Comments from anonymous Referee #3:

We would like to thank the reviewer for his/her helpful comments. We hope that we could address all questions and unclear points satisfactorily.

In the course of the revision, we have made the following important changes:

Based on a suggestion from Referee#2 we have looked into the TROPOMI AOT product. We added daily maps of the TROPOMI AOT in the Appendix of the manuscript. The lower branch visible in the TROPOMI PAL versus AirMAP comparison is mainly caused by data from 17 September (Fig. A9) and was discussed to be likely caused by a higher aerosol load which is identified as cloud in the retrieval and not treated adequately in the cloud correction, ending up with too high cloud pressures. This discussion can now be supported by the TROPOMI AOT data, which is showing a high AOT over a large area on 17 September.

During the corrections in the review process we found that the tropospheric NO2 VCD retrieval for the IUP car DOAS used an incorrect AMF of 1.5 instead of 1.3. This was corrected and Fig. 6 was updated. The correlation between the AirMAP and car DOAS measurements remains unchanged at 0.89, but the slope decreased from 0.98 to 0.89.

Referee#3 questioned the use of a NO2 box profile for the AMF calculations for the AirMAP flights, which we have stated in the text. This was an outdated information which we overlooked during the correction phase. The SCIATRAN tropospheric AMF calculations used in the AirMAP tropospheric NO2 VCD retrieval shown in the manuscript are not based on a 1 km box profile but are using a NO2 profile based on an old WRF-chem model run following a more typical urban profile, scaled to the ERA5 boundary layer height, which reached typical values of 1 km around noon. The NO2 profile is added to the Appendix.

Legend: Referee comments in black, author comments in blue

This manuscript uses DOAS data collected in September 2020 in a polluted region in western Germany from airborne, ground, and car-based instrumentation to validate the set of TROPOMI L2 NO2 products (both research and operational). The airborne datasets are first validated by the ground and car-based systems which then justifies the airborne use for validating TROPOMI. This paper fits the scope of AMT and will be valuable as a validation dataset for the TROPOMI NO2 product. However, before publishing, this manuscript requires some minor technical corrections/clarifications as detailed below but more reflection toward conclusions drawn about improvements in the S5P PAL product and the impact of clouds. Detailed comments below.

More significant comments:

Most of the results in this work are too heavily based on the slope of the regression, which is not representing the complete behavior of the validation activity. Table A1 has at least median difference in % which actually in some cases contradicts the results of the slope (e.g., having a 21% higher column from TROPOMI as a median from S5P PAL V02.03.01.). Consider more in-depth analysis based on statistics other than slope for all intercomparisons.

Thank you for the comment. We have included the median differences given in Table A1 with some additional comments in the text analyzing the comparisons and added a Figure with Box-and-whisker plots summarizing the bias and spread of the difference between the different TROPOMI versions and AirMAP tropospheric NO2 VCDs in the Appendix (see Fig. 1).

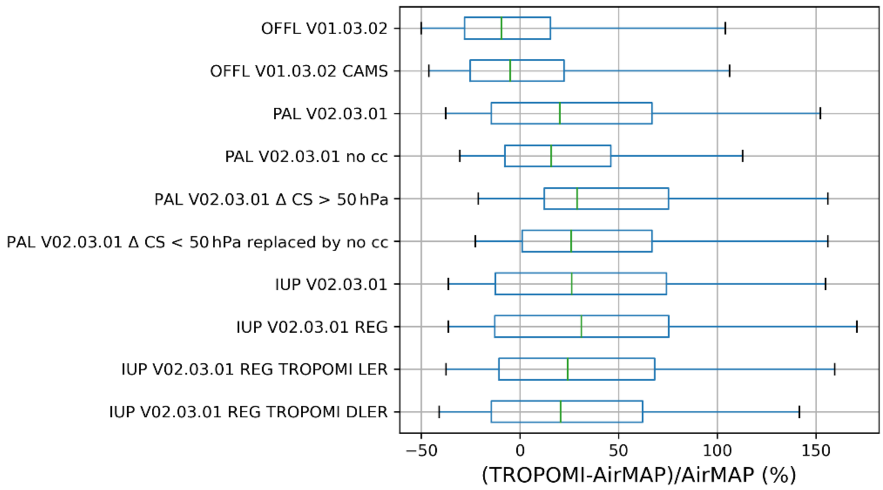


Figure 1: Box-and-whisker plots summarizing the bias and spread of the difference between the different TROPOMI versions and AirMAP tropospheric NO2 VCDs. The green line inside the box represents the median difference. Box bounds mark the 25 and 75 percentiles while whiskers represent the 5 and 95 percentiles.

