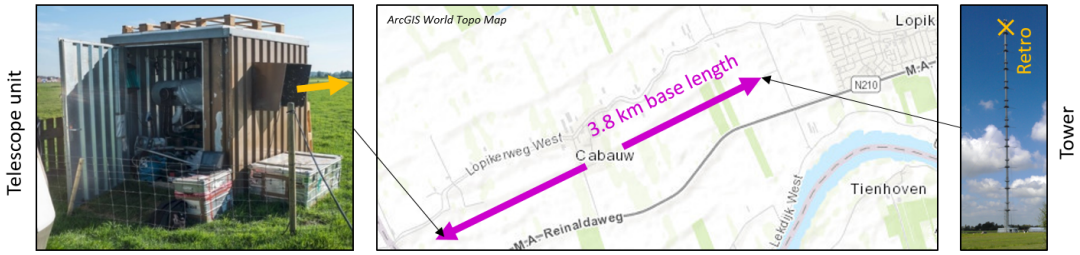


Figure S8. Setup of the LP-DOAS system. As shown on the map (Esri et al., 2018), the light sending and receiving telescope unit (left) was located at 3.8 km distance to the meteorological tower (right), resulting in a total light path of 7.6 km. There were several retroreflectors installed on the tower at different altitudes. However for this study, only the one at the very top (207 m altitude) was used to obtain the average gas in the lowest retrieval layer, extending from 0 to 200 m altitude.

S4.5 Consistency of supporting observations

The agreement of redundant supporting observations (processed as described in the main text Sect. 2.2) gives an impression of their reliability and/ or representativeness. In the case of NO_2 several observations of total vertical columns and surface concentration (note again, that throughout this paper "surface concentration" refers to the average concentration in the lowest retrieval layer) are available and compared in the correlation plots in Fig. S9 below. Corresponding time series plots are already shown in the main text in Fig. 17 and 20, respectively. Tables S4 and S5 show the RMSD (as observed) and σ (the expected deviation according to the specified measurement uncertainties) between each possible pair of observations. For the VCDs, the RMSD is close to σ or below. A maximum RMSD of 1.5 σ is found between NO_2 lidar and direct-sun DOAS. For the surface concentrations however, there seem to be systematic deviations which split the observations into two pairs: radiosonde and lidar observations agree well but are both systematically lower than LP-DOAS and tower measurements. MAX-DOAS UV agrees better with LP-DOAS and tower observations, while MAX-DOAS Vis agrees more with sonde and lidar. Between the LP-DOAS and the NO_2 lidar, an RMSD of more than 4 σ is observed. There are several potential explanations:

1. Biases are introduced due to data processing (temporal and spatial regridding, for instance for the lidar profiles described in Sect. S4.3).
2. Spatio-temporal variability of the real gas abundances
3. Imperfect estimates of the measurement uncertainties (in particular systematic deviations)

For the NO_2 lidar and the radiosondes, there are four simultaneously recorded NO_2 profiles available over the campaign (simultaneous in the sense that for a single MAX-DOAS profile timestamp, profiles from both systems are available according to the definitions in Sect. 2.2.3 in the main text). They are compared in the top row of Fig. S10. For the first situation, where good spatial and temporal overlap is given, there is mostly an agreement within the specified errors. In the case of bad temporal and/ or spatial overlap, strong deviations occur. For the 2nd and the 4th plot, there are several lidar profiles available, which

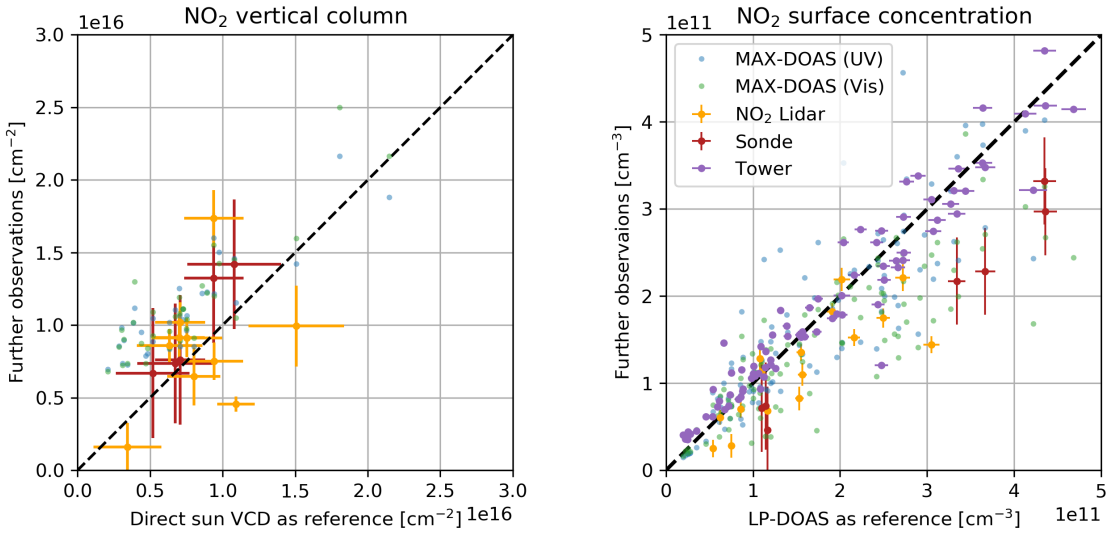


Figure S9. Comparison of redundant supporting observations of NO_2 VCDs (left panel) and surface concentration (right panel). MAX-DOAS retrieved values are plotted in the background. To improve visibility, tower measurement uncertainties (vertical error bars of typically $(6.0 \pm 0.5) \times 10^{10}\,\mathrm{molec\,cm^{-3}}$) are not shown.

are temporally closer to the radiosonde (however further away from the corresponding MAX-DOAS profile), which in contrast show very good agreement again. This shows that the real NO_2 profile varies strongly even on timescales of ≈ 30 minutes (see also (Peters et al., 2019)) and that improved synchronisation between MAX-DOAS and supporting observations should be considered for future campaigns.

Table S4. Comparison of redundant measurements of the NO_2 surface concentration (in $10^{11}\,\mathrm{molec\,cm^{-3}}$). For each pair of observations, the observed scatter (RMS) is compared to the specified uncertainty (σ).

	Tower in-situ (0.56)		Radiosonde (0.50)		NO_2-Lidar (0.13)	
	RMSD	σ	RMSD	σ	RMSD	σ
LP-DOAS (0.06)	0.32	0.56	1.01	0.51	0.57	0.13
NO_2-Lidar (0.13)	0.72	0.57	0.40	0.52	-	-
Radiosonde (0.50)	0.99	0.78	-	-	-	-

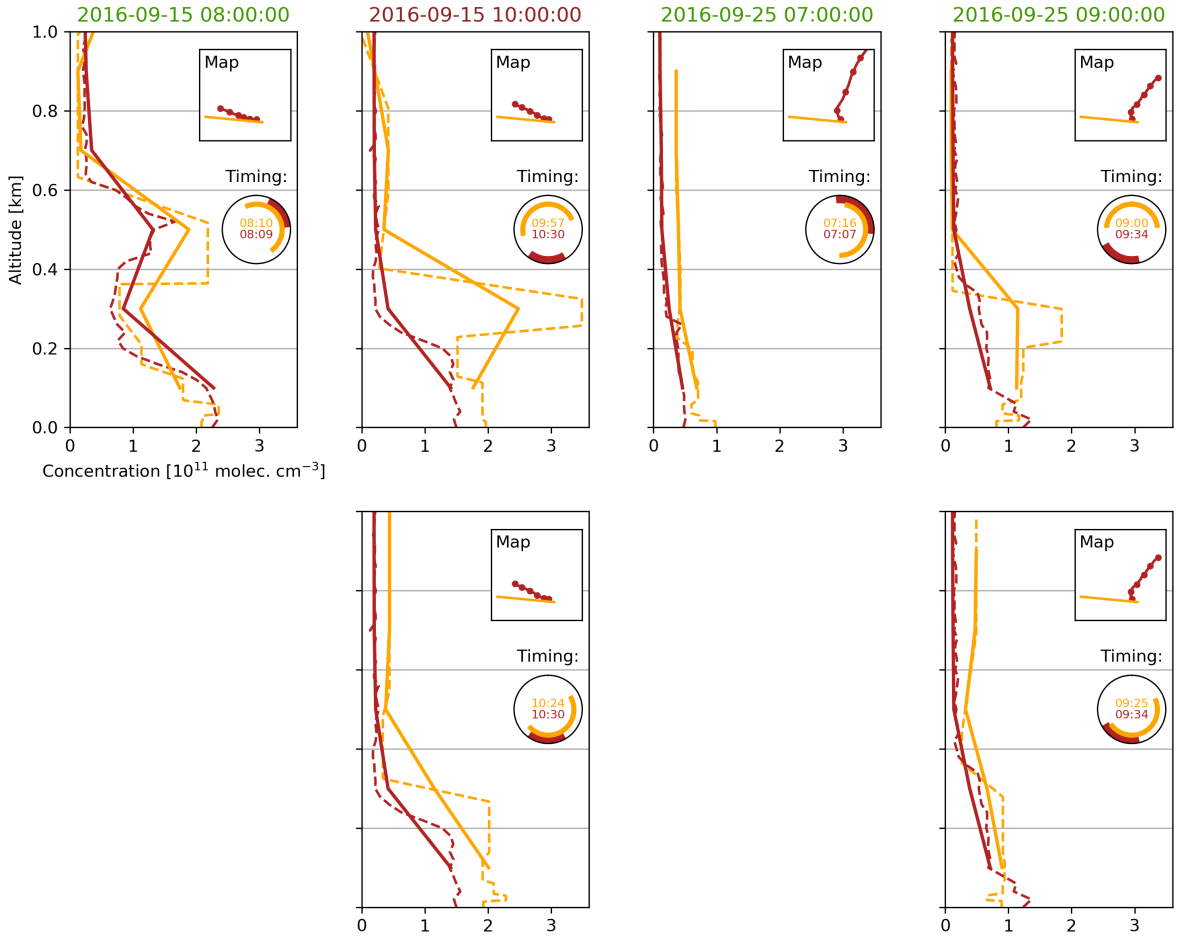


Figure S10. Top row: Comparison of NO_2 lidar (orange) and radiosonde (red) profiles, that were assigned to a common MAX-DOAS profile timestamp according to Sect. 2.2.3 in the main text. Dashed lines represent original instrument resolution while thick lines show the concentrations averaged to the MAX-DOAS altitude grid. The colour of the dates indicates the cloud conditions. The rectangular little subplots show a map (4 x 4 km) of the lidar line of sight and the sonde flight path. The dots on the sonde flight path mark the transitions between the retrieval layers. The polar plot (to be read like a clock) shows the temporal overlap between the two observations, together with the middle timestamps of each observation. Lower row: For the 2nd and 4th timestamp, there were lidar profiles available with improved temporal overlap (however, with a worse overlap with the corresponding MAX-DOAS profile).

Table S5. Comparison of redundant measurements of the NO_2 total columns (in $10^{16}\,\mathrm{molec\,cm^{-2}}$). For each pair of observation, the observed scatter (RMS) is compared to the specified uncertainty (σ).

	Radiosonde (0.44)		NO_2-Lidar (0.15)	
	RMSD	σ	RMSD	σ
Direct-sun DOAS (0.23)	0.24	0.51	0.40	0.26
NO_2-Lidar (0.15)	0.34	0.48	-	-

S5 MAX-DOAS viewing distance

Wagner and Beirle (2016) derived polynomial relationships between the "horizontal sensitivity range" (HSR, defined as the distance, at which the box airmass factors dropped to 1/e) and O_4 differential airmass factors (dAMF). Applying this approach to the CINDI-2 O_4 dAMFs yields the HSRs shown in Fig. S12. A constant vertical O_4 column of $1.19 \cdot 10^{43}\,\mathrm{molec^2\,cm^{-5}}$ was assumed. The HSR for the actually retrieved layers is more complicated and not assessed here, as information aspects (which elevation contributes to information on which layer), geometrical limitations and the atmospheric state (trace gas and aerosol layer height) would have to be taken into account. Depending on the conditions, HSRs vary between a few and tens of kilometres (as shown in Fig. S11) defining whether only air masses over rural areas and/or urban areas (Gouda at 15 km distance, Zoetermeer at 30 km distance and The Hague at 40 km distance. to the measurement site) are sampled. Further, depending on the wind, plumes of Utrecht, Rotterdam or Amsterdam might be sampled.

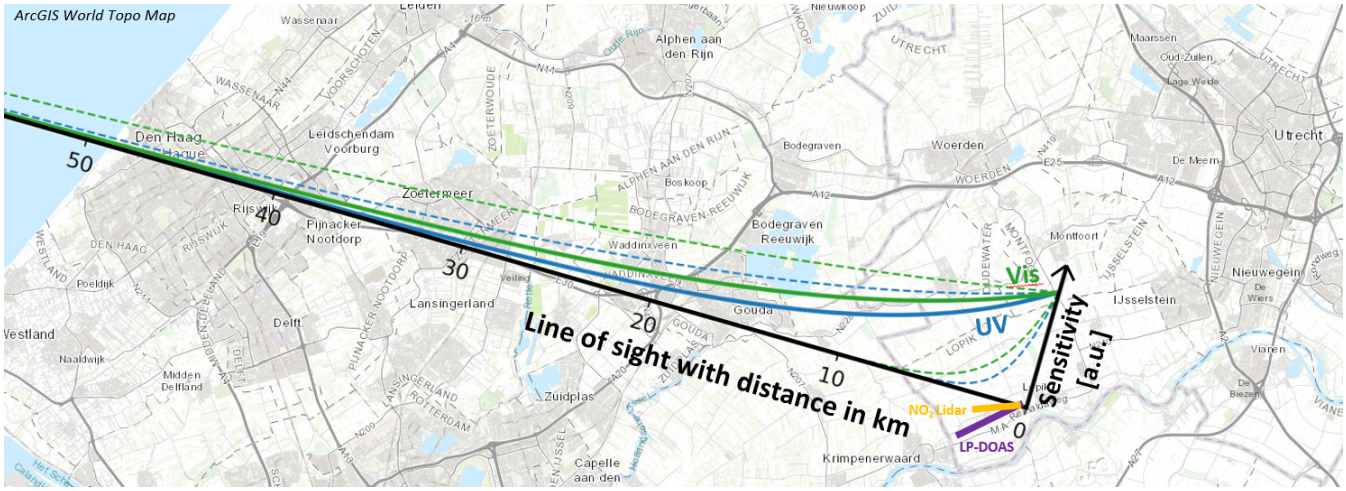


Figure S11. Line of sight of the MAX-DOAS instruments on a map (Esri et al., 2018). The coloured curves indicate the sensitivity for the ~~two extreme~~ average (~~shortest~~ solid), the minimum (dashed) and ~~longest~~maximum (dashed) viewing distances encountered during the campaign ~~. Dashed lines indicate~~ in the ~~1/e-length~~UV (blue) and Vis (green).

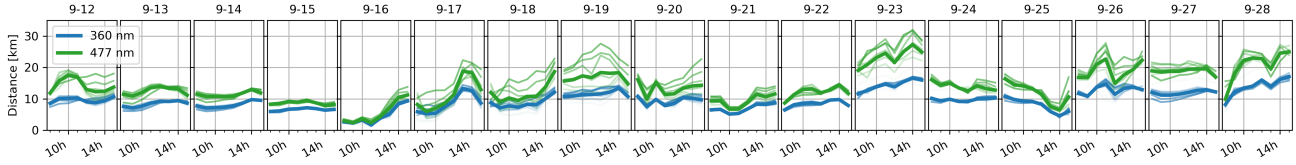


Figure S12. Viewing distance (HSR) of MAX-DOAS instruments during CINDI-2. It was calculated for different ~~elevations~~ elevation angles (1, 2, 3, 4, 5, 6 and 8° with increasing transparency of the curves) and the average value for UV and Vis (thick lines).

S6 ~~Spatio-temporal variability~~

~~For an impression of the impact of~~

S6 Spatio-temporal mismatch and variability

Given the MAX-DOAS horizontal sensitivity ranges determined in Sect. S5, approximate values for the spatio-temporal ~~variability on the comparisons, different~~ mismatch of MAX-DOAS and different supporting observations can be derived. They are given in Table S6. The potential impact of these mismatches can be demonstrated by means of the NO_2 surface concentration. The left panel of Fig. S13 shows observations of the NO_2 surface concentrations at their original temporal resolution Δt and integration time t_{int}~~are shown in the left panel of Fig. S13.~~ . The CE-DOAS as a point measurement with $\Delta t = t_{int} = 1\,\mathrm{min}$ shows very strong variability on short timescales. However, for the tower measurements (all in situ instruments in the tower vertically integrated as described in main text Sect. 2.2.5 at $\Delta t = 20\,\mathrm{min}$ and $t_{int} \approx 5\,\mathrm{min}$), the LP-DOAS

Table S6. Estimates for the average spatio-temporal mismatch of different supporting observations w.r.t. to the MAX-DOAS measurements. For the location of the MAX-DOAS observations the centers of mass of the horizontal sensitivity curves from Sect. S5 were used. For the location of sun photometer and direct-sun DOAS observations, the center of the lines of sight towards the sun up to 2 km atitude were considered.

Observation	Spatial mismatch [km]	Temporal mismatch [min]
Sun photometer	13	8
Ceilometer	11	0
Direct-sun DOAS	13	23
NO_2-Lidar	10	9
Radiosonde	6	13
LP-DOAS	10	6
In-situ in tower	11	0

($\Delta t = 32\,\mathrm{min}$, $t_{int} \approx 100\,\mathrm{s}$) there is already significant smoothing. The 1D-MAX-DOAS data was recorded by DLR (see Supplement S10), who retrieved profiles in the nominal azimuth direction (287°) more or less continuously ($\Delta t = 15\,\mathrm{min}$, $t_{int} \approx 10\,\mathrm{min}$). In all measurements there is significant variation on the sub-hour timescale. Further, spatial variability might be observed in the form of disagreement between UV and Vis observations of the 1D-MAX-DOAS as viewing distance and thus the sampled air volume changes between the two spectral ranges (see Supplement S5). To estimate the order of magnitude, the right panel of Fig. S13 shows a kind of autocorrelation of the total campaign time series of each observation. The RMSD between the original and a temporally shifted signal is calculated. ~~Note, that the~~ 1D MAX-DOAS Vis data is not ~~reliable as it features a lot of gaps . As described in main text Sect. 2.2, typical temporal shifts/ interpolation intervals between different observations during CINDI-2 are in the order of 30 minutes. RMSD values at 30 minutes shift are~~ shown, as multiple gaps in the data complicated the autocorrelation. Comparing this figure with values from Table S6 yields that spatio-temporal variability causes RMSD values of around $3.5 \times 10^{10}\,\mathrm{molec\,cm^{-3}}$ ~~and thus indeed of similar order as typical~~ in the NO_2 surface concentration, which is indeed of the order of the observed RMSD values in the NO_2 surface concentration comparisons within this study (approx. $5 \times 10^{10}\,\mathrm{molec\,cm^{-3}}$, compare to main text Fig. 22). ~~For an impression on~~

For another demonstration of the spatial variability, we refer to data from the IMPACT instrument (Peters et al., 2019), an imaging MAX-DOAS operated by IUP-Bremen (IUPB) which allows to perform elevation "scans" in different azimuth viewing directions in quick succession. During CINDI-2, the IMPACT performed full-azimuthal scans in 10° steps every 15 minutes. Figure S14 exemplarily shows the observed NO_2 Vis dSCDs at 4° elevation on the 20 September 2016 together with dSCDs measured by the IUPB standard MAX-DOAS instrument in the nominal azimuth direction (287°, compare main text Sect. 2.1). The red shaded area depicts the variation of the dSCD with azimuth viewing direction. In particular around local noon this variation is tremendous, exceeding a factor of five. Further investigation on this issue can be found in Peters et al. (2019).

