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.

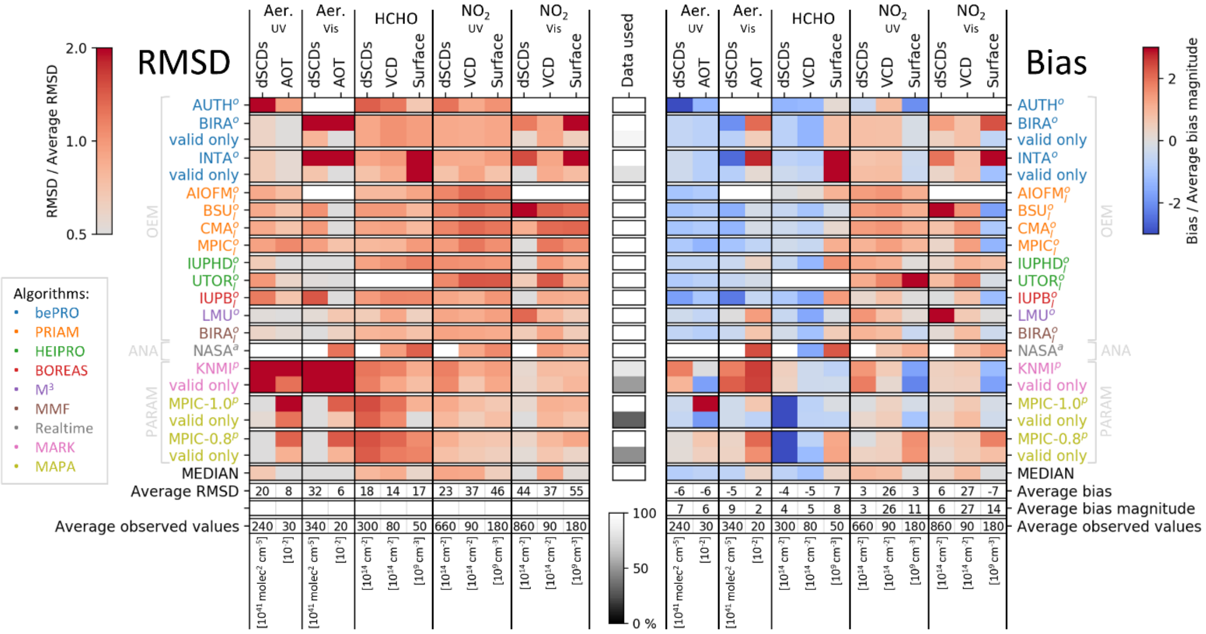
Response:
The conclusions were revised as stated above and according to the specific comments below (see also response to reviewer #1).

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:

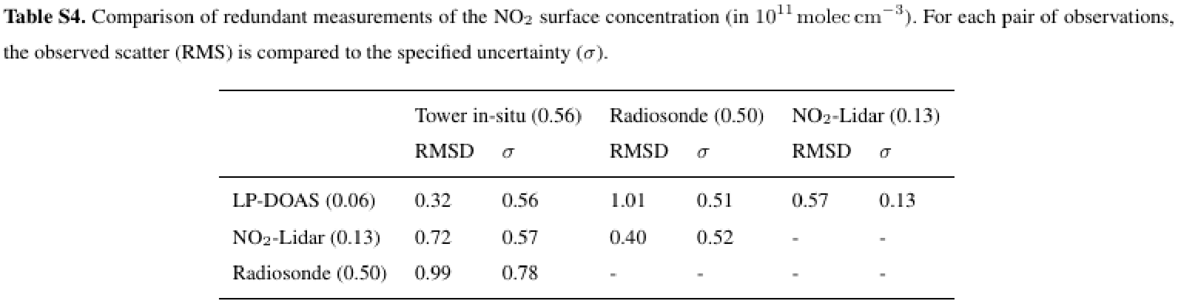


Table S4. Comparison of redundant measurements of the NO_2 surface concentration (in 10^{11} molec cm^{-3}). For each pair of observations, the observed scatter (RMS) is compared to the specified uncertainty (σ).