Some conclusions drawn in section 6 are either overgeneralized or not quite technically correct. These comments do not specify lines in the text but more so in general comments that need to be kept in mind when adding to and editing the analysis based on the suggestions below:

With the data presented, conclusions about the S5P PAL product are only stated as an improvement. This is an overgeneralized conclusion, and the authors should do some more detailed analysis from other statistics. Some of this is already done with discussion of the lower lobe results but it is missing discussion on the higher lobe. Additionally, with the loss in precision, some users may find this result more detrimental than having a predictable low slope and this is not commented upon in the results, abstract, or conclusions.

Thank you for pointing this out. We included comments in the abstract, results and conclusion section highlighting that while the slope improved with the PAL product, the correlation of the data has decreased.

"With the modifications in the NO2 retrieval implemented in the PAL V02.03.01 product the slope and median relative difference increased to 0.83±0.06 and +20 %. However, the modifications resulted in larger scatter and the correlation decreased significantly to r = 0.72."

"The comparison of this TROPOMI product PAL V02.03.01 with the AirMAP data in Fig. 9c shows much more scatter with a correlation coefficient which is significantly poorer than for the OFFL V01.03.02 product, changing from 0.86 to 0.76. The slope, however, increased by more than a factor of 2 from 0.38±0.02 to 0.83±0.06, demonstrating that the updates in the new TROPOMI NO2 data version have a large impact on the analyzed data set from the Rhine-Ruhr region. Due to the large scatter and driven by the large amount of measurements with tropospheric NO2 VCDs of about less than $7 \pm 0.15 \cdot 10^{15}$ molec cm^{-2} the PAL V02.03.01 product has a positive median relative difference of 20% with an interquartile range of -14% to 66% (see Fig. A11)."

In this context we also changed the often used "improved retrieval" to a more neutral form like "modifications/updates in the retrieval". The higher lobe is not as much discussed as the lower lobe since we could not identify completely what is causing this higher lobe, except that it is reduced for the TROPOMI data version without cloud correction (see Sect. 6.1). Nevertheless, we added additional comments highlighting this in the results and conclusion section.

The main reason concluded about the improved PAL product is due to the cloud correction. It is stated that all these changes are due to more 'realistic' cloud pressures or more 'realistic' cloud corrections but this more 'realistic' outcome is not demonstrated in this region. Therefore, these conclusions cannot be stated unless they are proven with the data available in that specific region (e.g., could look at imagery from satellites or other creative sources and reflect on what it should be in reality). In fact, removing the cloud correction all together (Figure 9b) shows that the massive improvement in slope is something else removed from the cloud correction as this is the best result in terms of conserving precision (correlation) and a higher slope.

Previous studies showed that for scenes with low clouds, i.e. close to the surface, a height that is even closer to the surface was retrieved by the original FRESCO implementation. Since the algorithm does not discriminate between clouds and aerosols, this also holds for low aerosol layers. In many cases, FRESCO then retrieves the surface height, which is incorrect (Compernolle et al., 2021, van Geffen et al., 2022). In the old OFFL V01.03.02 product, 110 out of 117 pixels and thus 97 % of the TROPOMI observations were found to have cloud heights very close to the surface (within 50 hPa), which is not realistic and especially not for such a large amount of observations. In the new PAL product, the cloud retrieval yields only for 23% of the observations (28 out of 117 pixels) a cloud height close to the surface, which can be considered more physically realistic, resulting in a better slope of the regression line. However, since some scenes remain problematic, more scatter results. See also Figure A8 and A9 in the manuscript Appendix which show daily TROPOMI versus AirMAP scatter plots with points color coded by the difference of the surface and cloud pressure.