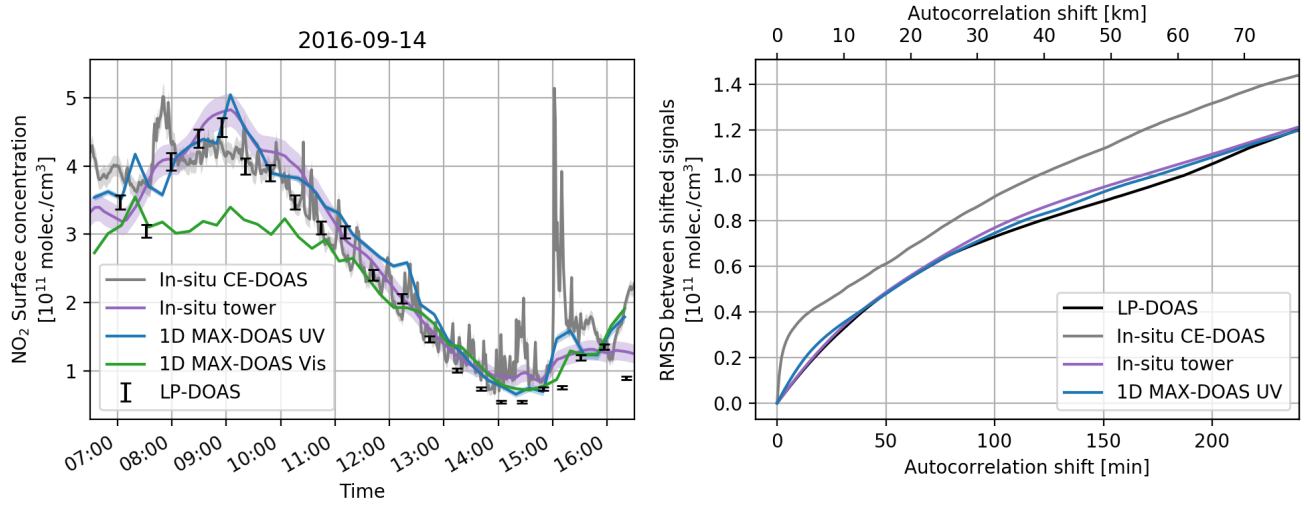


Figure S13. Left: Different observations of the NO_2 surface concentrations on 14 September 2016, each at its original temporal resolution to reveal short-term variations. Coloured areas behind the lines indicate the specified uncertainties. Right: RMSD values obtained from a kind of autocorrelation analysis over the whole campaign (night times excluded). For each observation, the RMSD between the original and a temporally shifted signal is calculated. The temporal shift (bottom horizontal axis) was varied between 0 and 4 hours. The temporal shift was roughly converted to its spatial equivalent by multiplication with the average observed wind speed in the surface layer ($\approx 5\,\mathrm{m/s}$), yielding the top horizontal axis.

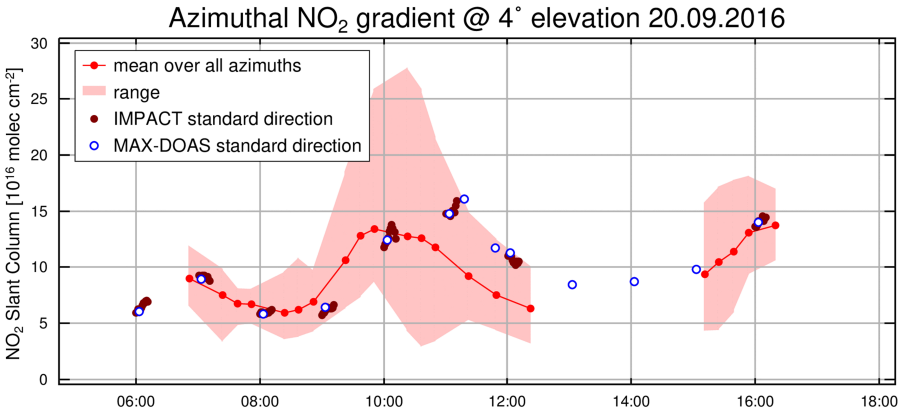


Figure S14. Variation in the NO_2 Vis dSCDs with different azimuth viewing directions at 4° elevation, as observed by the IMPACT imaging MAX-DOAS (Peters et al., 2019). Around local noon this variation is largest, exceeding a factor of 5. The time is UTC.

S7 Impact of the choice of pressure and temperature profiles for the RTMs

Pressure (p) and temperature (T) profiles used for the RTMs within this study are averaged sonde measurements performed in De Bilt by KNMI during September months of the years 2013-2015 (see main text Sect. 2.1.3). To estimate the effect of this approximation on the results, IUPHD/ HEIPRO retrieved an additional set of profiles, using p and T information from radiosondes launched at KNMI (De Bilt) during the campaign. Between one and three sondes were launched every day except on 16 September. For each profile inversion, the temporally closest sonde observation was used. Table S7 shows the difference in RMSD and Bias magnitude between these results and the "standard" results of IUPHD/ HEIPRO (that used the prescribed averaged p and T profiles from years before) relative to the average RMSDs and average Bias magnitude for all participants.

The impact on the dSCD comparison is less than 5% for both, RMSDs and Bias magnitudes. For AOTs, VCDs and surface concentrations, significant improvement ($> 10\,\%$ in RMSD) is only observed for HCHO surface concentrations ($17\,\%$) that contrasts with a deterioration for UV AOTs by 13%. The average improvement in RMSD for AOTs, VCDs and surface concentrations is $3.2\,\%$. The overall consistency between MAX-DOAS and supporting observations can thus be considered to remain similar, despite larger changes in some Bias magnitudes are observed (up to $51\,\%$ improvement for NO_2 Vis surface concentrations and up to $20\,\%$ deterioration for UV AOTs).

Table S7. The differences in RMSDs and Bias magnitudes for the IUPHD/ HEIPRO results arising from using daily p and T profiles, relative to the average RMSDs and Bias magnitudes assessed within the main study. Values are given for the comparisons of modelled and measured dSCDs ("dSCDs") and the comparisons against the supporting observations of AOTs, VCDs and surface concentrations as described in the main text. Minus signs indicate improvement. Only clear sky conditions were considered.

	dSCDs		AOT/VCD		Surface	
	ΔRMSD [%]	ΔBias [%]	ΔRMSD [%]	ΔBias [%]	ΔRMSD [%]	ΔBias [%]
HCHO	2.7	3.5	6.8	10.5	-17.4	-22.0
NO_2 UV	-0.7	-1.1	-2.7	-2.6	-3.5	8.7
NO_2 Vis	-0.7	-3.3	-0.8	-1.0	-2.8	-50.9
Aerosol UV	-0.7	0.7	12.5	20.2	-	-
Aerosol Vis	-0.2	2.1	-8.7	-40.1	-	-

S8 Further details on the comparison results

S8.1 AVKs of individual participants

Figures S35 to S39 show the averaging kernels (AVKs) and retrieved degrees of freedom of signal (DOFS) of each participant for aerosol UV. For explanation of colours and symbols please refer to main text Sect. 3.1. The DOFS values in brackets were calculated considering valid data only.

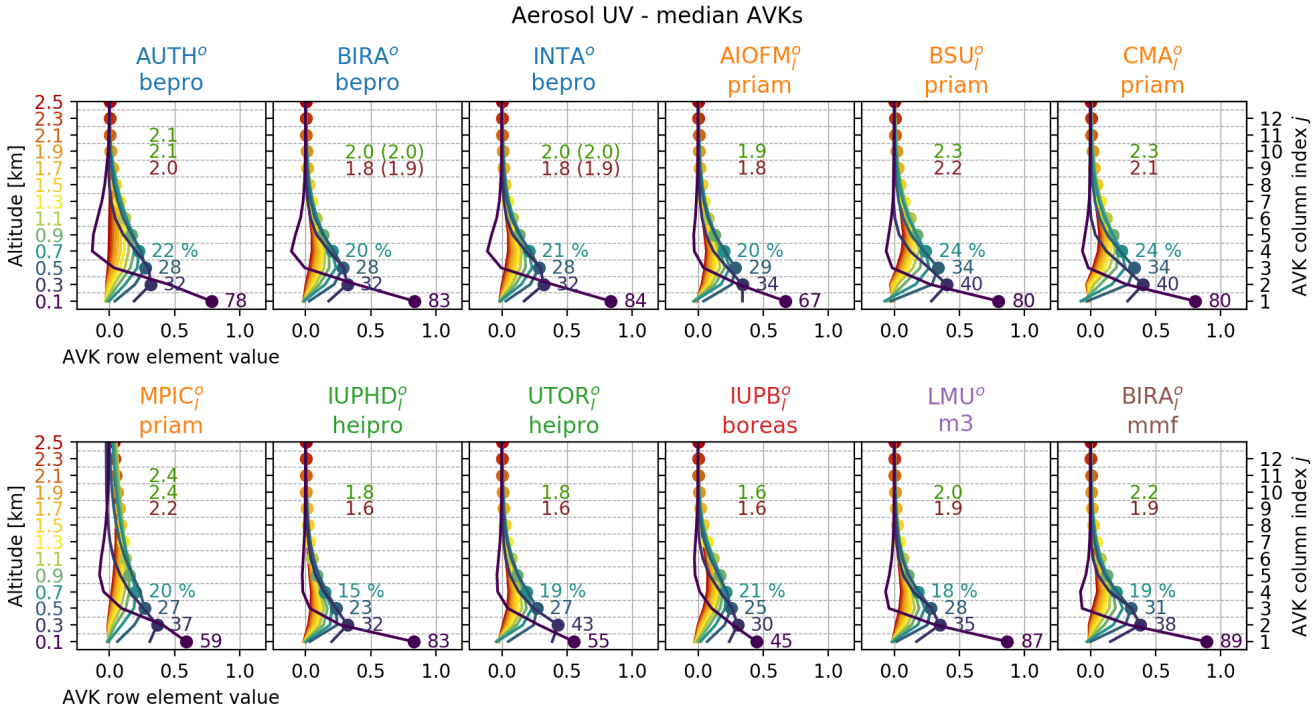


Figure S15. Mean averaging kernels for Aerosol UV for each participant. Coloured values at AVK peaks show the amount of retrieved information on the respective layer in percent. "DOFS" numbers are given for clear-sky (green) and cloudy (red) conditions. Values in brackets are DOFS including flagging.

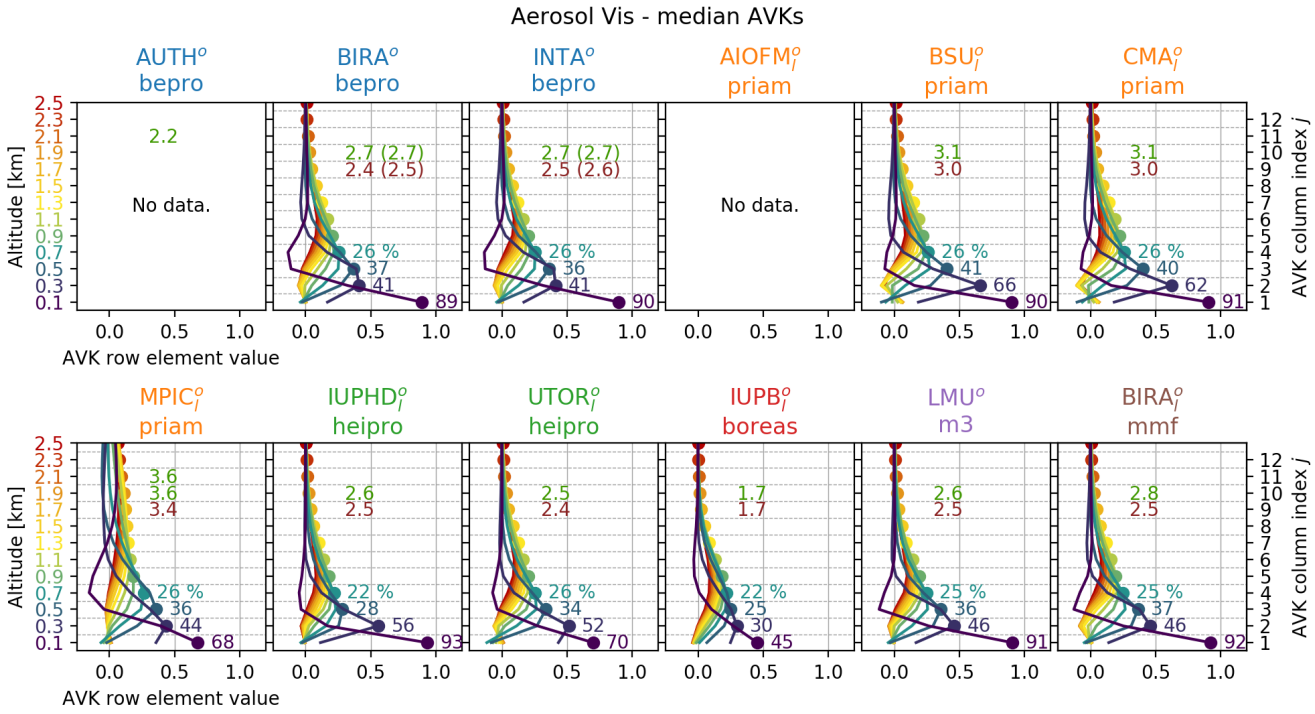


Figure S16. Mean averaging kernels for Aerosol Vis for each participant. Description of Fig. S35 applies.

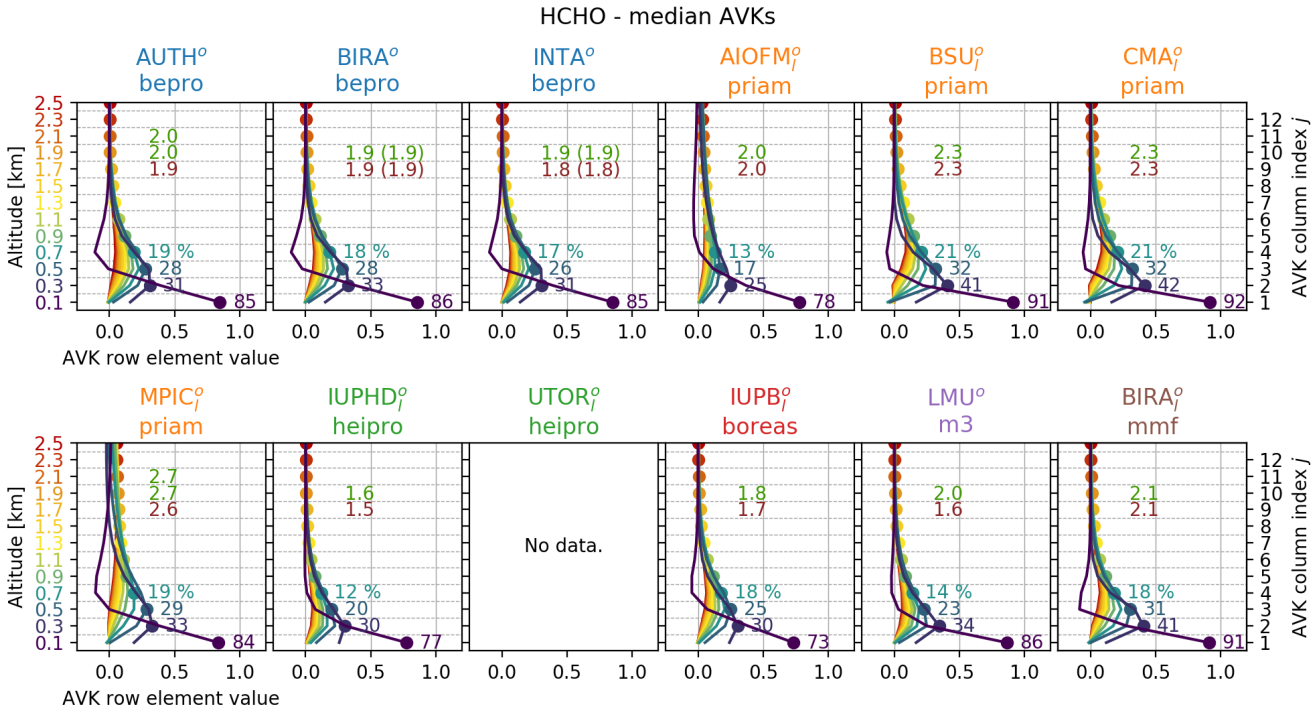


Figure S17. Mean averaging kernels for HCHO for each participant. Description of Fig. S35 applies.

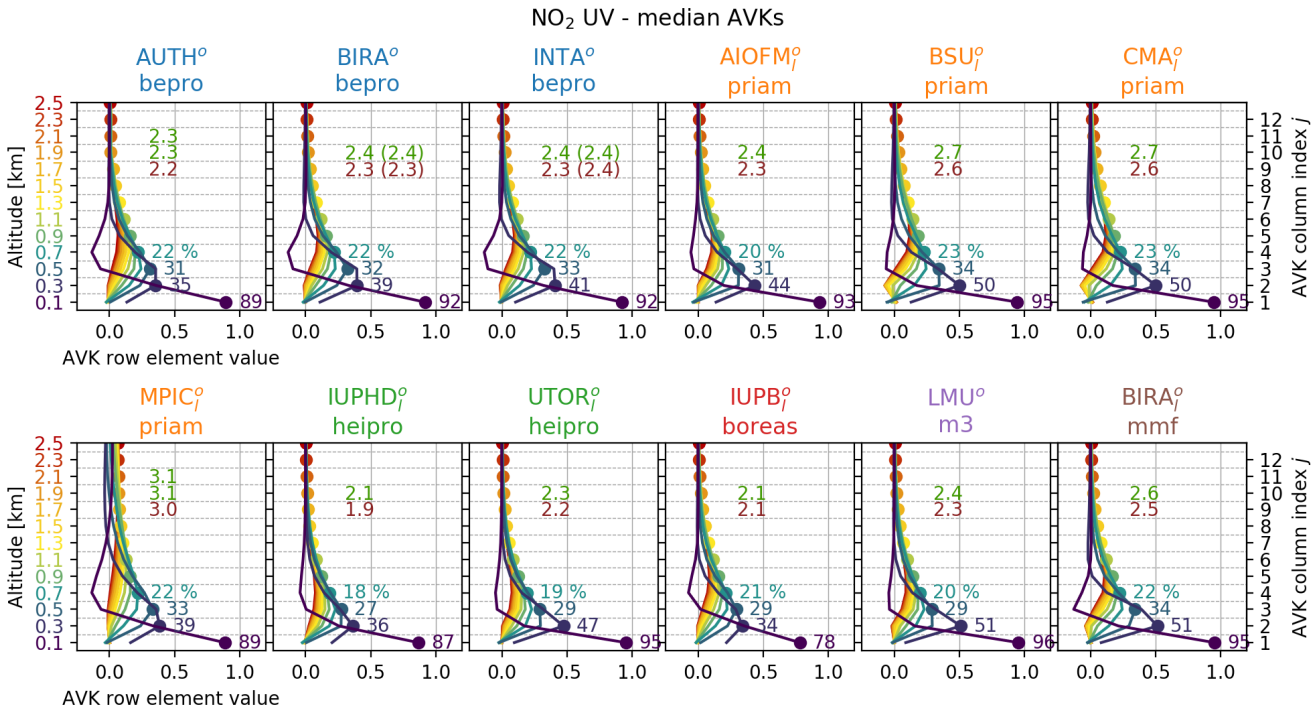


Figure S18. Mean averaging kernels for NO_2 UV for each participant. Description of Fig. S35 applies.