	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	-	-	-	-

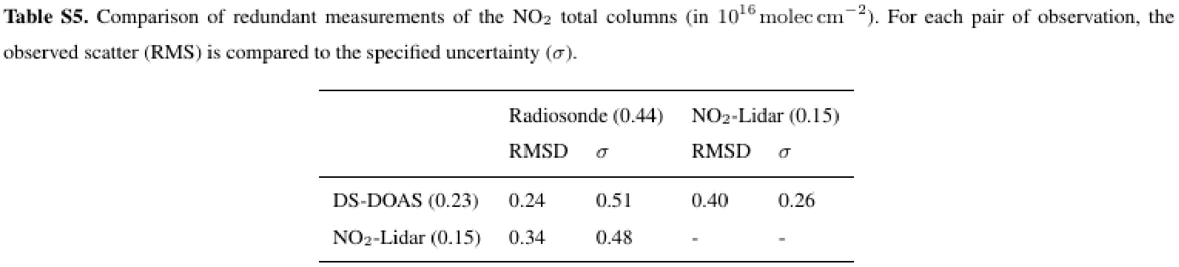


Table S5. Comparison of redundant measurements of the NO_2 total columns (in 10^{16} molec cm^{-2}). For each pair of observation, the observed scatter (RMS) is compared to the specified uncertainty (σ).

	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?

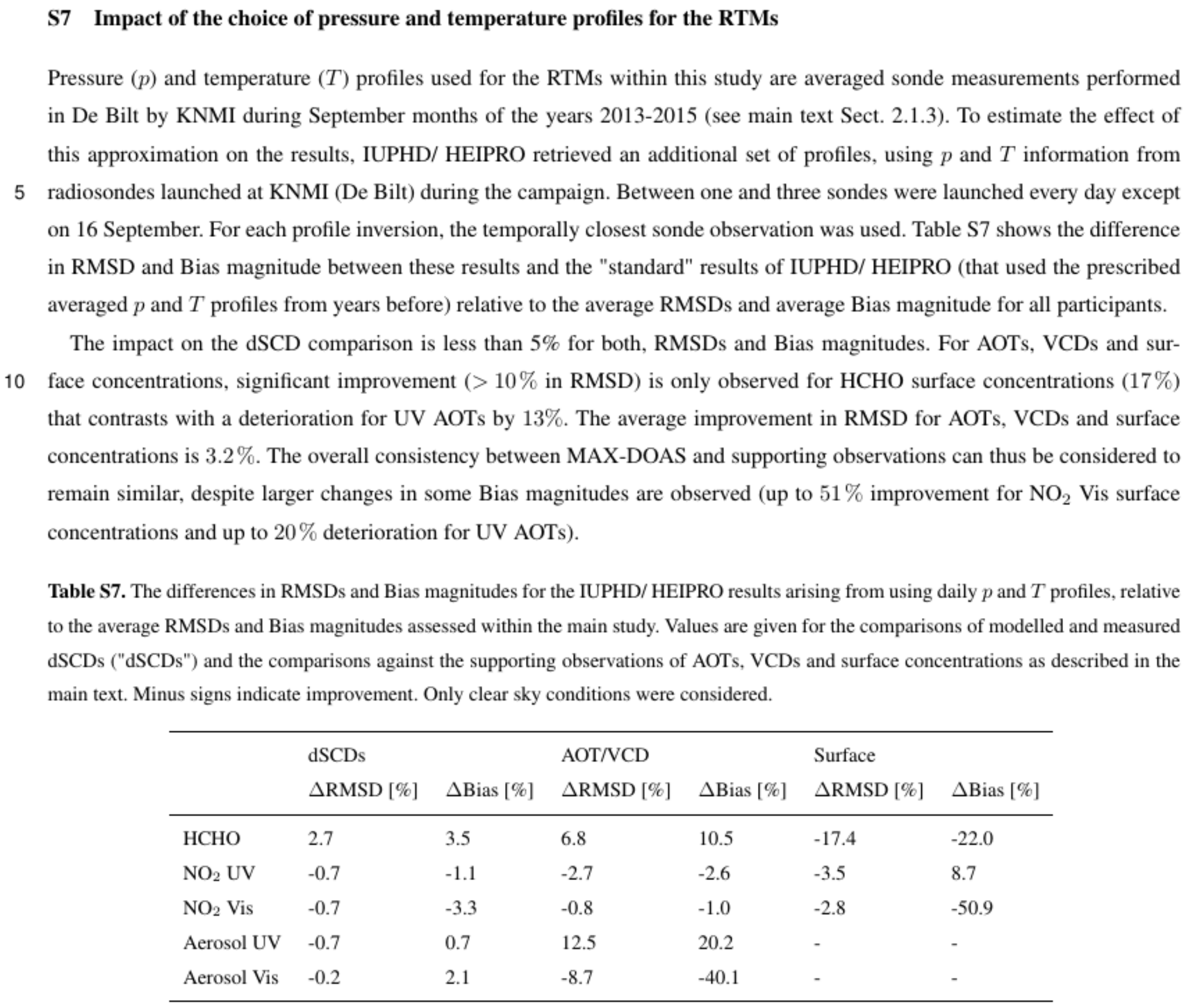
We agree with the reviewer. To make this point clearer we added a sentence very similar to the reviewer’s comment: “*Note, that this approach gives a large weight to the uppermost measurements, as they are representative for the majority of the relevant layer.*”