VIIRS images of the campaign measurement days support the on flight observations of nearly cloud free conditions during the measurement flights over the target areas. "Clouds" detected in the cloud retrieval must therefore be aerosols, which are treated as clouds in the cloud correction. For nearly cloud free observations, the cloud correction is more an aerosol correction. Whether the cloud correction actually improves the NO2 results in the presence of aerosols depends on the details of the vertical distributions of aerosols and NO2. Therefore, in some cases, the results can be better if no cloud correction is made see Fig. 9 (now Fig. 10) in the manuscript.

We have added/highlighted the mentioned points in the results and conclusion of the manuscript. We hope it is now more comprehensible.

Conclusions drawn about the cloud pressure in some cases seems to not be interpreted correctly as written. For example, discussions from line 555-563 talk about the low lobe. (1) It is stated that cloud pressures are too low, but looking at imagery online there seems to be zero clouds seen by VIIRS on this afternoon, so cloud pressures shouldn't be low to start with. (2) Aerosols are also pointed at as a potential cause, but the sensitivity results in Figure A2 show that the impact of aerosols would not be large enough to create this bias in this lobe.

Yes, the campaign days have been nearly cloud free, but since the cloud algorithm does not discriminate between clouds and aerosols, aerosols are treated as clouds in the retrieval and we have to discuss the retrieved „cloud" pressures also for these cases.

Based on a suggestion from Referee#2 we have looked into the TROPOMI AOT product. We added daily maps of the TROPOMI AOT in the Appendix. The lower branch visible in the TROPOMI PAL versus AirMAP comparison is mainly caused by data from 17 September (see Fig. A9) and was discussed to be likely caused by a higher aerosol load which is treated as clouds in the retrieval and not corrected for adequately by the cloud correction, ending up with too high cloud pressures. This discussion can now be supported by the TROPOMI AOT data, which is showing a high AOT over a large area on the 17 September.

Figure A2 (now Fig. A3) shows the impact of aerosols on the AirMAP retrieval, not on the TROPOMI retrieval. Due to different observation heights, the TROPOMI retrieval is expected to be more sensitive to aerosols than the AirMAP retrieval. Also, the effect discussed here is introduced by the TROPOMI cloud correction algorithm which is not applied to the AirMAP data.

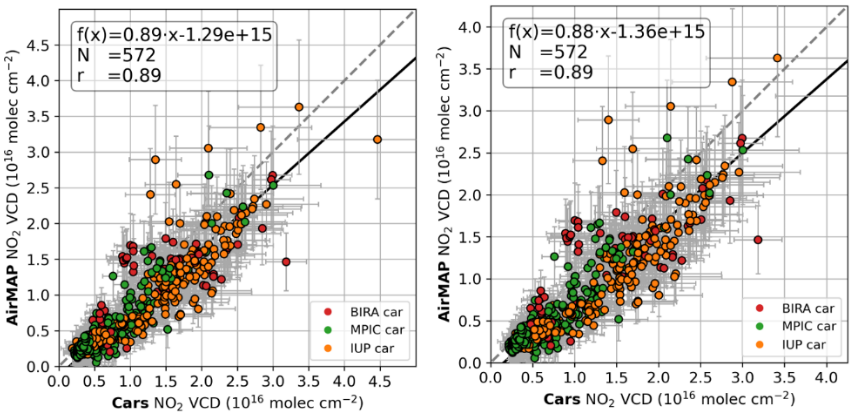
We have added/highlighted the mentioned points in the results and conclusion of the manuscript. We hope it is now more comprehensible.

Technical comments in relation to the AirMAP retrieval that need more justification or clarification.

It is said that the reference VCD in the troposphere for AirMAP is 1e15. One of the MAX-DOAS retrievals has a different value of 1.5E15 but they are referred to as similar. It is different by 50% rather than similar. Please clarify these difference or explain them.
Can the reference value be justified with any other data from this work? (i.e., What does the CAMS model say the tropospheric amount is?)
Could this reference assumption be the cause for a low offset between the car DOAS systems and the airborne dataset?

