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:

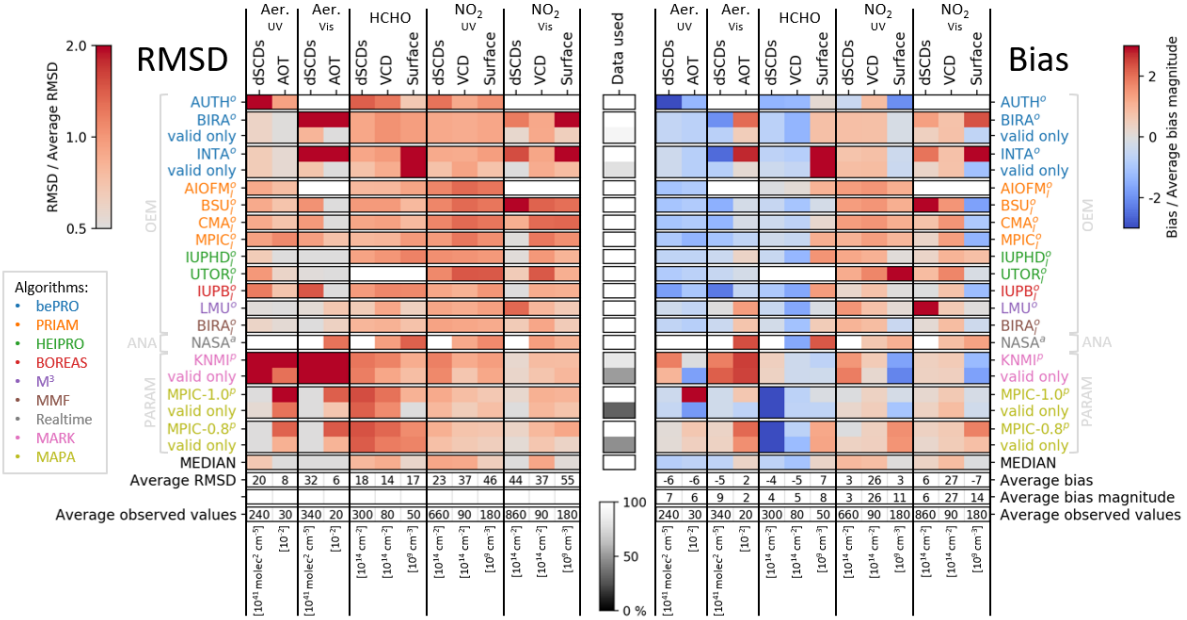
1. A summarising figure similar to Fig. 23 was created also for the own dSCD comparison and is contained in the supplementary material
2. In the corresponding Section (3.8) in the main text, "Bias" values (description below) were added in Table 5. Further, we now directly compare the impact of the use 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.

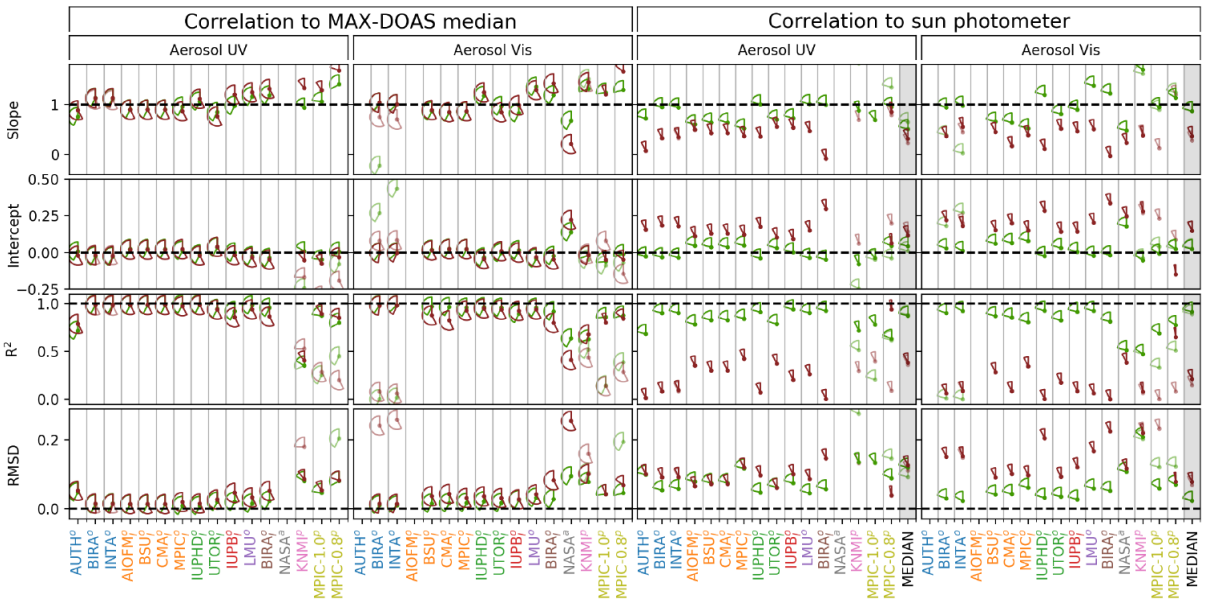
P29, L9. "underprivileged" sounds weird, please change it.

