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

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.
    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.
    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.
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.
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.

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.
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.

Specific Comments

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

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

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.

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.

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.

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.

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

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.

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

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

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.

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.

P6 L14 DSCDs are reported for five data products, however the UV and Vis retrievals of O_4 and NO_2 retrieve the same chemical species

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.

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.

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.

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

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.

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?

P12 L8 temperature and pressure should be spelled out here.

P12 L9 Wagner et al., (2019) find effects of up to 7% on the modeled O_4 profile when using a standard atmosphere. This could be a significant contributor or the retrieved RMSD, can the authors comment?

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

P12 Eq7 $1/N_p$ here should be in parentheses for clarity

P14 L24 replace "not given" with "inaccurate"

P15 L1-2 "A_{ij} describes the sensitivity of the **measured** concentration in the i^{th} layer to **small** changes in the real concentration in the j^{th} layer.,"

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.

P15 L16-18 The increase in information content reflects the 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.

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.

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

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

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

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

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

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?

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?

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

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.

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.

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?

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.

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.

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

P31 14 The first sentence should be reworded, the VCDs are compared to different standards or “assessed”, but the NO_2 VCDs are not compared to the HCHO VCDs

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

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.

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

P34 L3 The language here should be more precise. The surface concentration does reflect the ability of MAX-DOAS retrieval to solate 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.

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?

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.

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

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.

P4 L12-14 Clear-sky O_4 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.

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

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.

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.