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: $$\sigma_{arms,p} = 1/N_P \sum_{p} \sigma_{rms,p}$$

To:

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

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:

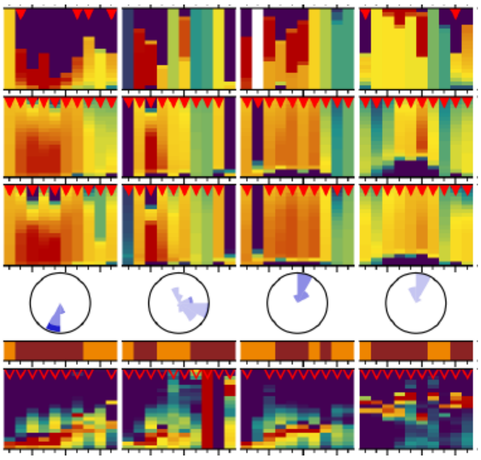
“Department of Physics, University of Toronto, Toronto, Canada” → “Department of Physics, University of Toronto, Canada”

“NASA-Goddard, Greenbelt, Maryland” → “NASA-Goddard, Greenbelt, United States”

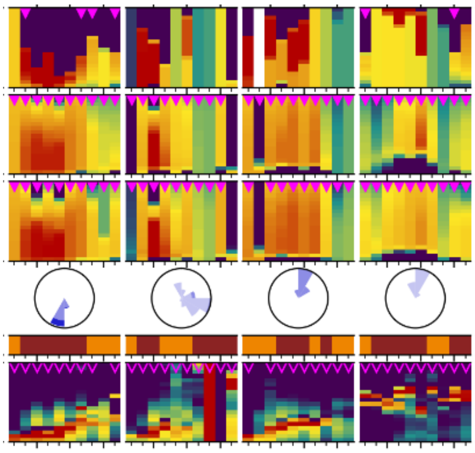
Figs. 3-7. The red triangles are not readily seen against the color scale.

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:

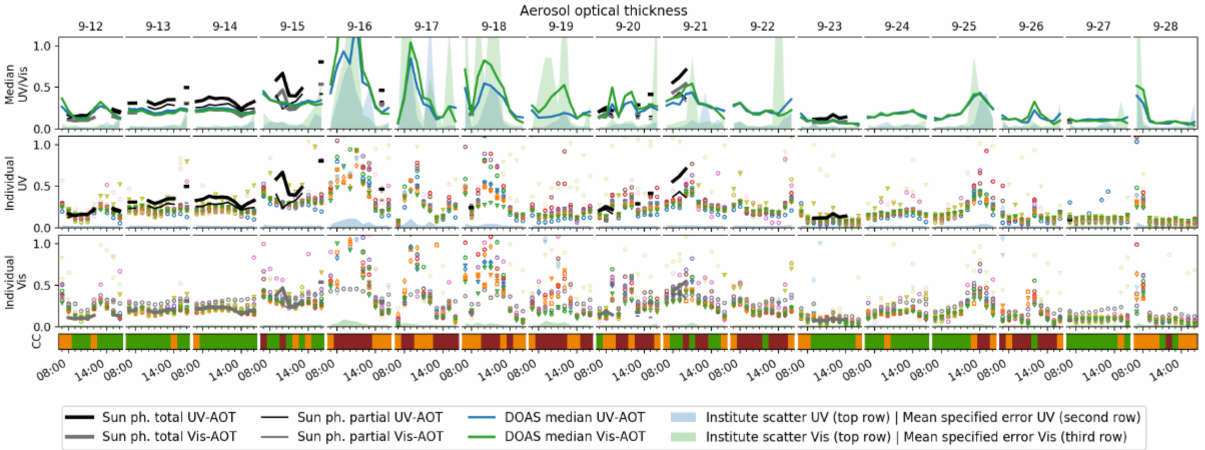
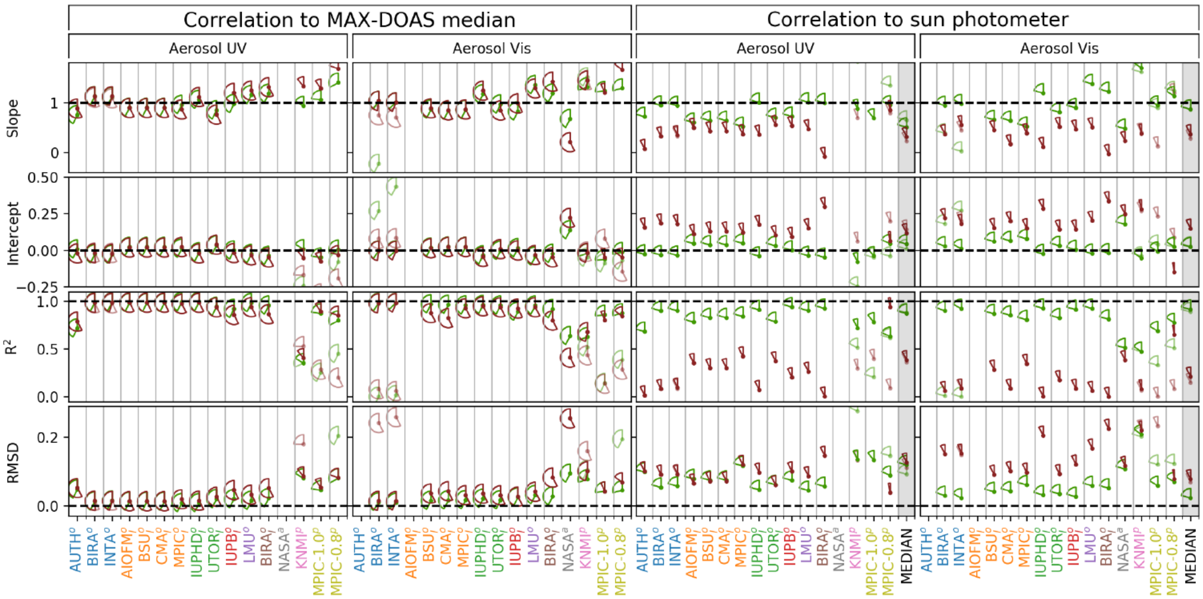


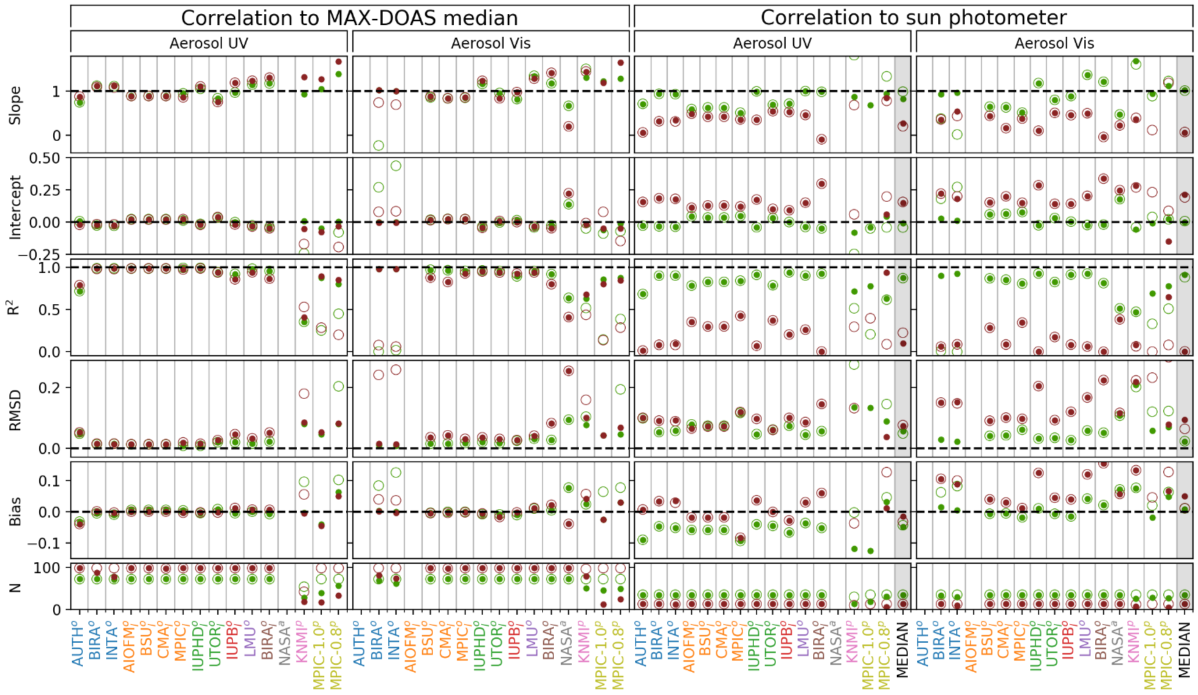
Fig. 14 and other Figs following same format. While I can appreciate what the authors are trying to communicate with the pie chart symbols, the clear and cloudy data are drawn from the same total and the symbols repeat within a given column. This should be simplified in some way.