Thank you for pointing this out. Since there is no reason for using different reference values, we have decided to use the value of 1 x 10^{15} molec/cm^2 for both, the AirMAP and IUP car DOAS tropospheric VCD retrieval. The other car DOAS instruments do not rely on this value as they use dedicated measurements taken at lower elevation angle to directly estimate the tropospheric column in the reference measurement.

The influence of the mentioned difference of 0.5 x 10^{15} molec/cm^2 is not very large and cannot explain the offset between the car DOAS and AirMAP dataset, respectively only a very small part of it. Figure 2 shows scatter plots of AirMAP versus car DOAS comparisons. The AirMAP tropospheric NO2 VCDs are retrieved with a $VCD_{trop, ref}$ of 1 x 10^{15} molec/cm^2 for both plots. The IUP car DOAS data are retrieved with (a) 1 x 10^{15} molec/cm^2 and (b) 1.5 x10^{15} molec/cm^2.



(a) (b)

Figure 2: Scatter plots of AirMAP versus car DOAS tropospheric NO2 VCDs. The AirMAP tropospheric NO2 VCDs are retrieved with a VCD_{ref} of 1 x 10^{15} molec/cm^2 for both plots. The IUP car DOAS data are retrieved with (a) 1 x 10^{15} molec/cm^2 and (b) 1.5 x 10^{15} molec/cm^2.

Due to larger differences between the CAMS model and TROPOMI respectively AirMAP tropospheric NO2 VCD in distribution and amount (see paper Fig. 8 and Fig. A1) we have decided not to use the CAMS model data for the determination of the VCD_{ref}. Instead we checked the TROPOMI tropospheric NO2 VCD closest in time and space to the AirMAP reference measurement. Figure 3 shows the daily maps of TROPOMI and AirMAP tropospheric NO2 VCDs. The red cross marks the location over which AirMAP took the reference measurement. The pink cross marks the TROPOMI pixel covering this reference area. Using these TROPOMI observations would yield in a VCD_{ref} of 4.1 ± 1.5 x 10^{15} molec/cm^2. Due to the time difference between the AirMAP reference measurement and the TROPOMI observation, variations are expected and often pixel with lower values can be found close to selected pixel.

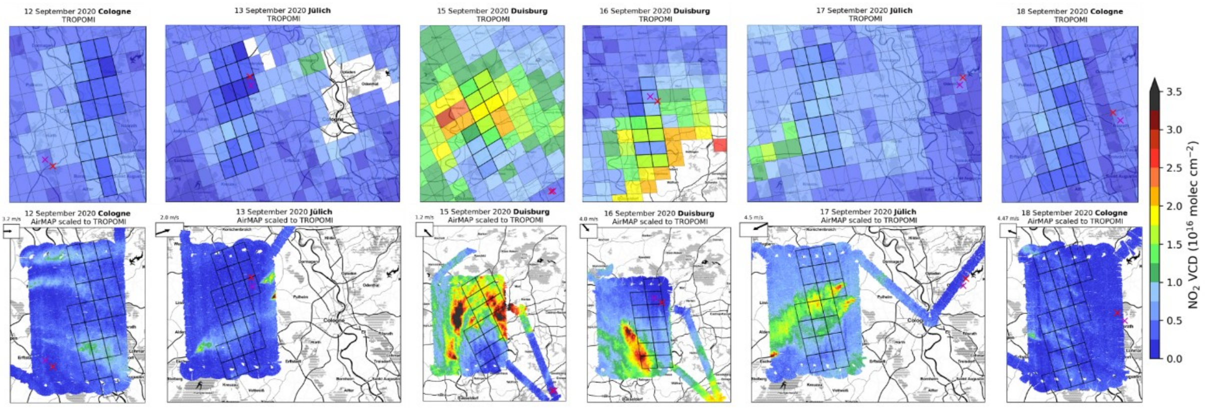


Figure 3: Daily maps of TROPOMI and AirMAP tropospheric NO2 VCDs. Red crosses mark the location over which AirMAP took the reference measurement. Pink crosses mark the TROPOMI pixel covering this reference location.