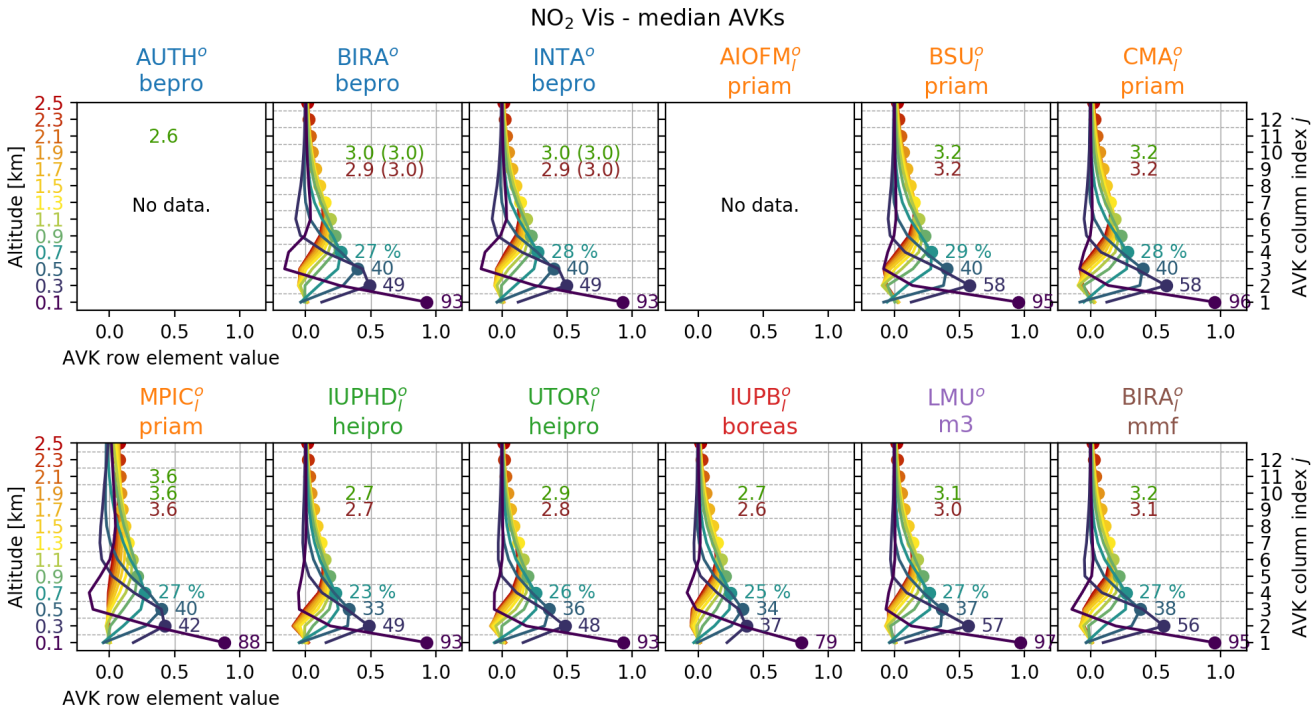


Figure S19. Mean averaging kernels for NO_2 Vis for each participant. Description of Fig. S35 applies.

S8.2 Profile deviation statistics

Figures S20 to S24 show statistics on the observed differences in the retrieved profiles for all five species. The plots on the left compare the retrieved profiles of individual participants $\boldsymbol{x}$ to the median MAX-DOAS profiles $\bar{\boldsymbol{x}}$. While the vertical axes represent altitude, the horizontal axes depicts the difference $\boldsymbol{x} - \bar{\boldsymbol{x}}$. The coloured boxes indicate the $25\,\% - 75\,\%$ percentile, whiskers are $5\,\% - 95\,\%$. Black dots indicate the mean value. For each layer there are boxplots for clear-sky (green) and cloudy conditions (red). Note that for aerosol there are two different horizontal axes defined for the two cloud conditions: the green scale at the bottom and the red scale on the top of each plot. Only valid data (flagged) was considered. For aerosol and NO_2 a plot on the very right shows statistics of the difference of supporting measurements $\boldsymbol{x_{anc}}$ (lidar/ radiosonde for NO_2, sun photometer scaled ceilometer for aerosol) to the median $\bar{\boldsymbol{x}}$, hence $\boldsymbol{x_{anc}} - \bar{\boldsymbol{x}}$. The numbers in all the plots show RMSD deviation of the three lowest (most sensitive) layers. Dashed lines indicate the median retrieval uncertainty as specified by the participants.

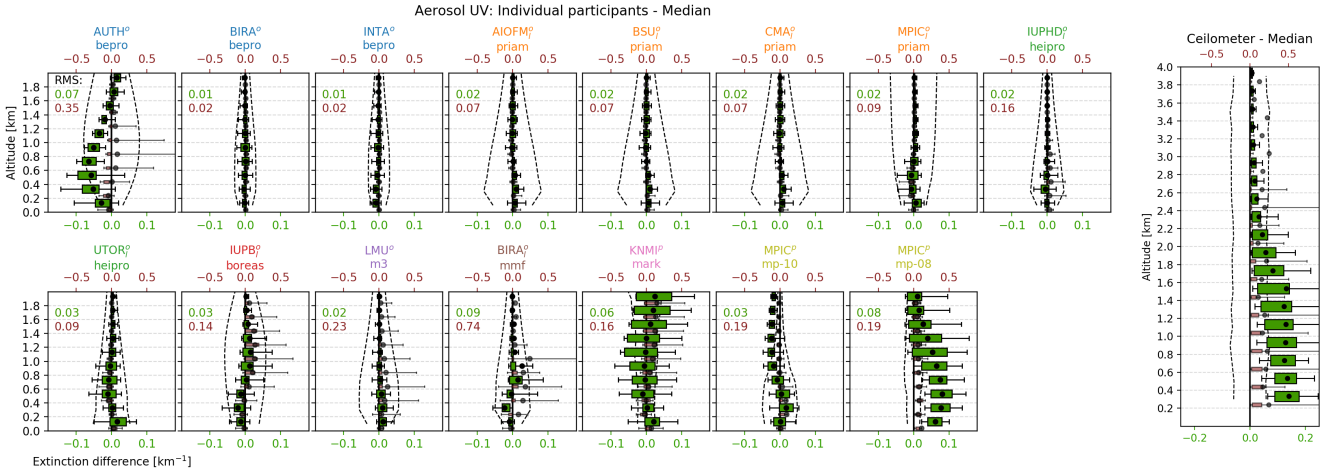


Figure S20. Left: ~~Flagged deviations for~~ Deviations of Aerosol UV profiles (valid only) of individual participants from the MAX-DOAS median profiles. Dots show the mean, boxes indicate the (25%-75%) percentile, and whiskers show (25%-75%) percentile. Green (red) box-whiskers represent clear (cloudy) conditions. Note, that there are different x-scales (on top and bottom of the plot) for different cloud conditions. The average standard deviations specified by the participants are indicated by the dashed lines. Right: Deviation of the AOT scaled ceilometer backscatter signal to the MAX-DOAS median profiles. The numbers in the plots indicate RMSD values for clear sky (green) and cloudy (red) conditions. Further details are given in the related text in Sect. S8.2

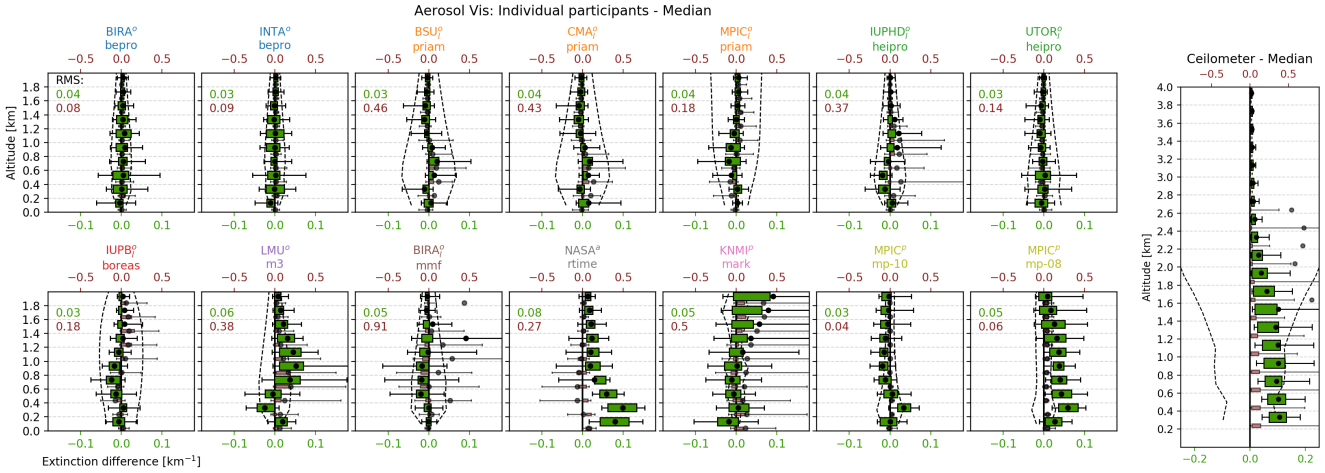


Figure S21. ~~Flagged deviations for~~ Deviations of Aerosol Vis profiles (valid only) of individual participants from the MAX-DOAS median profiles. Description of Fig. S20 applies.

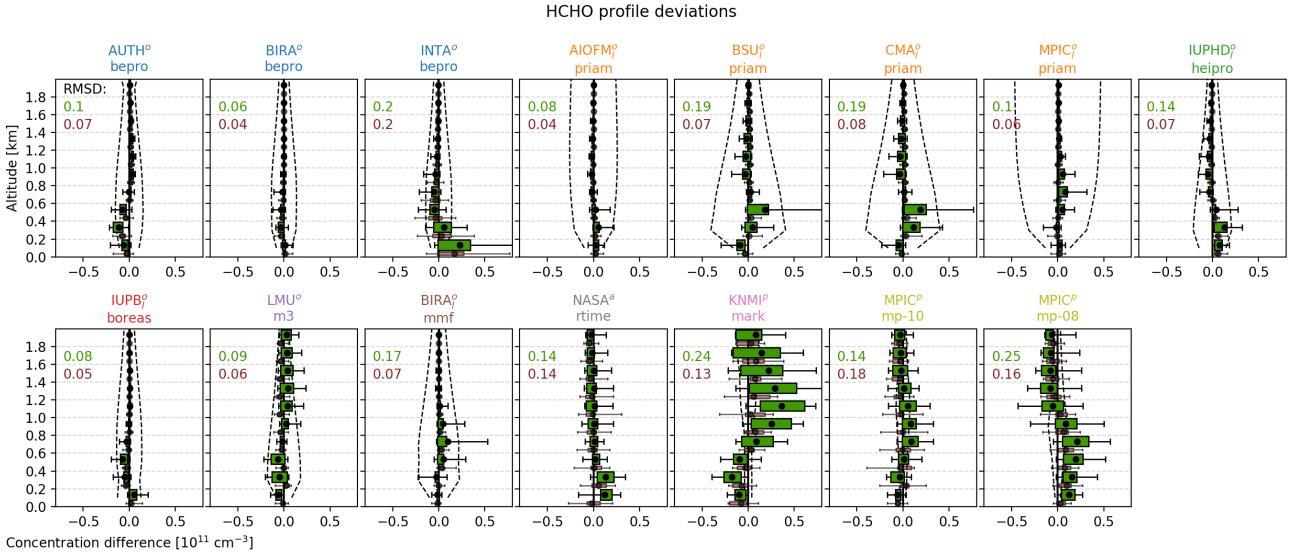


Figure S22. ~~Flagged deviations for~~ Deviations of HCHO profiles (valid only) of individual participants from the MAX-DOAS median profiles. The description of Fig. S20 applies but for HCHO, there is no independent reference profile available.

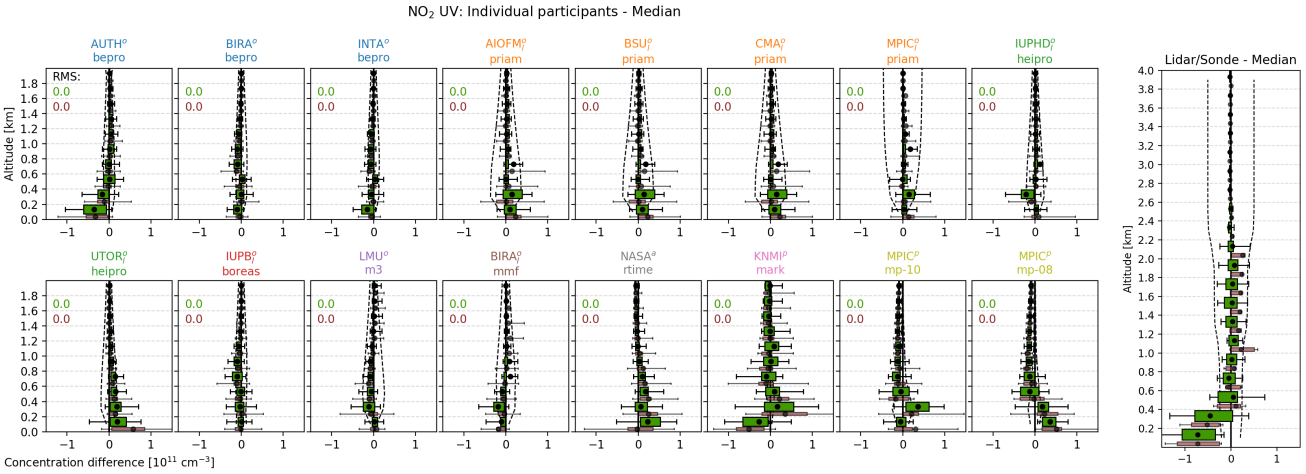


Figure S23. ~~Flagged deviations for~~ Deviations of NO_2 UV profiles (valid only) of individual participants from the MAX-DOAS median profiles. The description of Fig. S20 applies. On the right, deviation of the median retrieved profiles from the few available NO_2 lidar and sonde profiles are shown.

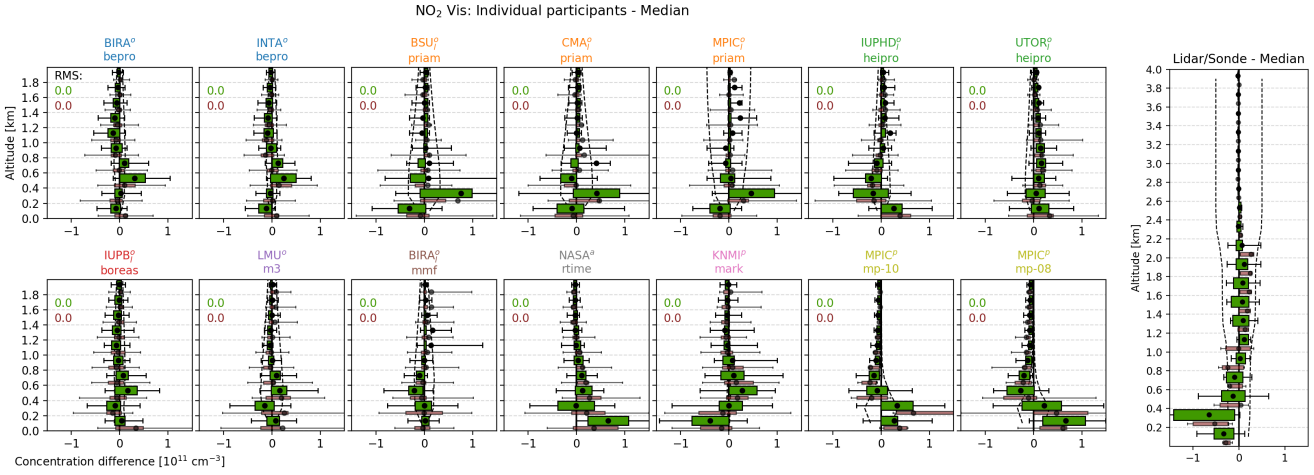


Figure S24. ~~Flagged deviations for~~ Deviations of NO_2 ~~UV~~ Vis profiles (valid only) of individual participants from the MAX-DOAS median profiles. The description of Fig. S23 applies.

S8.3 Correlation plots for AOTs, VCDs and surface concentrations

Figures S25 to S32 show the individual correlation plots for the comparisons of AOTs, VCDs and surface concentrations as summarized in Sect. 3.4, 3.5 and 3.6 in the main text, respectively. The colours indicate cloud conditions: clear-sky (green) and cloudy conditions (red). Transparent markers represent data points flagged as invalid. The small grey bars indicate uncertainties in the measurement. For the median values, the bars show the standard deviation among the participants (valid data only). Dashed lines in the correlation plots represent the ideal 1:1 line. Correlations were performed separately against MAX-DOAS median values and supporting observations. For AOTs a third correlation is shown for the partial AOT (with PAC applied).