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*"

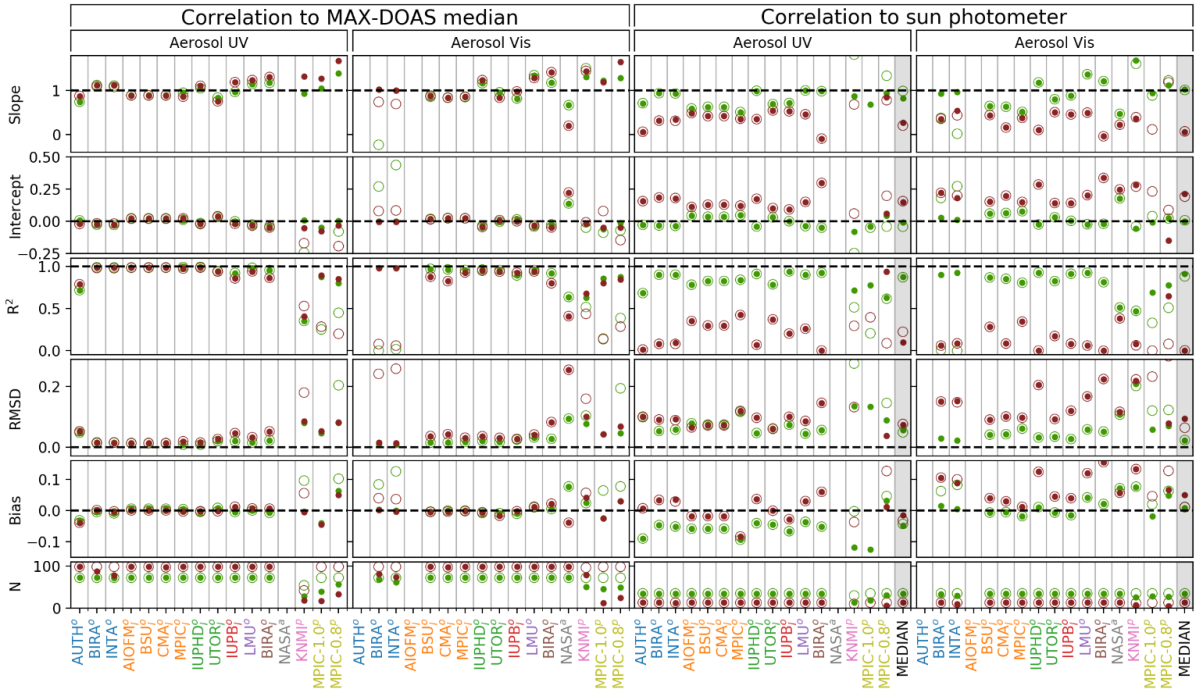
P30. figure 14. Please add bias in % (negative/positive) as mentioned above. Additionally, light vs opaque are not distinguishable, maybe using other colors might help? Furthermore, symbols on the two right column plots are not shown in the legend, maybe you meant to use the same symbols?

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.