Since the TROPOMI data are indicating a higher value for the VCD_{ref}, we recalculated the IUP car DOAS data with a VCD_{ref} of 3.13 x 10^{15} molec/cm^2. Figure 3 shows scatter plots of collocated car DOAS measurements with IUP car VCDs retrieved with (a) VCD_{ref} = 1.0 x 10^{15} molec/cm^2 and (b) VCD_{ref} = 3.13 x 10^{15} molec/cm^2. The MPIC and BIRA car DOAS tropospheric NO2 VCDs are determined independently with their additional off-axis measurements in 22° respectively 30° as described in the corresponding instrument sections. As illustrated in Fig. 4, the IUP car DOAS VCDs calculated with the larger VCD_{ref} of 3.13 x 10^{15} molec/cm^2 are causing a significantly larger offset of -2.27 x 10^{15} molec/cm^2 than with the VCD_{ref} of 1 x 10^{15} molec/cm^2. Based on this comparison the IUP car DOAS and AirMAP tropospheric NO2 VCD calculations are based on the VCD_{ref} of 1 x 10^{15} molec/cm^2.

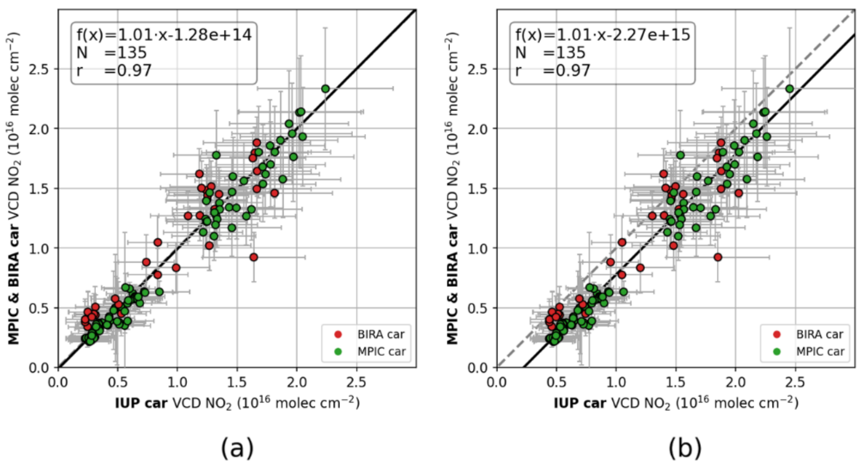


Figure 4: Scatter plot between collocated car DOAS measurements (5 min time window) of MPIC and BIRA car DOAS data against IUP car DOAS tropospheric NO2 VCDs averaged within 200 m x 200 m grid boxes and 5 min time intervals. IUP car DOAS VCDs are retrieved with (a) VCD_{ref} = 1.0 x 10^{15} molec/cm^2 and (b) VCD_{ref} = 3.13 x 10^{15} molec/cm^2.

We added a comment in the manuscript, discussing the remaining offsets in the comparisons of AirMAP versus car and stationary data:

"The comparison shows an offset of $-1.29 \pm 0.15 \cdot 10^{15}$ molec cm^{-2}. This offset could be adjusted to be closer to zero by increasing the estimated $VCD_{trop, ref}$ in the AirMAP retrieval by more than a factor of 2. However, the offset in the comparison of AirMAP and ground-based stationary data of $1.16 \pm 0.15 \cdot 10^{15}$ molec cm^{-2}. is positive instead of negative, and a larger $VCD_{trop, ref}$ in the AirMAP retrieval would further increase this offset. Because of this, and a lack of justification for a large difference between the $VCD_{trop, ref}$ for the car and AirMAP retrieval, we chose to leave the $VCD_{trop, ref}$ as it is."