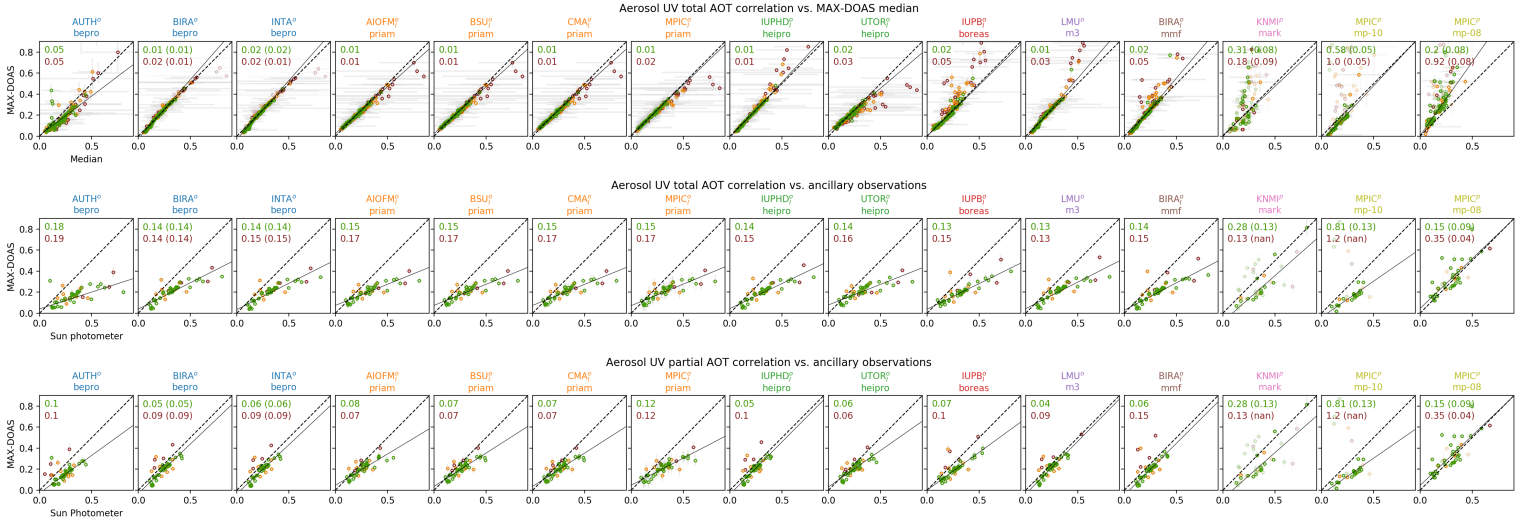


Figure S25. Correlation of Aerosol UV AOTs. Complementary to main text Sect. 3.4

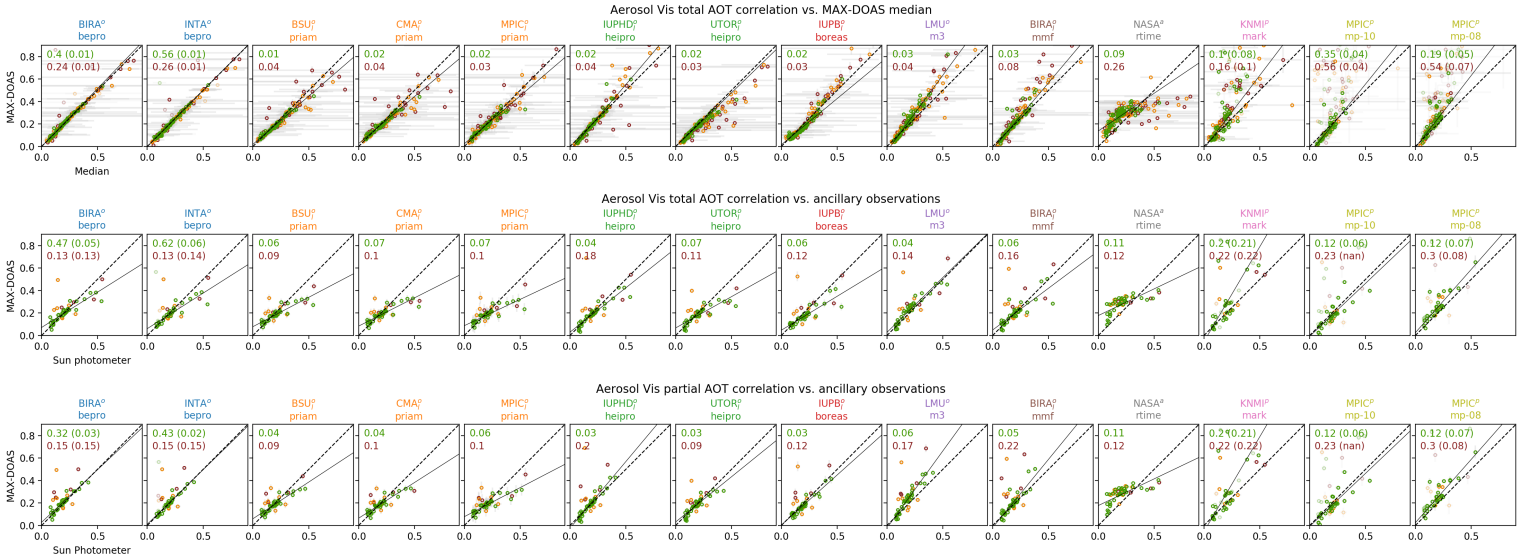


Figure S26. Correlation of Aerosol Vis AOTs. Complementary to main text Sect. 3.4

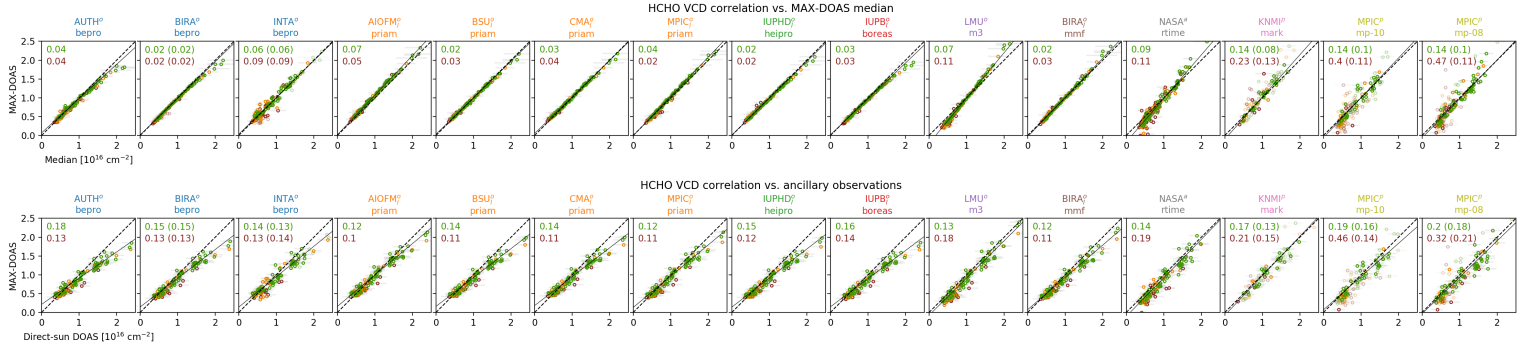


Figure S27. HCHO VCD correlation plots. Complementary to main text Sect. 3.5

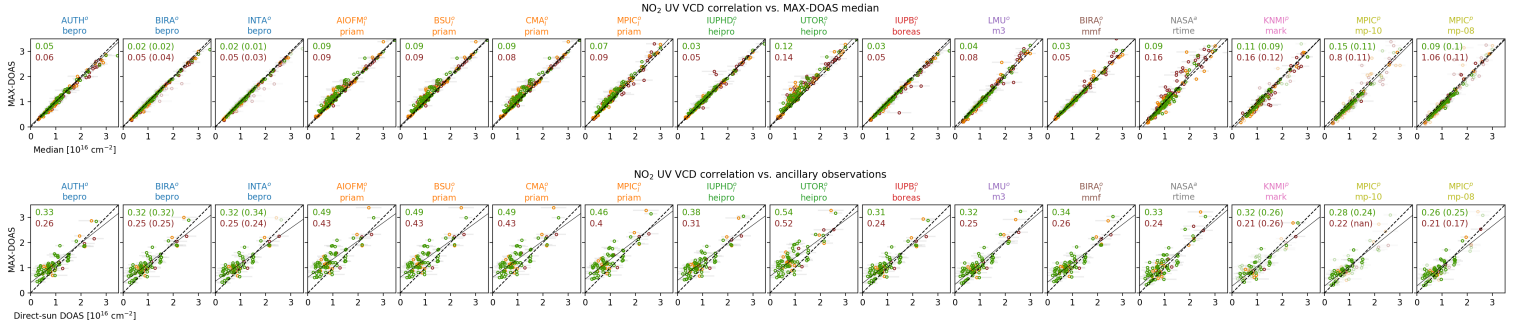


Figure S28. NO_2 UV VCD correlation plots. Complementary to main text Sect. 3.5

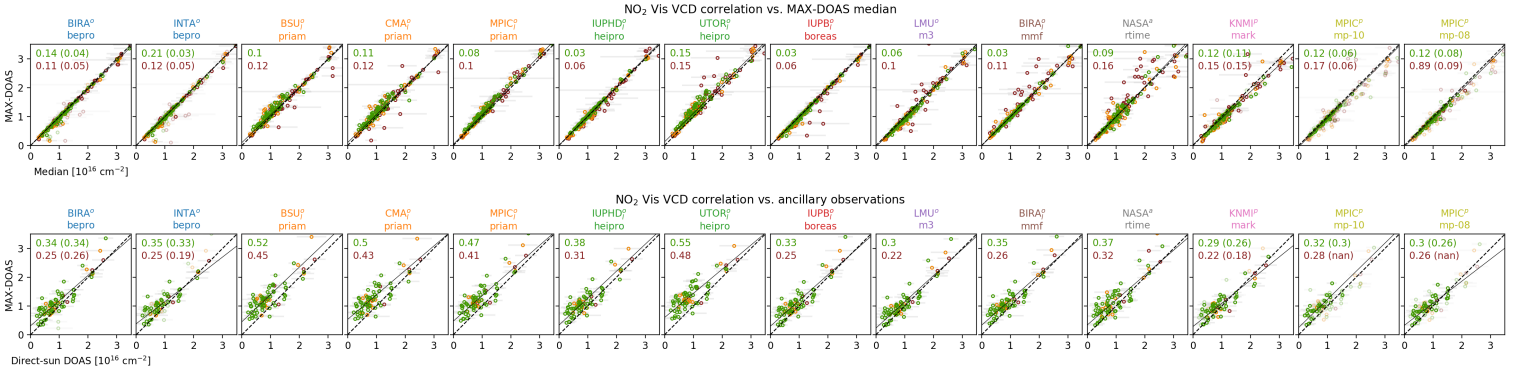


Figure S29. NO_2 Vis VCD correlation plots. Complementary to main text Sect. 3.5

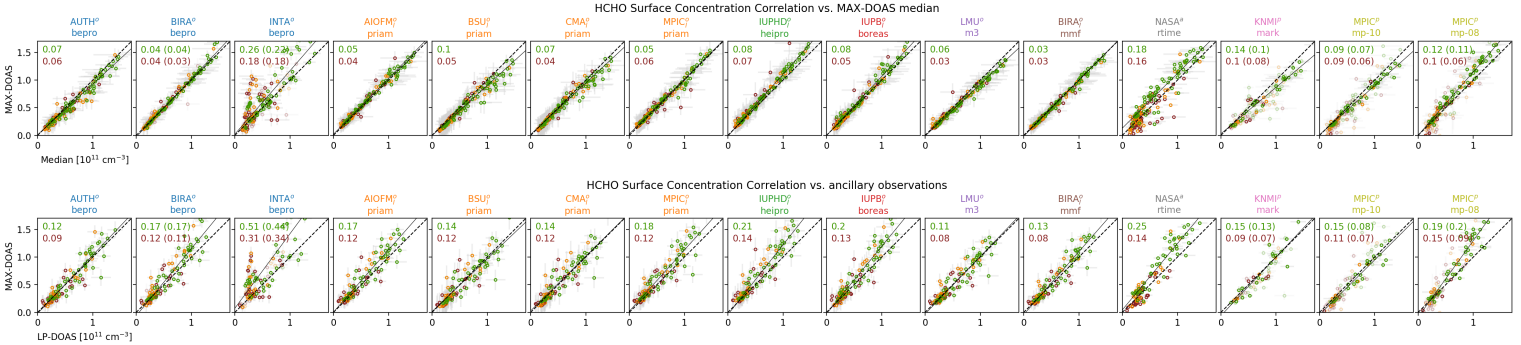


Figure S30. HCHO surface concentration correlation plots. Complementary to main text Sect. 3.6

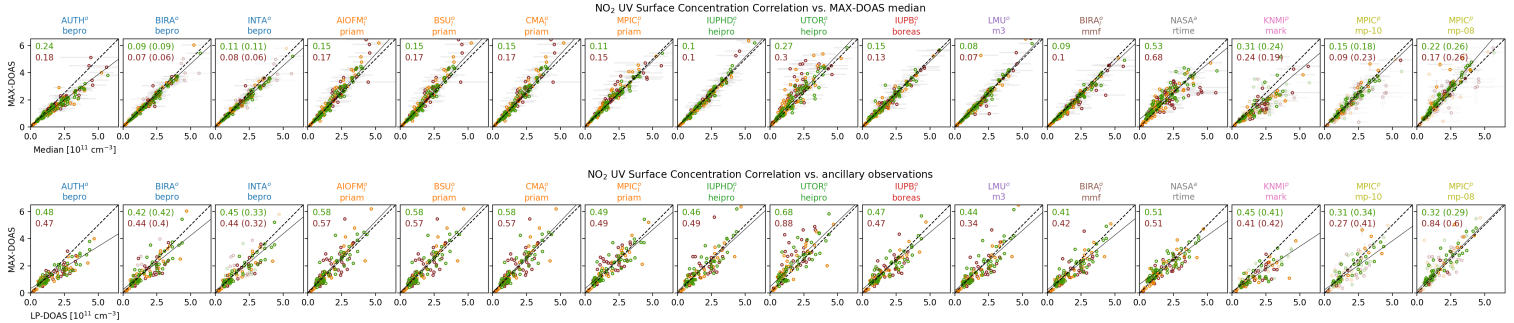


Figure S31. NO_2 UV surface concentration correlation plots. Complementary to main text Sect. 3.6

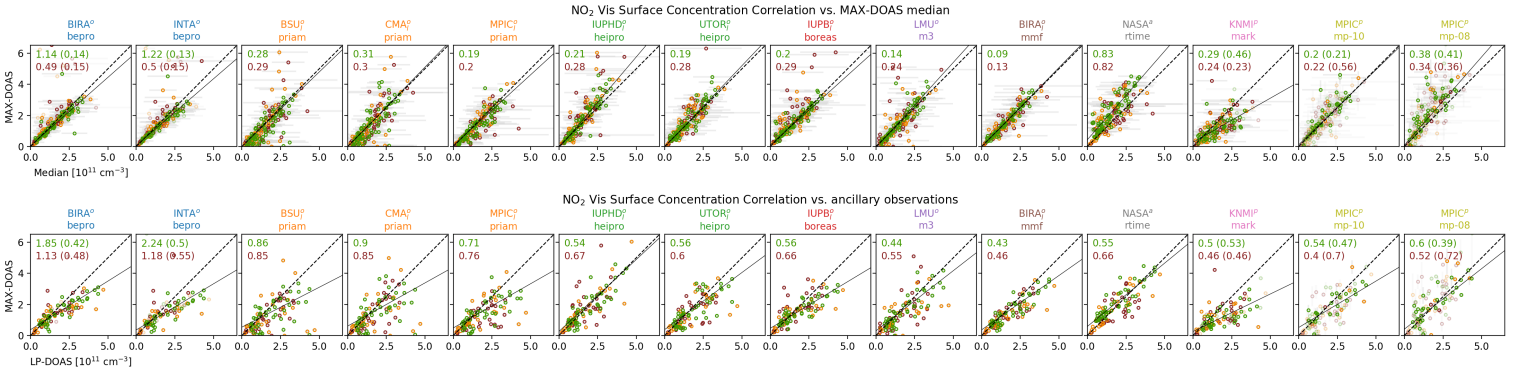


Figure S32. NO_2 Vis surface concentration correlation plots. Complementary to main text Sect. 3.6

S9 Impact of smoothing on surface concentration

For NO_2 the impact of smoothing effects on the surface concentration retrieved by OEM algorithms can be estimated from profiles of the NO_2 lidar and radiosondes. Each profile is smoothed according to Eq. (9) in the main text and the difference in surface concentration between the smoothed and the unsmoothed profile is calculated. ~~Figure~~ Fig. S33 shows histograms of the calculated differences. The standard deviation is about 5×10^9 molec cm^{-3} which is only about 10 % of the total average RMSD between MAX-DOAS and LP-DOAS observations. An estimate of the impact of smoothing on the retrieval results is actually provided by the OEM retrievals themselves as the "smoothing error". The specified smoothing errors are also indicated in Fig. S33 and ~~indeed slightly larger than~~ are similar to the standard deviation observed in in this test, meaning that for the surface layer they are well representative for the real impact of smoothing.

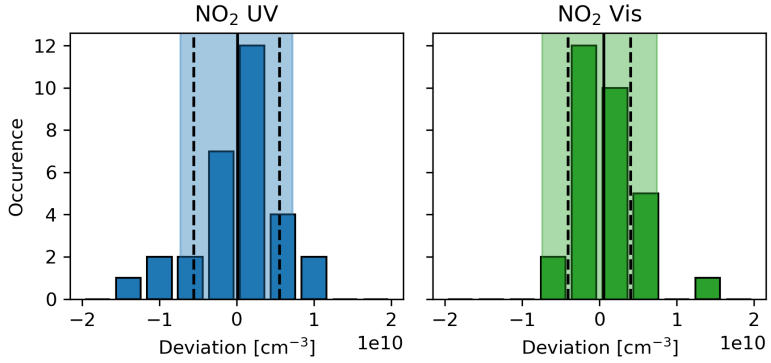


Figure S33. Histograms of the observed deviations in surface concentration between raw and smoothed lidar/ radiosonde NO_2 profiles. Solid and dashed lines indicate mean value and standard deviation, respectively. Coloured areas represent the median smoothing errors as specified by the OEM retrievals, which is in good agreement with the deviations obtained from the supporting NO_2 profiles.

S10 Participant's own dSCD comparison results

This section shows the comparison results for the case where each participant uses dSCDs measured with his own instrument. Evaluation and plots are fully equivalent to Sect. 3 in the main text. With DLR (German Aerospace Center, Oberpfaffenhofen, Germany, marked by blue squares) and USTC (University of Science and Technology of China, Hefei, China, marked by green squares), two other participants were included here, retrieving profiles with bePRO and HEIPRO, respectively. Gaps in the data are mostly related to instrument malfunction during the campaign. Further, not all instruments covered the spectral range to detect all desired species and the corresponding participants therefore do not appear in the respective plots.

S10.1 Information content

Information on the averaging kernels and DOFS. This section is equivalent to Sect. S10.1 in the main text and Sect. S8.1 in this supplement, respectively.

