Point-by-point answers to Anonymous Reviewer #2

Note: Reviewers comments are printed in black, author's replies in blue (and italic).

General Evaluation:
The paper by Peters et al. investigates differences in NO2 DOAS retrieval due to differences in different DOAS retrieval codes. A same set of MAX-DOAS observations and DOAS retrieval settings were provided to various groups using different DOAS retrieval codes to retrieve NO2 differential slant column densities (dSCD). Resulting NO2 dSCD and RMS were then compared and a range of possible sources for their differences were investigated. In the end, the authors provide a list of 5 recommendations aimed at improving the NO2 DOAS retrievals. These recommendations are fairly straight forward and largely constitute best practices that are already followed to improve DOAS retrievals. Overall, the paper is very nicely written but way too long. I suggest the authors make changes to shorten the length of the paper.

Recommendation:
The paper fits the scope of AMT journal as it documents differences is DOAS retrieval codes and hence the paper is acceptable for publication in AMT with modifications to (i) make the differences between the codes more explicit and (ii) shorten the manuscript.
We like to thank the reviewer for the generally encouraging comments. We also share the opinion that the manuscript should be shortened. We did this following the specific reviewer's suggestions as described below.

Specific Comments:
There are two parts to the paper and they could very easily be two separate papers. The first part focuses on the intercomparison of NO2 from different retrieval codes. While in the surface it seems like a nice and interesting idea to compare different DOAS retrieval codes, it becomes very clear after section 3.1 (Part 1) that differences due to retrieval codes are very small compared to differences due to instrumental design (see Roscoe et al, 2010). The second part investigates the potential sources of these differences. However it appears like a sensitivity study to determine best features to have in a retrieval code. The authors performed the sensitivity tests first, and then compared the results from the sensitivity tests to the results from various groups. Based on the results the authors went back to the groups to verify their findings. There is disconnect between section 3 and 4 as there is lack of basis for the tests being performed. I suggest the authors include an overview table (using the survey data) which highlights the differences between different retrieval codes to connect section 3 and 4. This table could replace most of the text describing the retrieval codes.
We changed the motivation for Part2 to make a stronger connection and better explanation for Part 2 following from Part1 (we did this already in response to comments of Anonymous Reviewer #1).
It is not true that we performed the sensitivity tests first. The intercomparison between groups was performed first, then a survey to identify possible reasons for observed differences, and then the sensitivity studies in order to evaluate the effect of each of the potential reasons. We state this more clearly in the revised manuscript.
It is true that differences are smaller than observed in Roscoe et al. 2010 and we agree that instrumental differences are supposed to be larger. However, we cannot strictly conclude that remaining differences come from instrumental design alone. For example, during CINDI (Roscoe et al, 2010), the instruments did not point at exactly the same time into the same direction (and also in real measurements all kinds of misalignments are potentially present). In this aspect, the recent CINDI-2 campaign is interesting because instruments followed a strict measurement protocol, i.e. coinciding measurements are assured. However, there are no CINDI-2 results published yet.

In order to meet the reviewer's comments, we changed and shortened Sect. 2.4. In particular, instead of the list of participating groups we now introduce each retrieval code and emphasize important differences (we also tried to put all content into a table, as suggested by the reviewer, but it turned out that the table format is not suited). We did not include an additional table for the survey as this would again increase the length of the manuscript while providing little new insight. We therefore directly mention the result of the survey leading to the 5 systematic differences which are then subject of the performed sensitivity tests.

The amount of details for different retrieval codes are not comparable. Some codes are described in details while others (e.g. IUPHD) barely include a sentence.
This whole section has been changed in the revised manuscript (see above). Explanations of retrieval codes still differ a bit in length, but sufficient references are always included.

The authors make 5 recommendations to improve fit quality and harmonization between MAX-DOAS retrievals. Is there one/many DOAS retrieval code which already have these features? If so please include this/these codes as the current state of the art. This would be especially useful for new users.
Of course some of the retrieval codes allow all recommended options, e.g. NLIN (obviously, as it was used for performing the sensitivity tests) or QDOAS in its latest version (which features the use of an interpolated reference spectrum). However, it is a bit delicate to nominate one or a subset of participating retrieval codes as the current state of the art. It is also not the objective of this work to recommend any retrieval for new users (also, some of them are in-house software and not even publicly available). Moreover, while some of the recommended best practices are intrinsic features of the code (e.g. numerical computation) others are normally options for the user (e.g. choice of background spectrum). The given recommendations are therefore valuable for both, users (best practices independent from specific code) as well as designers (intrinsic features) that either programmed the participating codes (some faults have already been found and corrected within this study) or intend to design an own retrieval in the future.