Can the authors justify why a 1km box profile used if CAMS analysis is available for these flights to provide a profile shape and what that assumption impact may be in the results? A 1 km box profile assumes that NO2 is well mixed through that 1km boundary layer which has been demonstrated as not the case with in situ measurements from aircraft near strong sources (which is the case here in many of these flights). (e.g., https://doi.org/10.1002/2015JD024203 and https://doi.org/10.1525/elementa.2020.00163). This paper also shows the impact of AMFs based on assuming a 1km box vs an urban profile atmos-meastech.net/3/475/2010/

Thank you for pointing out this mistake. This was an outdated information and was missed by us during correction phase. The SCIATRAN tropospheric AMF calculations used in the AirMAP VCD retrieval are not based on a 1 km box profile but are using a NO2 profile based on an old WRF-chem model run scaled to the ERA5 boundary layer height, which reached typical values of 1 km around noon (see Fig. 5). This assumed profile is following very well the modeled and in-situ aircraft profiles from the DISCOVER-AQ campaign 2011 (Zhang et al., 2016) and the averaged urban NO2 profile from CHIMERE model runs shown in Leitao et al. (2010). We have changed the text accordingly.

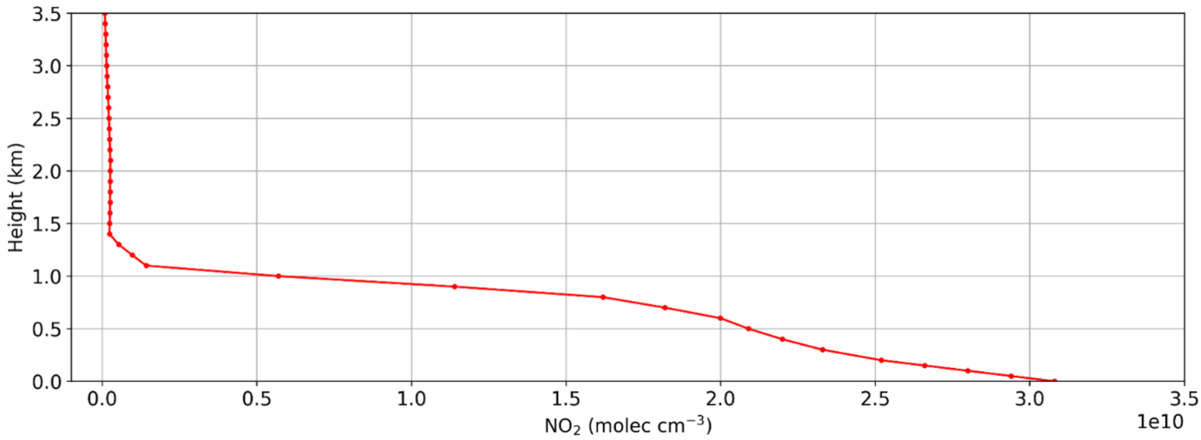


Figure 5: NO2 profile used in the SCIATRAN AMF calculations for the AirMAP measurement flights. The profile is based on WRF-Chem model runs and scaled to the typical boundary layer height during the measurement days around noon.

Line 291: 'Surfaces with different brightness introduce artefacts in the maps of NO2'. The impact isn't necessarily an artifact at the SCD stage. This is caused by the brighter surface increasing sensitivity in the lower parts of the atmosphere meaning a higher slant column if NO2 is present (if there is not any or minimal NO2 then this spatial pattern will not show up in the slant column). It only becomes an artifact if the surface reflectivity assumption in the AMF calculation doesn't account for this accurately.

Thank you for pointing this out, we have rewritten it and hope that it is clearer now.

"Bright surfaces enhance the relative contribution of light reflected from the surface to the signal received by the airborne instrument, increasing the sensitivity to NO_2 near the ground. Therefore, areas of high surface reflectance in the fitting window generally show larger dSCDs for the same amount of NO_2. Thus, differences in the surface reflectivity must be accounted for in the AMF calculations."

Minor comments:

When referring to the spatial resolution of TROPOMI as 3.5 km x 5.5 km, please specify that this is at nadir.

Thank you for the comment, we included that the resolution is given for nadir observations.

Line 74. Mention what version Verhoelst et al. validated to be consistent with this analysis and the other mentioned publications.

Done.

Line 94: the conclusion of 'low bias' is prematurely stated (before showing any results). Recommend just removing 'low' from the sentence.

Done.

Figure 2 is mentioned before Figure 1. Consider reordering figures to reflect this or consider combining Figures 1 and 2 for a more helpful side-by-side comparison.