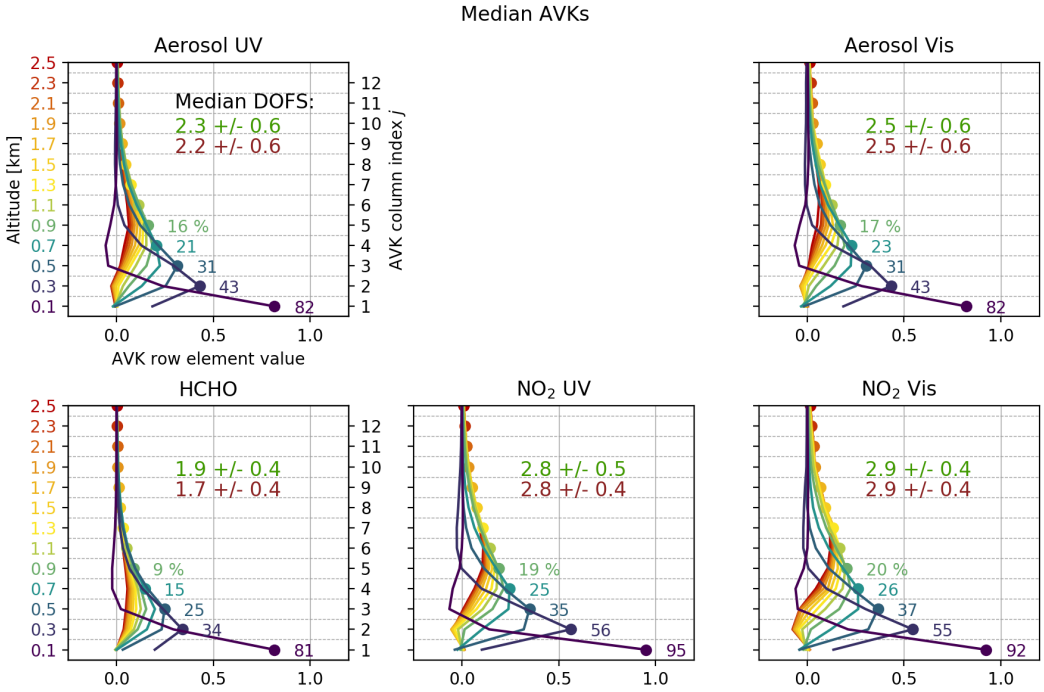


Figure S34. Average AVKs for the retrieved species (median over participants, mean over time). Each altitude and corresponding AVK line are associated with a colour, which is defined by the colour of the corresponding altitude-axis label. The dots mark the AVK diagonal elements. The number next to the dots show the exact value in percent, which corresponds to the amount of retrieved information on the respective layer. In the upper right of each panel, the DOFS (median among institutes, average over time) are given for clear-sky (green) and cloudy conditions (red).~~The vertical bars indicate the vertical resolution (the "spread" as defined by equation 3.23 in **?**) for the five lowest layers.~~

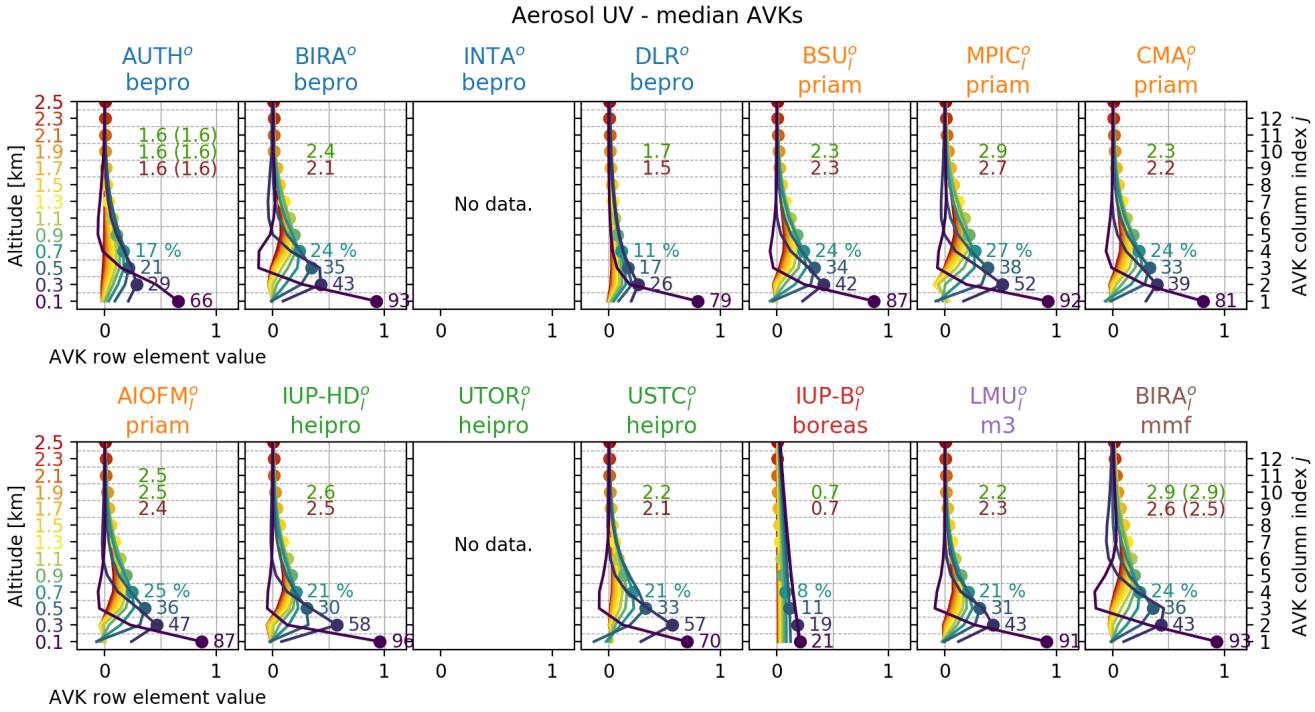


Figure S35. Mean averaging kernels for Aerosol UV for each participant retrieving from their own dSCDs. Coloured values at AVK peaks show the amount of retrieved information on the respective layer in percent. "DOFS" numbers are given for clear-sky (green) and cloudy (red) conditions. Values in brackets are DOFS including flagging.

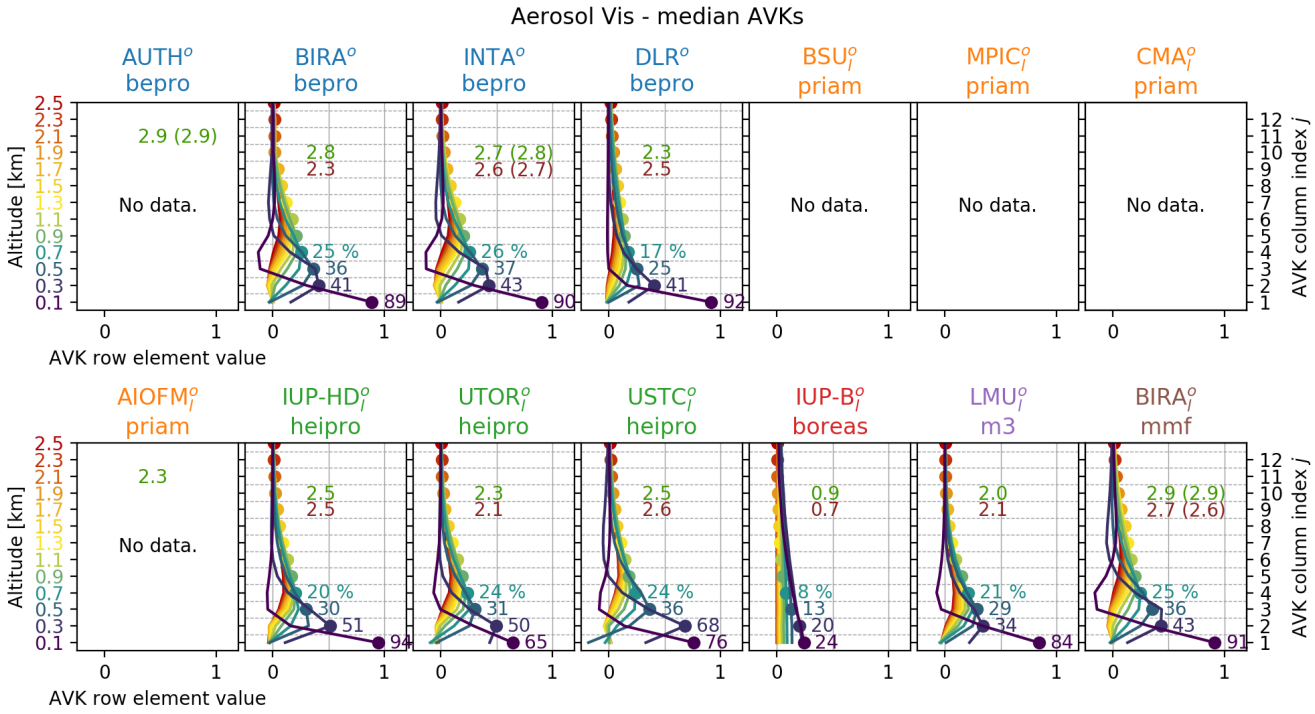


Figure S36. Mean averaging kernels for Aerosol Vis for each participant. Description of Fig. S35 applies.

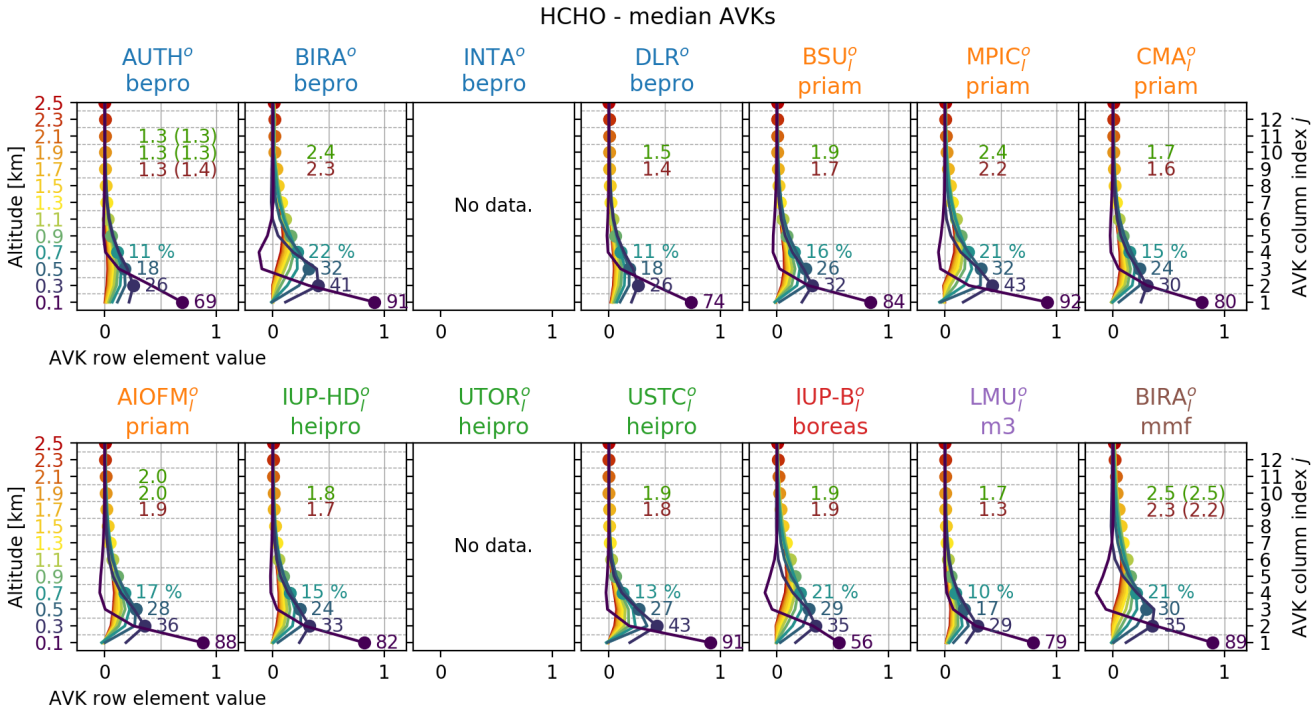


Figure S37. Mean averaging kernels for HCHO for each participant. Description of Fig. S35 applies.

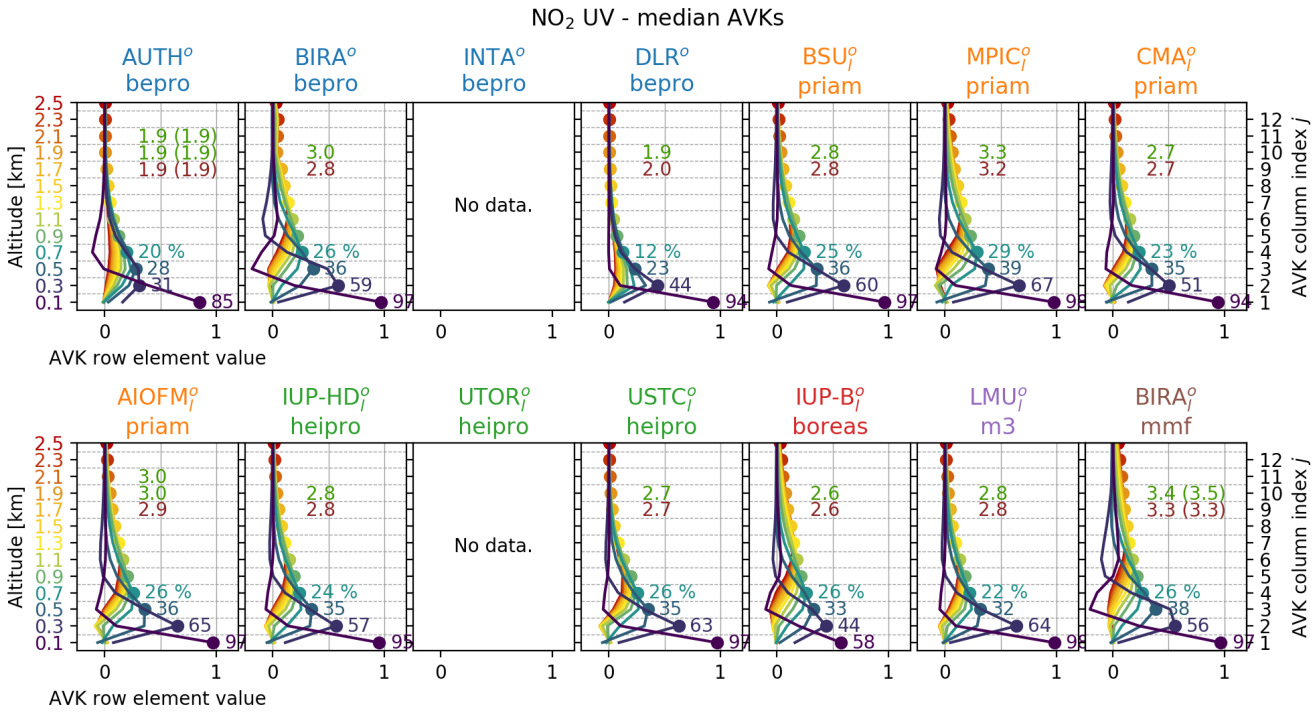


Figure S38. Mean averaging kernels for NO_2 UV for each participant. Description of Fig. S35 applies.

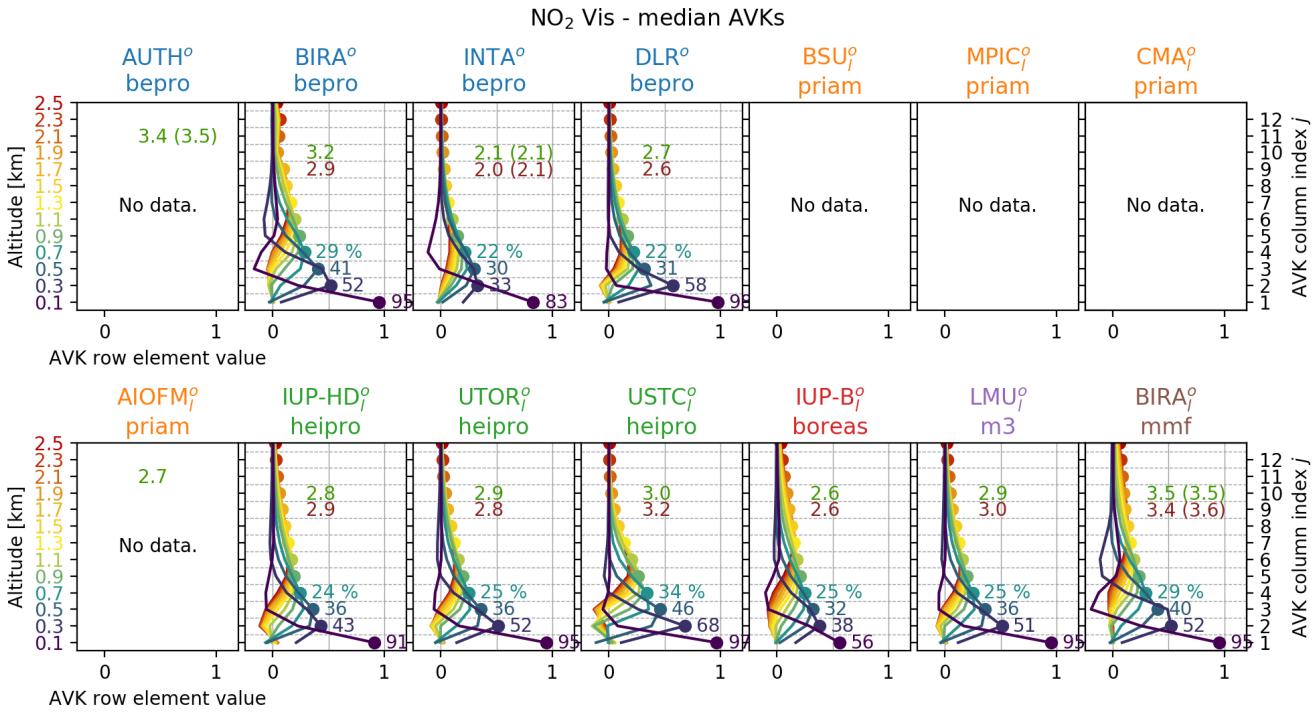


Figure S39. Mean averaging kernels for NO_2 Vis for each participant. Description of Fig. S35 applies.

S10.2 Overview plots

This section is equivalent to Sect. 3.2 in the main text.

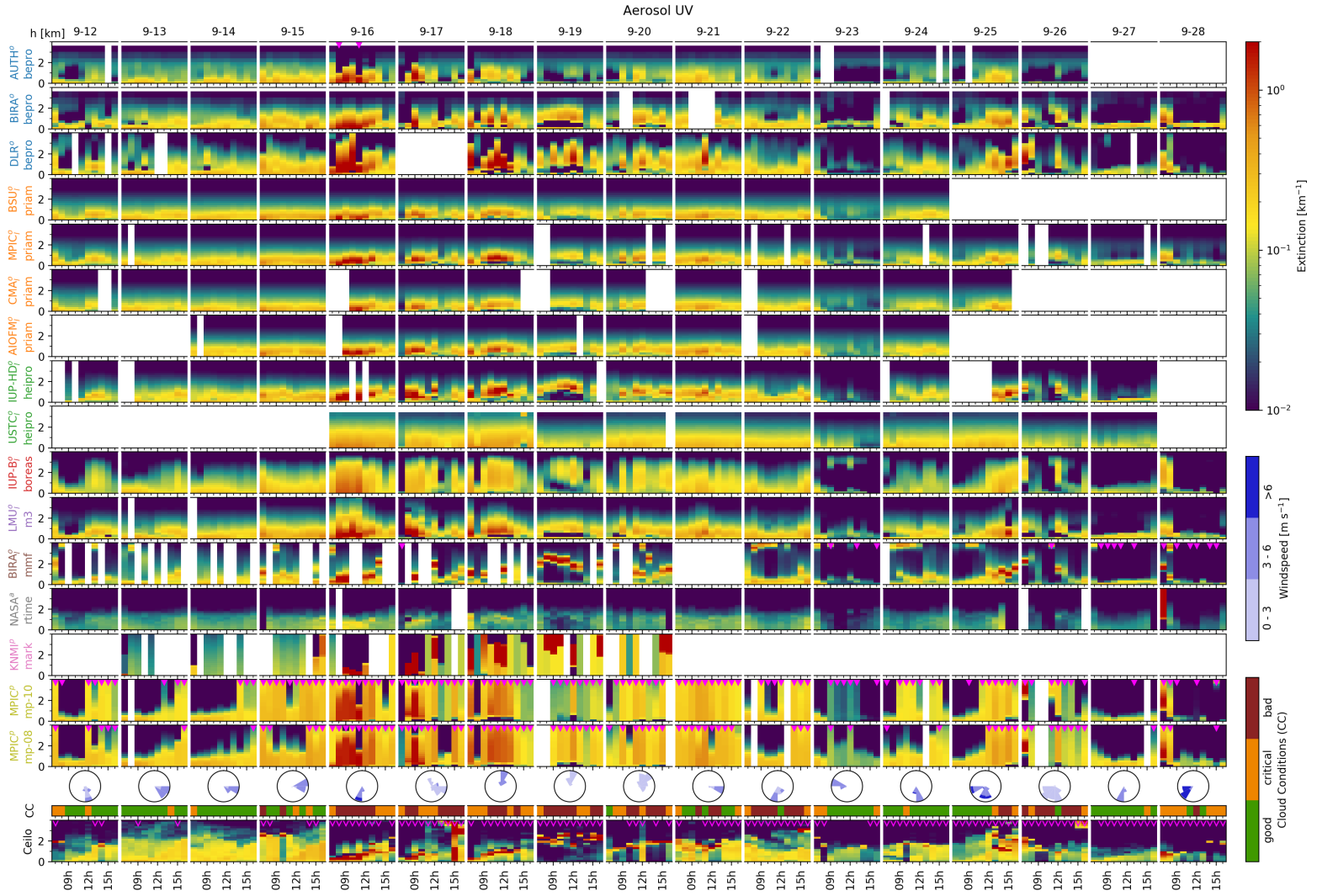


Figure S40. Aerosol UV extinction profiles retrieved from the participant's own dSCDs. The lowest row shows AOT scaled ceilometer backscatter profiles, calculated as described in Sect. S4.1. Backscatter profiles, which were scaled from MAX-DOAS AOTs (and which are therefore not fully independent) are marked by ~~red~~ pink triangles. Maximum extinction values reach $20\,\mathrm{km}^{-1}$, exceeding the colour scale.