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).*"

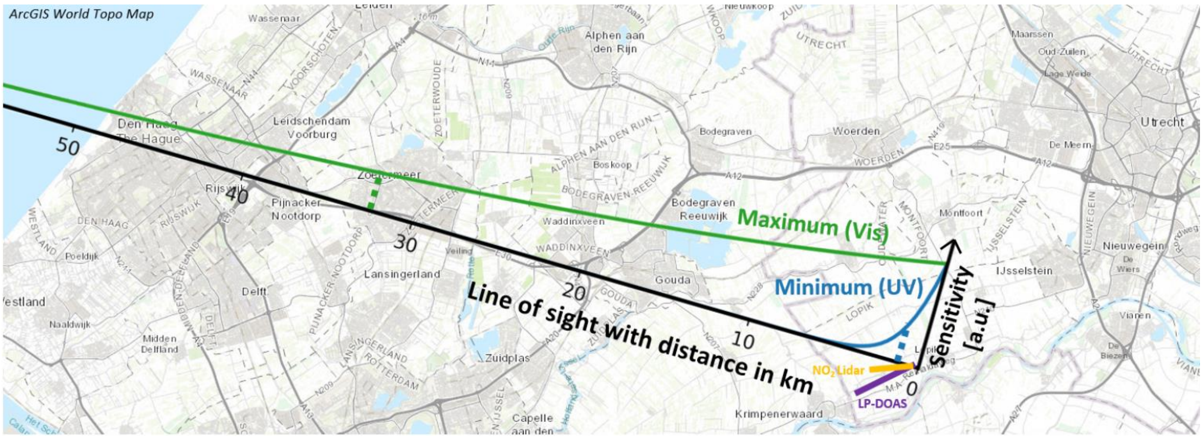
Fig S11 The color scheme makes this figure very difficult to read.

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:

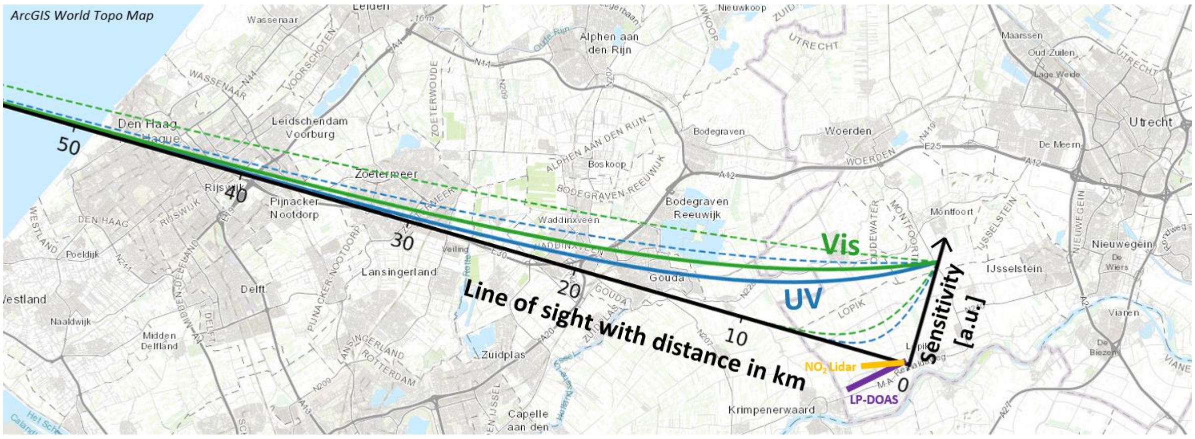


Fig. S34 If I understand this figure correctly virtually all data are within two standard deviations, is this not as expected. P33 L6-7 seems to imply something unexpected.

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.*”