We moved Fig. 2 before Fig. 1.

Line 159: capitalized Ozone Monitoring Instrument

Done.

Line 179-181: The sentence about V02.04.01 should either clearly state that this analysis does not include this product or should be removed.

We added a statement that this version is not included and discussed in this study since it is not yet reprocessed and thus not available for the campaign period.

Lines 173-177: The following sentence needs references: 'Other factors that could contribute to the underestimation are the low spatial resolution of the used a priori NO_2 profiles from the TM5-MP global chemistry transport model, the use of the OMI LER climatology given on a grid of 0.5° x 0.5° for the AMF and cloud fraction retrieval in the NO_2 fit window and the GOME-2 LER climatology used for the NIR-FRESCO cloud retrieval given on a grid of 0.25° x 0.25° measured at mid-morning.'

We added references to this paragraph.

Line 189: add the spatial resolution of the CAMS global analysis

Added the CAMS global resolution of 0.4° x 0.4° to the text.

Line 198-199: The sentence referring to 15% increases needs a reference.

Added the reference to van Geffen et al. (2022).

Line 308: define quantitatively what polluted means for this statistic.

We added the mean dSCD and mean dSCD error value in the text.

Equation 5 seems to be the same as equation 4. Is it needed?

Yes, this is right, we have deleted Eq.5 and are now referring to Eq. 4.

Consider making a table of all the various information of the retrievals for the AirMAP, car, and stationary DOAS retrievals as the sections get repetitive and there are small differences in places that are hard to keep straight.

Thank you for the suggestion, we added additional columns with spectrometer wavelength range, fitting window, and information about the VCD calculation and used AMF to Table 2, hopefully giving a better overview of all instruments and retrievals.

Are there references for all the individual car or ground-based systems? If so, please add in the sections that describe them.

We added references for the individual car or ground-based instruments or at least to a very similar setup as far as available.

The MAX-DOAS measurement truck is different from the rest in that it measures in the UV rather than the visible wavelengths of the other retrievals. Is it realistic for their AMFs to be the same as the other systems?

Thank you for pointing this out. We did some radiative transfer calculations using the following parameters, which are adjusted to the ground-based and AirMAP comparison times around noon regarding SZA and typical albedo and AOT values found during the campaign measurement days. Based on these calculations the dAMF in the UV is closer to 1.1 instead of the assumed 1.2 (see Fig. 6). Thus, we recalculated the MAX-DOAS measurement truck VCDs, updated the AirMAP versus stations scatter plot and the AMF information in the manuscript.

Table 1: Parameters and ranges used in the AMF calculations for ground-based measurements.

Wavelength (nm)	350, respectively 460
Viewing zenith angle (°)	90, respectively 30
Relative azimuth angle (°)	0, 45, 90, 135, 180
SZA (°)	40, 50
Albedo	0.01, 0.02, 0.03, 0.04, 0.05
AOT	0.0015, 0.16, 0.31, 0.47, 0.62
NO_2 profile	typical urban profile scaled to 1 km boundary layer height (see Fig. A3)
Aerosol profile	1.5 km box profile

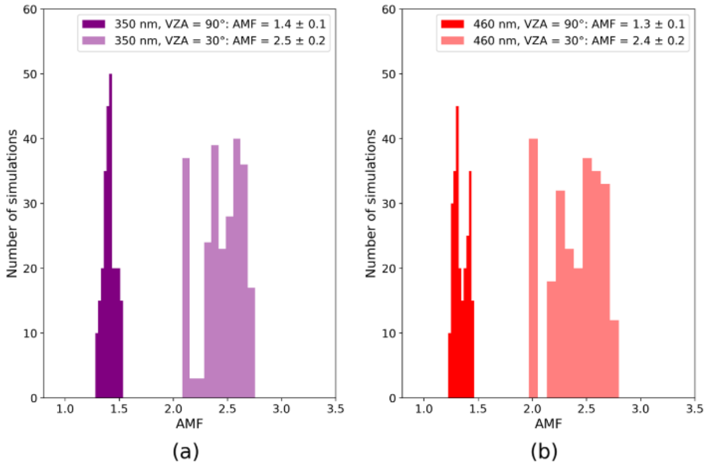


Figure 6: Distribution of AMFs calculated with the parameters of Table 1 for a wavelength of 350 nm (a) and 460 nm (b) for 90° VZA (dark color) and 30° VZA (light color).