Why does the reference after the scan (T6 settings) results in larger differences? Is it simply due to the time difference between 2 degree EA and reference spectra? What is the time difference between the two spectra? Also do you see similar behavior between references taken before the scan and spectra further away? For example between refA and spectra EAnA, or refA and spectra EA2B in the following scan sequence (refA,EA2A, . . . , EAnA, refB, EA2B,. . .., EAnB, refC, ...).
Please notice that this question affects fit TR6 in Fig. 7, which is in the revised manuscript Fig. 6 as one of the previous plots have been removed in order to shorten the manuscript (following a suggestion from Reviewer #1).
In the reference fit used for differences plotted here the I_0 spectrum is the zenith spectrum before and after the scan interpolated to the measurement time. As the vertical scanning sequence starts from low to high elevations, the one before is closer to the measurement time (additional comment: tests TR4 and TR5 are therefore almost identical) than the one after the scan. So the reviewer is correct. Unfortunately, all options for the sequential reference that are currently implemented in our retrieval code are shown in Fig. 6 and at the moment no option is possible to test the referee's question. Nevertheless, it is very likely that references further away but before the scan would result again in larger differences.

There is no specific need to include all the QDOAS results in the paper. I suggest the authors consolidate the QDOAS results. This could be done by either presenting select QDOAS results or grouping all QDOAS results together for clarity (e.g. similar symbol in figure 3 or one side of the plot in figure 4). It would help compare and contrast the results between QDOAS and other codes.

We partly agree and followed the referee's suggestion indicating all QDOAS groups by the same symbol in Fig. 3 and 5 (formally Fig. 6) for clarity and better comparison to non-QDOAS groups. We also indicated QDOAS groups in Fig. 4 for the same reason.
However, we do not think that consolidating all QDOAS results is a good option because groups using QDOAS show clearly different results and no systematic similar pattern, which is an interesting finding that is explicitly mentioned in the conclusion section and has some impact for other studies: The amount of prescribed fit settings in this study is comparable to intercomparison campaigns like CINDI or CINDI-2. Consequently, groups participating those campaigns will provide intrinsic (and non-systematic) differences in their results due to small differences in non-harmonized (detailed) settings (even if using the same retrieval code), which have to be expected in the same range as observed here. We point this out more clearly in the revised manuscript. Finally, the QDOAS versions used here are different.

Line 790: "differences of up to 8% have to be expected" – Does this also hold true for other elevation angles where dSCDs are smaller? To some extent quoting 8% as expected uncertainty is somewhat misleading knowing that the particular spectrum was affected by direct sunlight and such a scenario is not common in MAX-DOAS measurements. I suggest the authors make this distinction clear in the manuscript in order to avoid misuse of 8% as inherent uncertainty in DOAS retrievals.
The value of 8% does not correspond to the first data point in Fig. 3 (although this holds true as well) but to the sequential references seen in Fig. 5 where disagreements of 8% are frequently obtained. We stated this more clearly in the revised manuscript.
In general, the value of 8% is representative for small elevations as Figure 1 (below) shows. Fig. 5 (formally Fig. 6) in the paper has been reproduced here, but this time for 8° elevation instead of 2°. Although absolute differences between groups appear to be a bit smaller in 8° than in 2°, relative differences are even a bit larger because slant columns are smaller (but nevertheless, 8% appears to be reasonable even for 8° elevation measurements shown here).

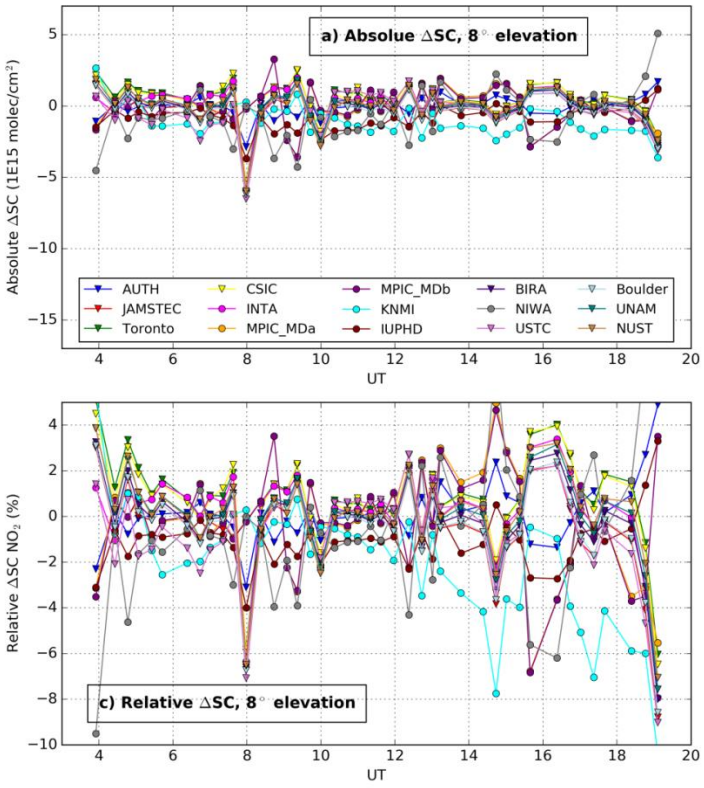

Figure 1: NO_2 differences between groups for 8° elevation angle.