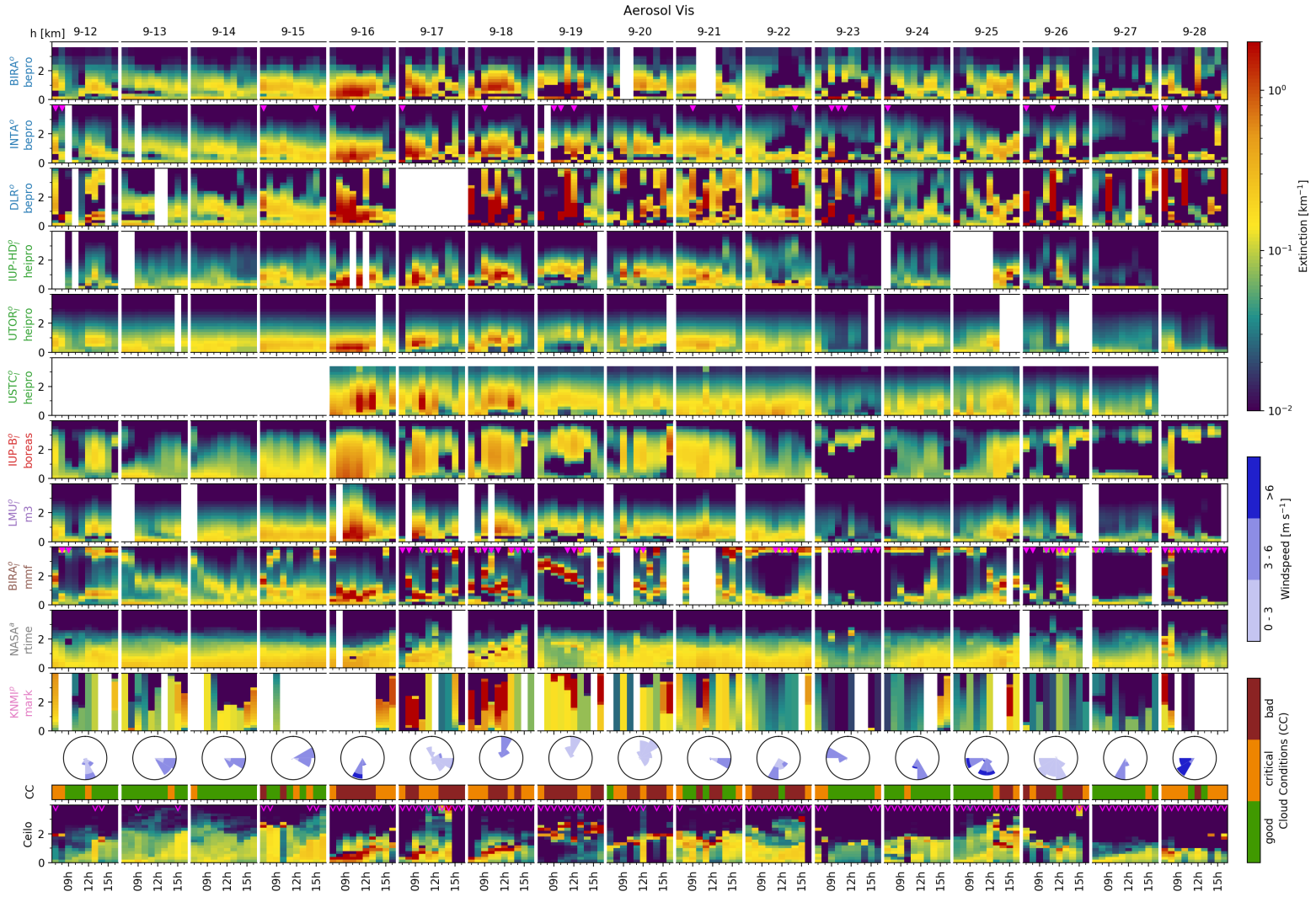


Figure S41. Aerosol Vis extinction profiles retrieved from the participant's own dSCDs. The lowest row shows AOT scaled ceilometer backscatter profiles, calculated as described in Sect. S4.1. Backscatter profiles, which were scaled from MAX-DOAS AOTs (and which are therefore not fully independent) are marked by ~~red~~ pink triangles. Maximum extinction values reach $20\,\mathrm{km}^{-1}$, exceeding the colour scale.

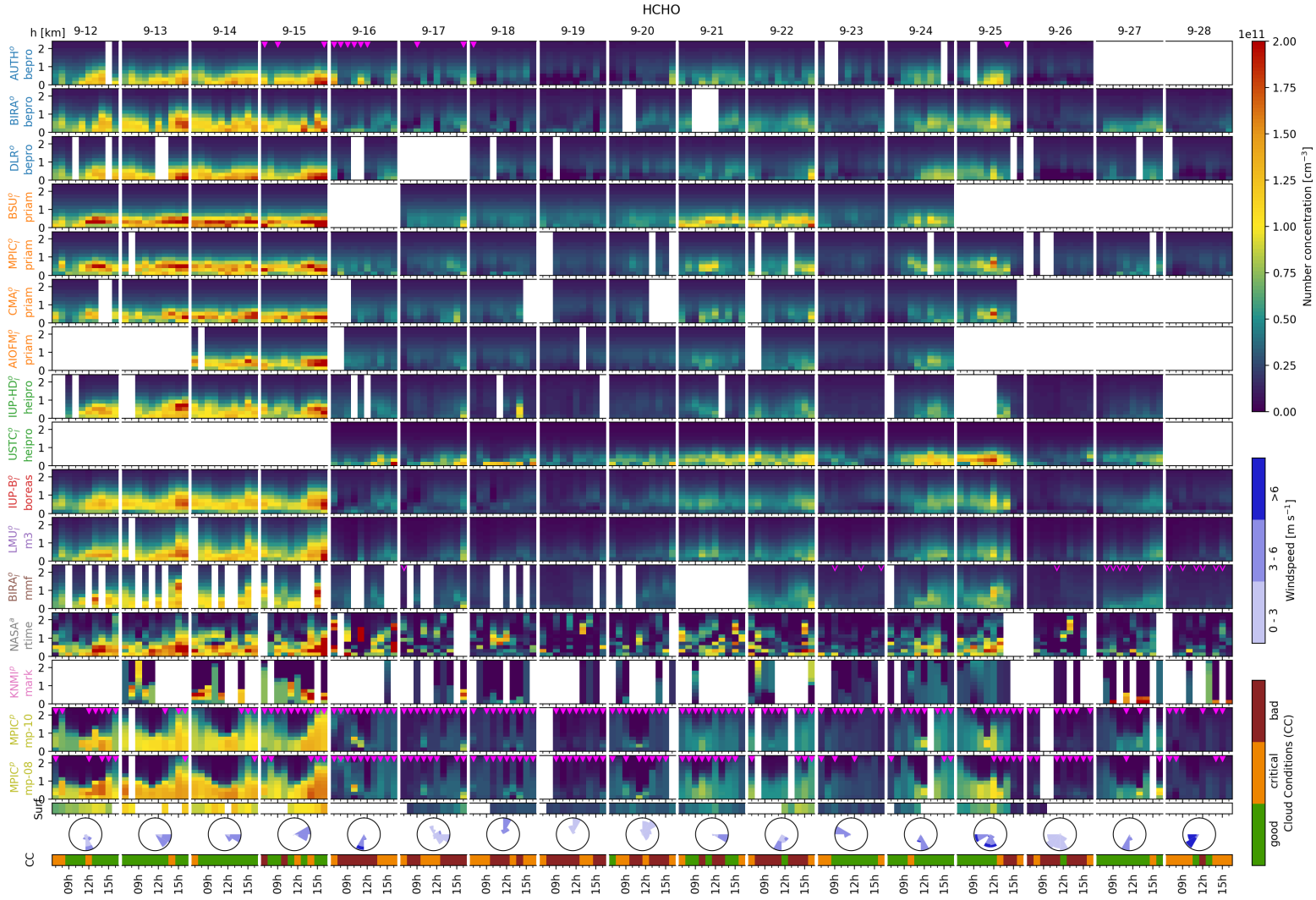


Figure S42. HCHO concentration profiles retrieved from the participant's own dSCDs. The "Surf"-row shows LP-DOAS surface concentrations.

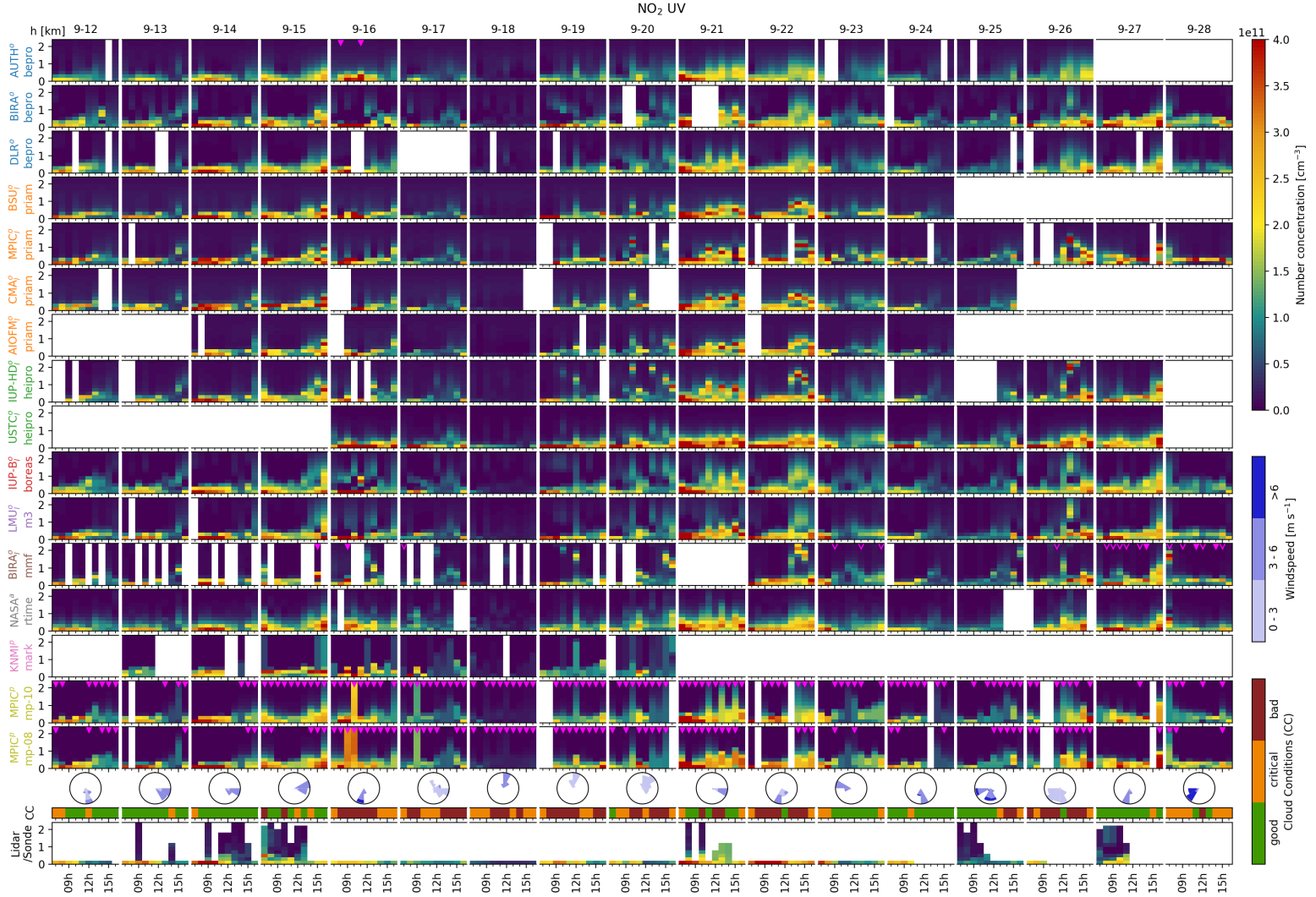


Figure S43. NO_2 UV concentration profiles retrieved from the participant's own dSCDs. The lowest row shows a combined dataset of NO_2 lidar, radiosonde, LP-DOAS and tower in-situ data. Redundant surface concentration measurements were averaged.

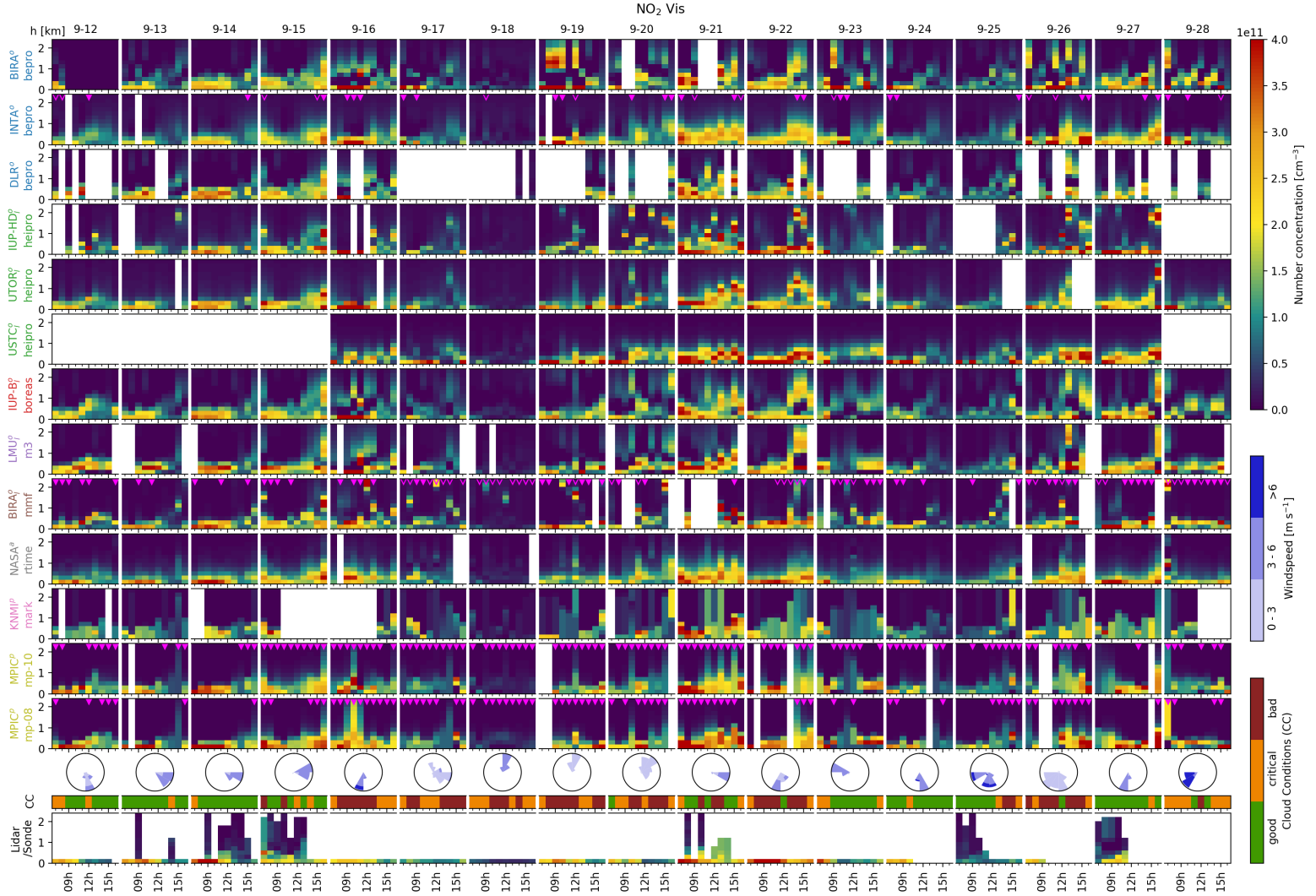


Figure S44. NO_2 Vis concentration profiles retrieved from the participant's own dSCDs. The lowest row shows a combined dataset of NO_2 lidar, radiosonde, LP-DOAS and tower in-situ data. Redundant surface concentration measurements were averaged.

S10.3 Modelled and measured dSCDs

This section is equivalent to Sect. S10.3 in the main text.

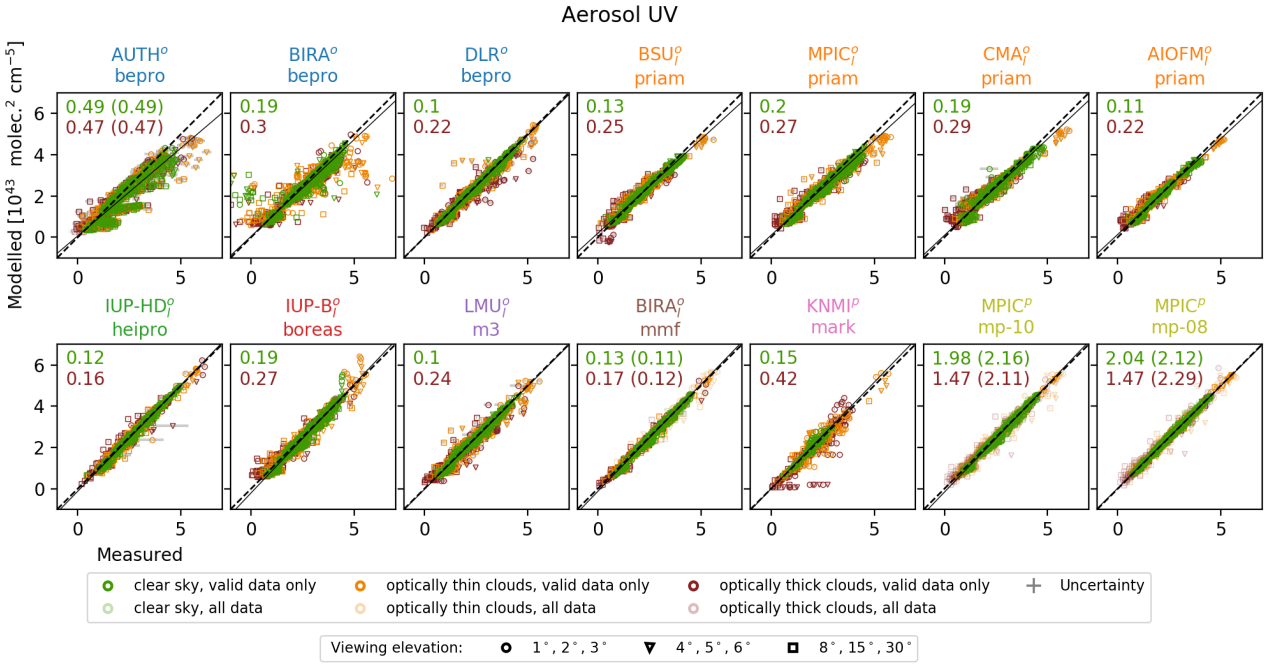


Figure S45. O_4 UV dSCD correlation when profiles are retrieved from the participant's own dSCDs. Marker colours and marker shapes indicate the cloud conditions and viewing elevation angles, respectively. Numbers represent the measurement error weighted RMSD between measured and modelled dSCDs for clear sky (green) and cloudy (red) conditions. Values in brackets were calculated only considering valid data.