Line 415-416. The SCD of the reference for this DOAS instrument seems quite large considering the statements that the AMFs for a zenith DOAS retrieval are about 1.3. Is this off by an order of magnitude or are the measurements just in a densely polluted area for the reference?

The SCD_{ref} given here includes the stratospheric and tropospheric NO2 in the reference spectrum. The different AMFs (tropospheric and stratospheric) and the stratospheric and tropospheric columns contributions must be considered. In order to estimate the tropospheric VCD from this value, it must be taken into account that the reference was taken during summer and therefore a relatively large part is stratospheric NO2.

Line 449-451. Is there a reference for the tropospheric NO2 product from Pandora that can be added to this section? This is the first publication I have seen use that product.

To our knowledge, there are no publications yet using the Pandora tropospheric NO2 product. We have added a reference to the Pandora readme document (Cede et al., 2021), which to our knowledge is the only document with further information on the relatively new tropospheric NO2 product.

Line 501: Is it +/- 1 hour or 30 minutes? The rest of the paper seems to reflect 30 minutes.

For the comparisons only data +/- 30 min around the S5P overpass are used. The 1 hour was given as an optimal measurement time over the target area around the S5P overpass, providing the option to adjust and investigate the effect of the temporal collocation criteria. We have restructured this paragraph to make this clearer.

Line 576: Before this line, it says that the criterion for comparison is the same as Judd et al. 2020 but at this location the authors should specify that this criterion (filtering for delta CS less/greater than 50 hpa) is the opposite of the filter applied by Judd et al. to avoid confusion. Bonus suggestion: it could be nice to have a comparison of what the results look like for the points with delta CS less than 50 hPa?

Thank you for pointing out the possible misunderstanding. We changed the text to:

"As in Judd et al. (2020) the criterion is looking for differences between the cloud pressure and the surface pressure (delta CS), but different from Judd et al. (2020), data with delta CS > 50hPa are kept and the observations where low clouds are retrieved are filtered out or replaced."

Line 604-606: 'This behavior is different from the small impact that we observed for changing the a priori NO2 profile information from TM5 to CAMS for the OFFL V01.03.02 dataset'. The change seems to be on the same order of magnitude rather than different.

Thank you for pointing this out, we changed it to: "With a relative difference of 14%, the change is showing a slightly larger impact than the 8% we found for changing the a priori NO2 profile information from TM5 to CAMS for the OFFL V01.03.02 data set."

Line 660. Saying cloud fractions are always lower than 0.14 contradicts from other examples in the text. (e.g., saying it was on average 0.21 in line 128).

Yes, right, this was a mistake. In the beginning we accidentally checked the cloud fraction instead of cloud radiance fraction to calculate the mean value. We thought we changed it everywhere in the text but have overseen it here. We changed it to the "on average 0.21 ± 0.10" as mentioned in line 128

Line 667: it is stated that on average TROPOMI is lower than air map but there are no averages reported in the manuscript.

Thank you for pointing out this formulation, we have changed it.

References:

van Geffen, J., Eskes, H., Compernolle, S., Pinardi, G., Verhoelst, T., Lambert, J.-C., Sneep, M., ter Linden, M., Ludewig, A., Boersma, K. F., and Veefkind, J. P.: Sentinel-5P TROPOMI NO2 retrieval: impact of version v2.2 improvements and comparisons with OMI and ground-based data, Atmospheric Measurement Techniques, 15, 2037–2060, https://doi.org/10.5194/amt-15-2037-2022, 2022.

Cede, A., Tiefengraber, M., Gebetsberger, M., and Spinei Lind, E.: Pandonia Global NetworkData Products Readme Document, Tech. rep., PGN-DataProducts-Readme, version 1.8-5, 31 December 2021, available at: https://www.pandonia-global-network.org/home/documents/reports/, last access: 2 December 2022, 2021.