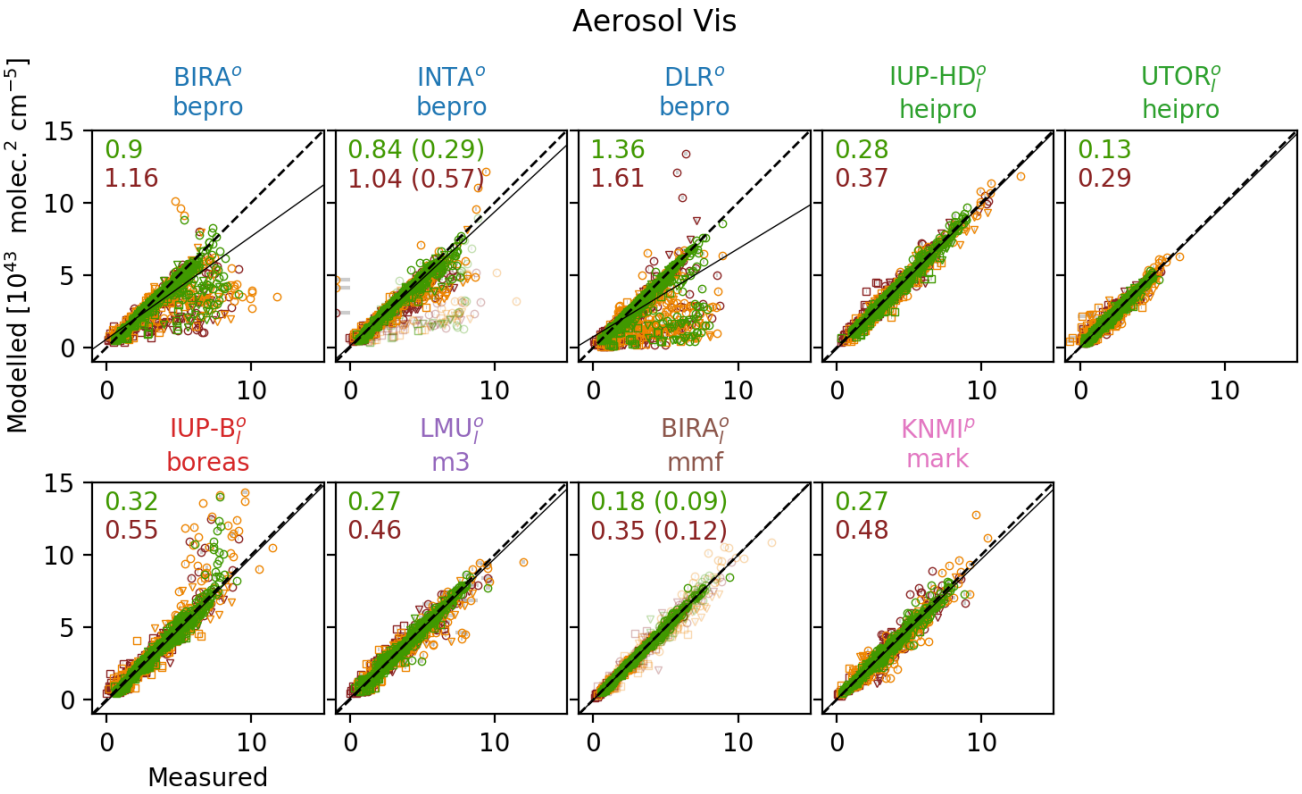


Figure S46. O_4 Vis dSCD correlation. Legends of Fig. S45 apply.

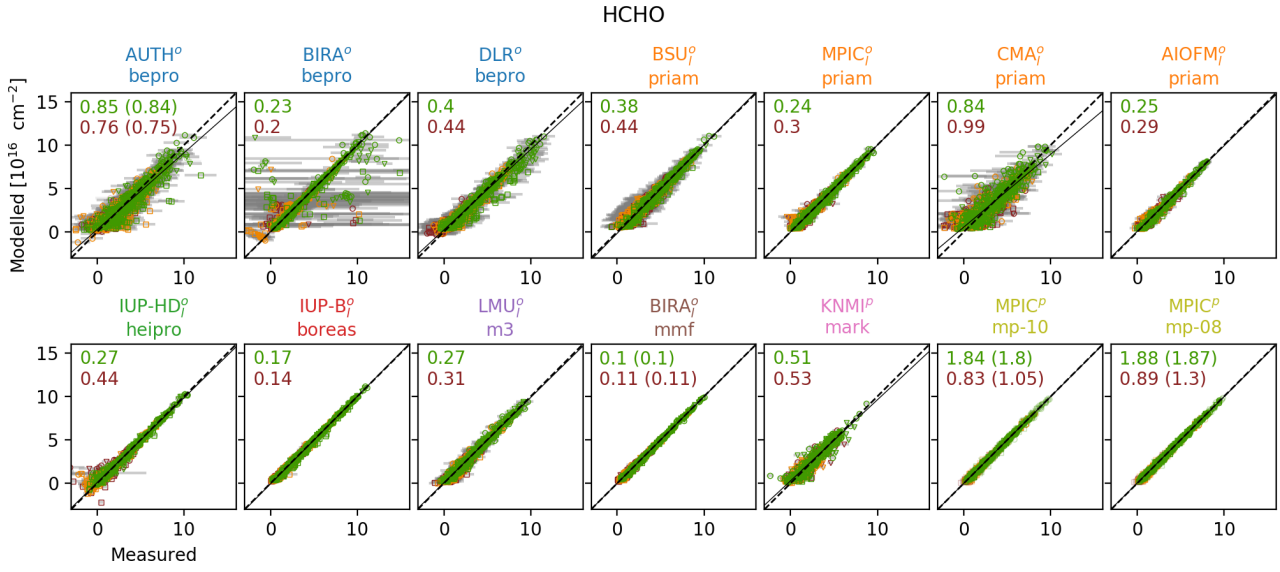


Figure S47. HCHO dSCD correlation. Legends of Fig. S45 apply.

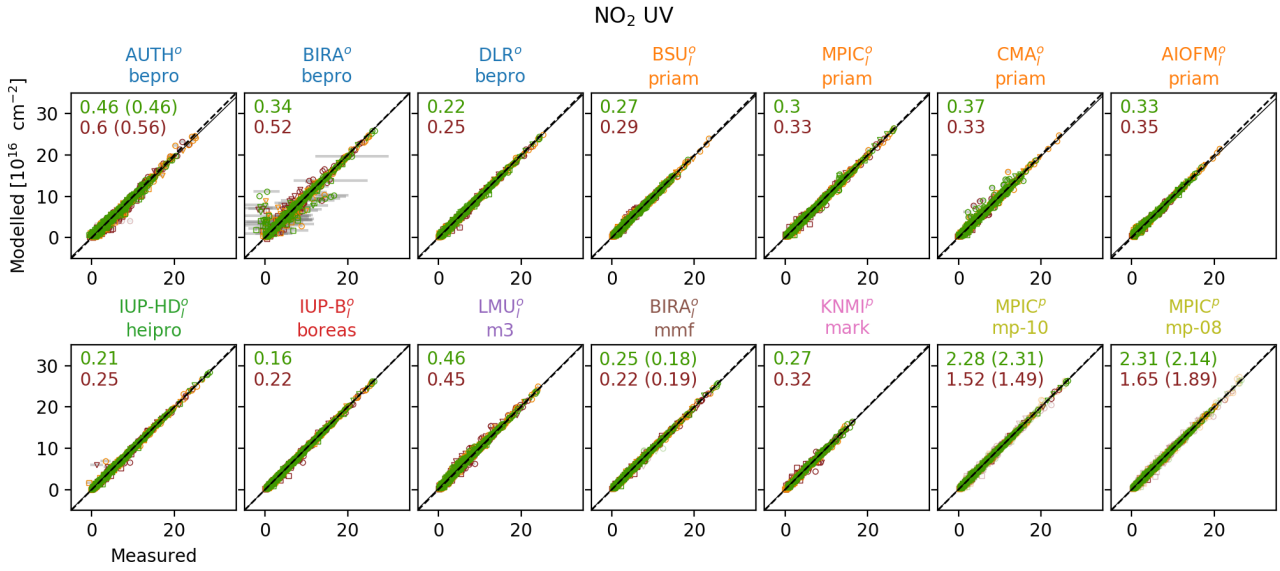


Figure S48. NO_2 UV dSCD correlation. Legends of Fig. S45 apply.

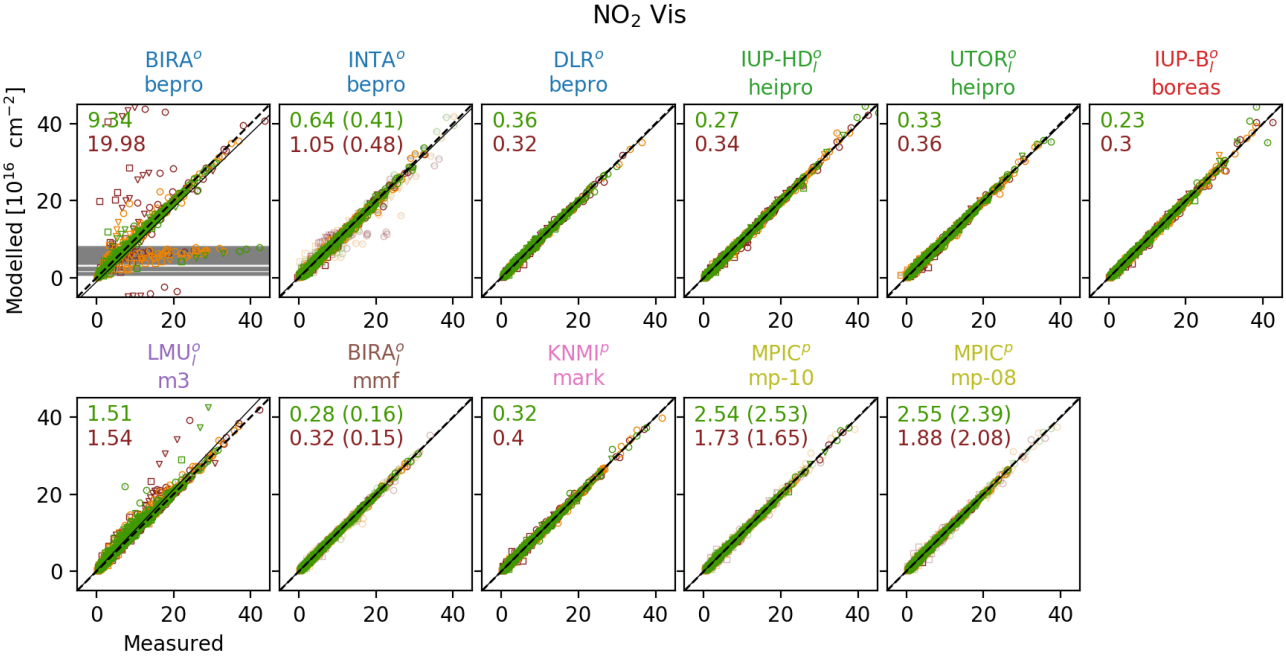


Figure S49. NO_2 Vis dSCD correlation. Legends of Fig. S45 apply.

S10.4 Aerosol optical thickness (AOT)

This section is equivalent to Sect. S10.4 in the main text.

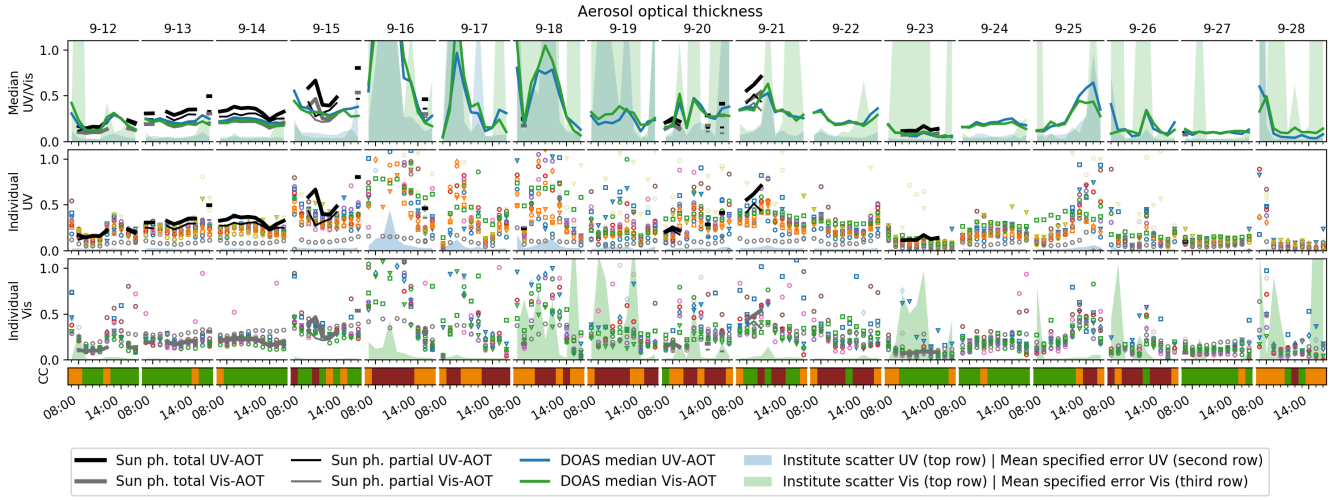


Figure S50. MAX-DOAS AOTs retrieved from the participant's own dSCDs in comparison to sun photometer data. Symbol and symbol colours are chosen according to Table 2 in the main text. Transparent symbols indicate data flagged as invalid. Top row: MAX-DOAS median results vs. the available supporting observations, according to the legend below the plot. ~~Hatched~~ The "institute scatter" areas ~~(starting at the top of the plot)~~ show the scattering ~~(standard deviation)~~ among the participants ~~(only~~ 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.

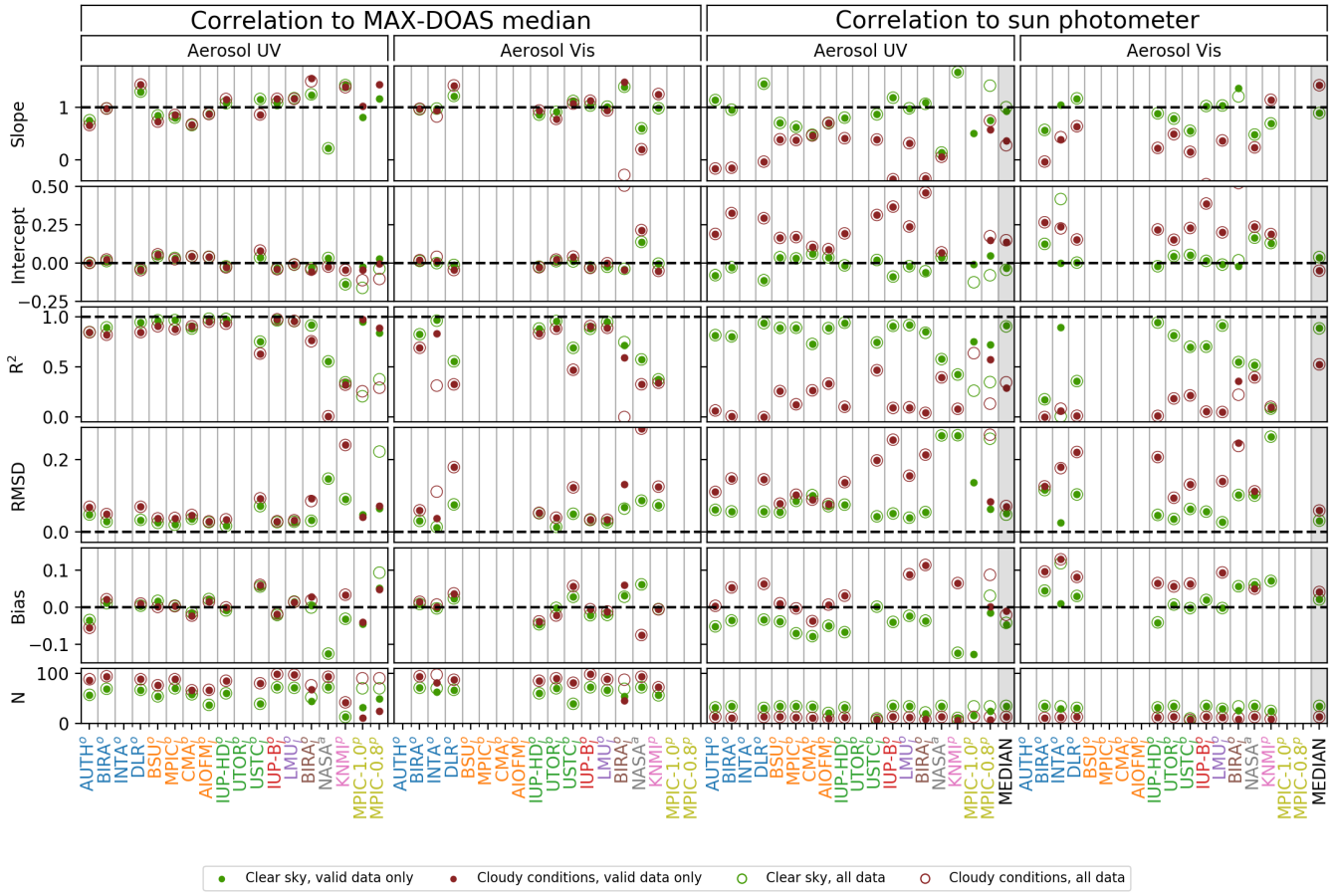


Figure S51. Correlation statistics for AOTs ~~retrieved from the participant's own dSCDs~~. The two left columns give an impression on the agreement among the institutes, as they show the correlation of the individual participant's retrieved AOT (ordinate of the underlying correlation plot) against the 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. ~~Transparent symbols~~ Hollow circles represent values for all submitted data, ~~opaque symbols~~ the dots only consider data points flagged as valid. ~~The pies indicate, which fraction of~~ N is the ~~total~~ number of profiles ~~(170)~~ 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).

S10.5 Trace gas vertical column densities (VCDs)

This section is equivalent to Sect. S10.5 in the main text.

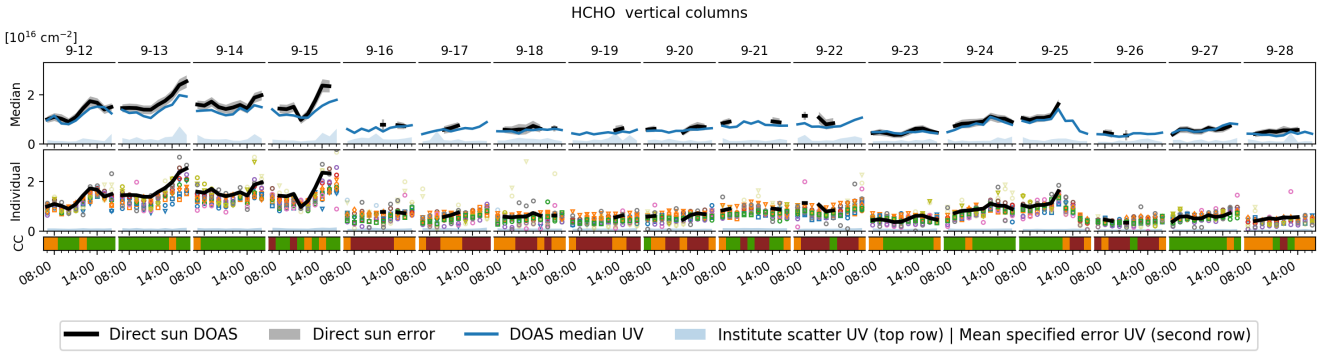


Figure S52. Comparison of MAX-DOAS HCHO VCDs retrieved from the participant's own dSCDs vs. direct-sun DOAS, NO_2 lidar and radiosonde. Descriptions of Fig. S50 apply.

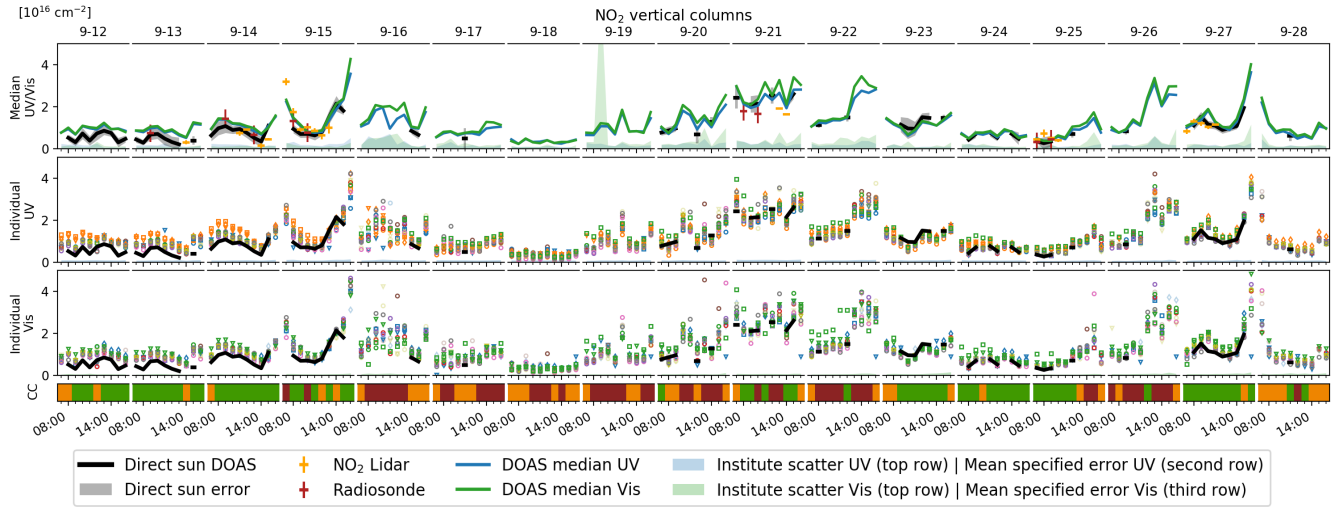


Figure S53. Comparison of MAX-DOAS NO_2 VCDs retrieved from the participant's own dSCDs vs. direct-sun DOAS. Descriptions of Fig. S50 apply.

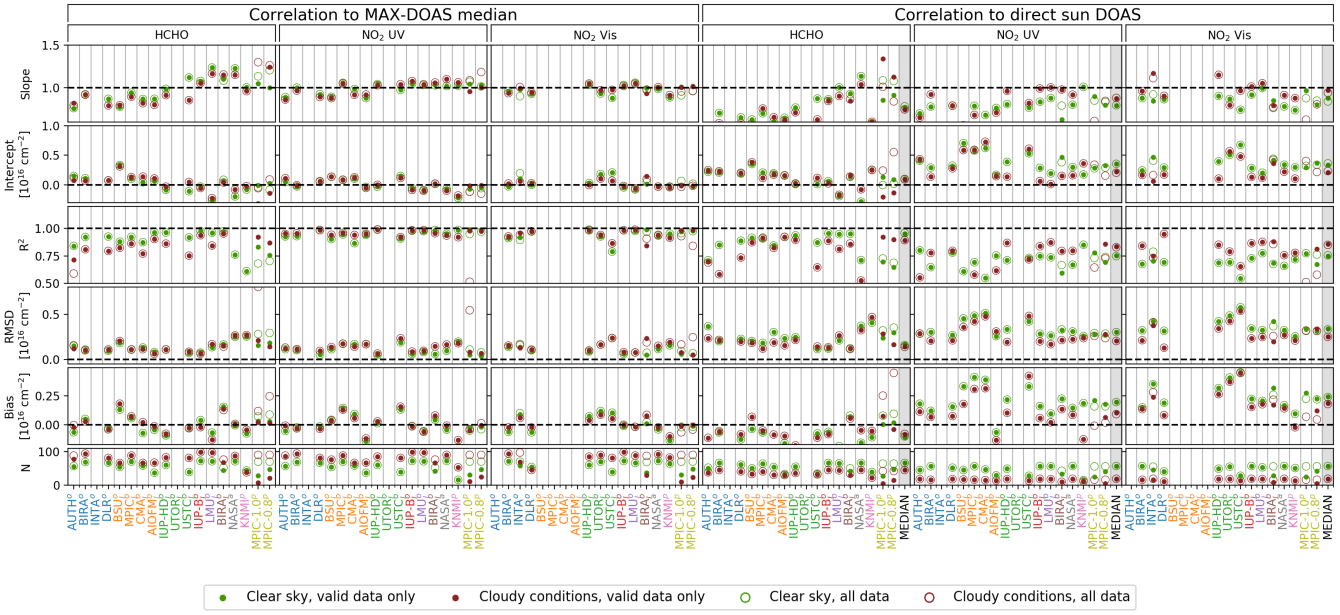


Figure S54. Correlation statistics of trace gas VCDs retrieved from the participant's own dSCDs. Legends and description of Fig. S51 apply.

S10.6 Trace gas surface concentrations

This section is equivalent to Sect. S10.6 in the main text.

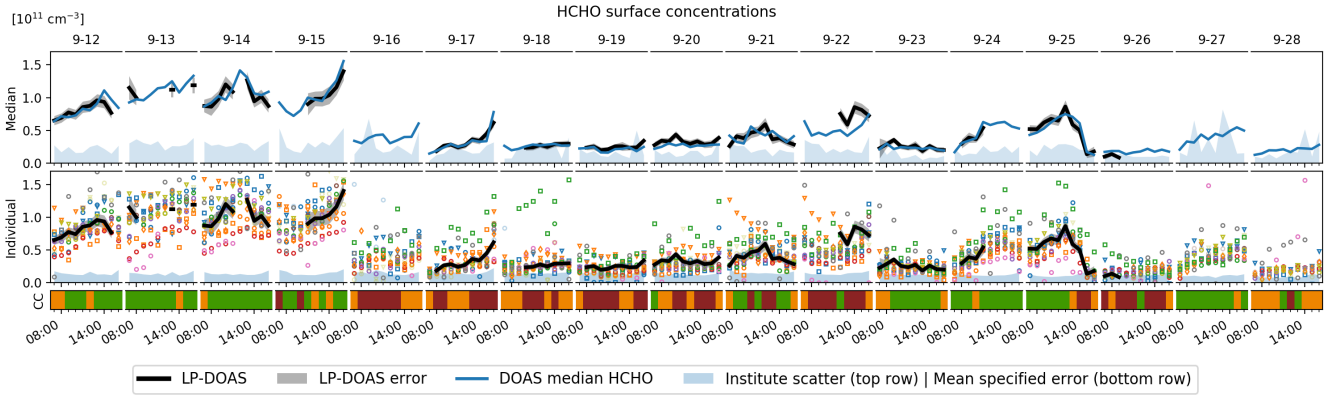


Figure S55. Comparison of MAX-DOAS HCHO surface concentrations retrieved from the participant's own dSCDs. Description of Fig. S50 applies.

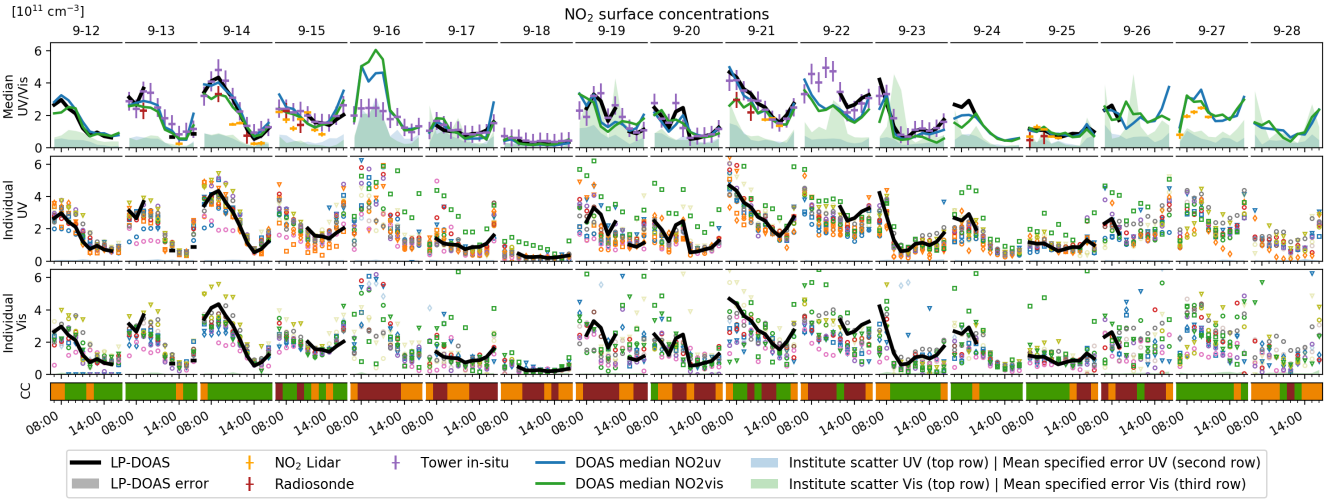


Figure S56. Comparison of MAX-DOAS NO_2 surface concentrations retrieved from the participant's own dSCDs. Description of Fig. S50 applies.

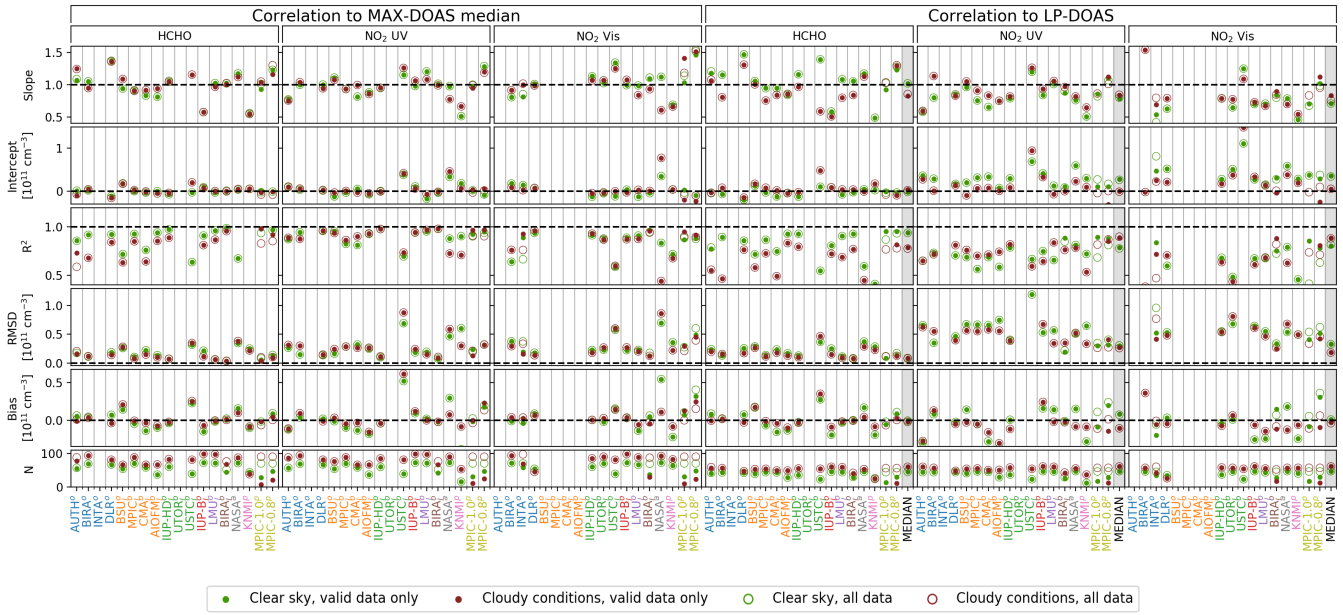


Figure S57. Correlation statistics of trace gas surface concentrations retrieved from the participant's own dSCDs. Basic description of Fig. S51 applies.

S10.7 ~~NO_2 UV-Vis comparison~~

~~This section is equivalent to Sect. ?? in the main text.~~

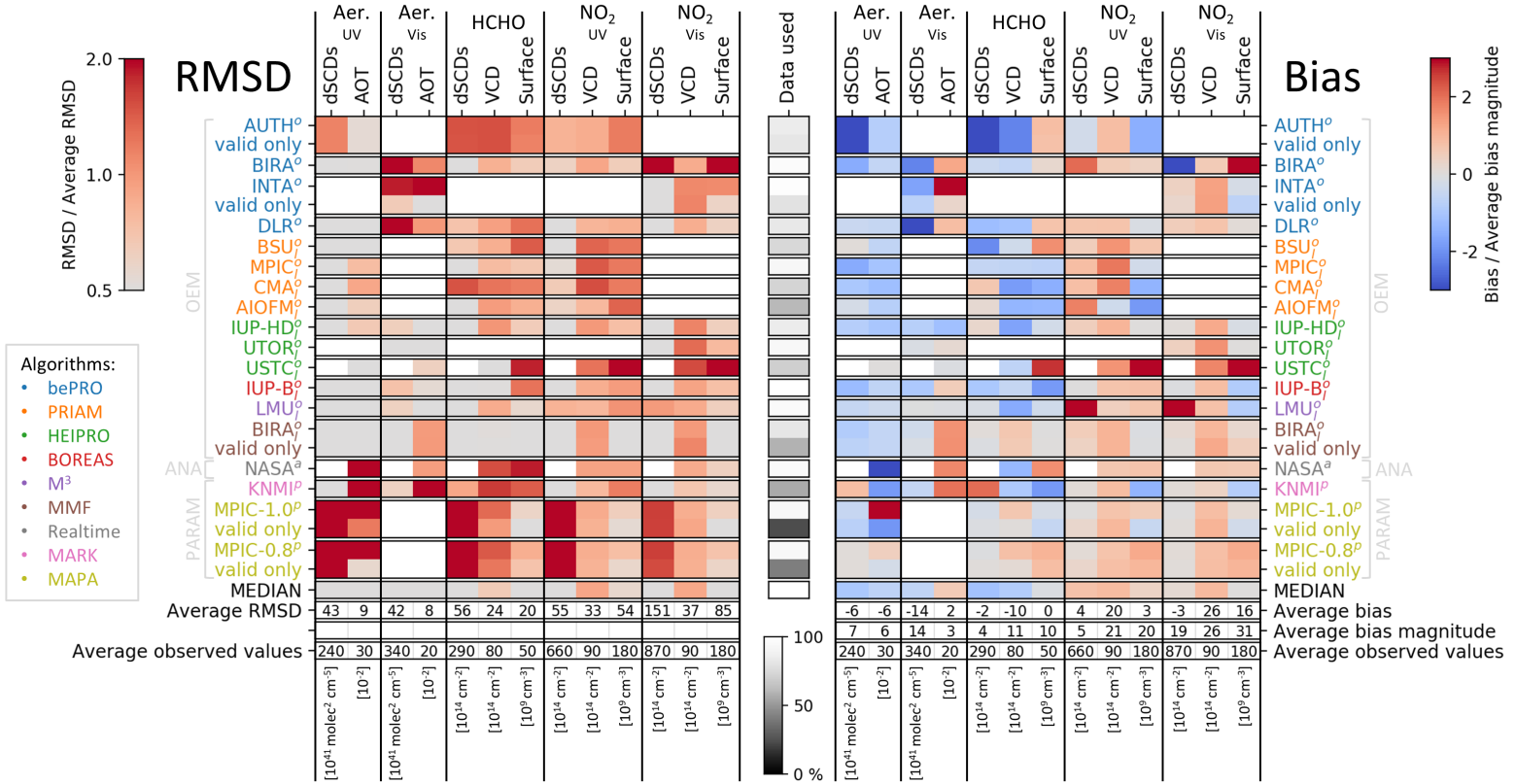


Figure S58. ~~Correlation~~ Summary of ~~MAX-DOAS VCDs retrieved from~~ the ~~participant's own dSCDs~~ comparisons in Sect. S10 for clear-sky conditions. Left panel shows RMSD, right panel shows Bias. Average values of RMSD (Bias) define the ~~UV and~~ colour scale of each column of the ~~Vis spectral range~~left (right) panel as indicated by the color bars on the top left (top right) of the figure. ~~Marker colours~~ Values of AOT, VCD and ~~marker shapes~~ surface concentration are given with respect to the corresponding supporting observations. White spaces indicate no data. Average observed values (bottom row) are rounded campaign averages of the ~~cloud conditions~~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").

~~Correlation of MAX-DOAS surface concentrations retrieved from the participant's own dSCDs in the UV and the Vis spectral range. Legends of Fig. **??** apply.~~

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.

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.

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.

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.

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.

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

Mayer, B. and Kylling, A.: Technical note: The libRadtran software package for radiative transfer calculations - description and examples of use, Atmospheric Chemistry and Physics, 5, 1855–1877, https://doi.org/10.5194/acp-5-1855-2005, https://www.atmos-chem-phys.net/5/1855/2005/, 2005.

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.

Peters, E., Ostendorf, M., Bösch, T., Seyler, A., Schönhardt, A., Schreier, S. F., Henzing, J. S., Wittrock, F., Richter, A., Vrekoussis, M., and Burrows, J. P.: Full-azimuthal imaging-DOAS observations of NO_2 and O_4 during CINDI-2, Atmospheric Measurement Techniques Discussions, 2019, 1–30, https://doi.org/10.5194/amt-2019-33, 2019.

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

Wagner, T. and Beirle, S.: Estimation of the horizontal sensitivity range from MAX-DOAS O_4 observations, Tech. rep., QA4ECV, http://www.qa4ecv.eu/sites/default/files, 2016.

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., 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.