Comments from anonymous Referee #1:

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. The lower branch visible in the TROPOMI PAL versus AirMAP comparison is mainly caused by data from 17 September (Fig. A8) and was discussed to be probably caused by a higher aerosol load treated as an effective 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 the 17 September.

During corrections in the review process it was found that in the tropospheric NO2 VCD retrieval for the IUP car DOAS an incorrect AMF of 1.5 was used instead of 1.3, this was corrected and Figure 7 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 out dated information and was overseen by us during correction phase. The SCIATRAN tropospheric AMF calculations used in the AirMAP 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.

Legend: Referee comments in black, author comments in blue

This manuscript provides an evaluation of TROPOMI NO2 vertical tropospheric columns (VTCs) against airborne NO2 observations from the AirMAP imaging spectrometer, which itself is compared with ground-based stationary and mobile DOAS instruments.

The study is excellent. The manuscript is very well written and clear and the topic fits well within the scope of AMT. I can only commend the authors for this thorough study and strongly recommend publication.

Excellence:

The study goes well beyond previous evaluation studies in several respects. The validation experiment combining ground-based and airborne remote sensing is very well thoughtout: The ground-based measurements provide a high-quality reference and the airborne observations, which effectively sampled the area covered by individual TROPOMI pixels over regions with strong NO2 gradients, provide the link to the satellite data. This setup allows for a quantitative (almost 1:1) comparison in contrast to more indirect/qualitative comparisons previous studies. Furthermore, the extensive aircraft observations over three different regions with varying NO2 levels allowed covering many individual TROPOMI pixels necessary, which is necessary to compute robust statistics. Finally, the study does not stop at presenting the comparisons but goes a long way towards explaining the reasons for errors and biases in different versions of TROPOMI data. By replicating the TROPOMI retrieval algorithm, the authors were able to analyze the influence of key input parameters such as a priori NO2 profiles and surface reflectance on the data. The main source of error was found to be the FRESCO cloud retrieval, which tended to place the cloud tops at too low elevation probably due the inability of the algorithm to properly account for the effect of aerosols. This finding is essential to guide further developments of the retrieval algorithms and improvements of the operational TROPOMI NO2 product in the future.

The study is very well written with almost no typos or grammatical errors, logically organized, well balanced in terms of conciseness and detail, the figures and tables are of high quality, and the Appendices add valuable information.

We would like to thank the reviewer for his/her nice comments.

I have only two small points to consider:

The differences in the NO2 VTCs between the former operational offline algorithm (OFFL V01.03.02) and the improved new algorithm PAL V02.03.01 are very large. The authors mention that the main change was a switch from the FRESCO-S to the FRESCOwide cloud retrieval algorithm, but there is little information on what else changed what the (potential) influence of these changes were. It would be useful to get some more insight into the changes.

Thank you for the comment. We added more details about the changes from the old OFFL V01.03.02 to the new PAL V02.03.01 version with reference to the van Geffen et al. (2022) paper in the "TROPOMI NO2 PAL V02.03.01 product version" section.

"The main change compared to the OFFL V01.03.02 impacting the tropospheric NO2 VCD data is the use of the FRESCO-wide algorithm instead of the FRESCO-S algorithm, which was already introduced in V01.04 and was operational from 29 November 2020 to 1 July 2021. The FRESCO-wide algorithm provides lower and therefore more realistic cloud pressures (i.e. clouds are at higher altitudes), especially for scenes when cloud fractions are low. This change results in decreased tropospheric AMFs, which leads to higher tropospheric NO2 VCDs (van Geffen et al., 2022b). Another update that can have a significant impact is the correction of the surface albedo over cloud free scenes by using the observed reflectance. This increases the tropospheric NO2 VCDs by about 15% over polluted regions in case the retrieved cloud fraction is zero (van Geffen et al., 2022b). For this study the effect is negligible since only 1 out of the here analyzed 117 TROPOMI 210 pixels is observed as cloud free. van Geffen et al. (2022b) also describes the following other modifications, which have only a small or no impact on the tropospheric NO2 VCD data. Level-1b v2.0 (ir)radiance spectra are updated in the new version, and are increasing the NO2 SCD of about 3 %, from which most of it ends up in a slightly increased stratospheric VCD. The improved level-1b v2.0 also leads to a small increase of completely cloud-free pixels and to slightly lower cloud pressures for pixels with a small cloud fraction, resulting in tropospheric NO2 VCDs being about 5% higher for these ground pixels. An introduced outlier removal is increasing the amount of good quality retrievals over the South Atlantic Anomaly and over bright clouds where saturation can occur. The change to new spatially higher resolved snow and ice information is increasing the amount of valid retrievals at high latitudes."

The comparisons between AirMAP and the ground-based mobile and stationary suggest that the ground-based measurements (separately analyzed by the different groups) provide a consistent set of reference measurements. Nevertheless the question arises whether there has been no direct comparison between the ground-based instruments, e.g. when a car DOAS passed by a the location of a stationary instrument or when several car instrument were placed at the same location. If such intercomparisons have been made, it would be good to learn about them and add the results e.g. in an Appendix.

Thank you for the suggestion. We added a comparison plot of the three car DOAS instruments in the Appendix (see Fig. 1).

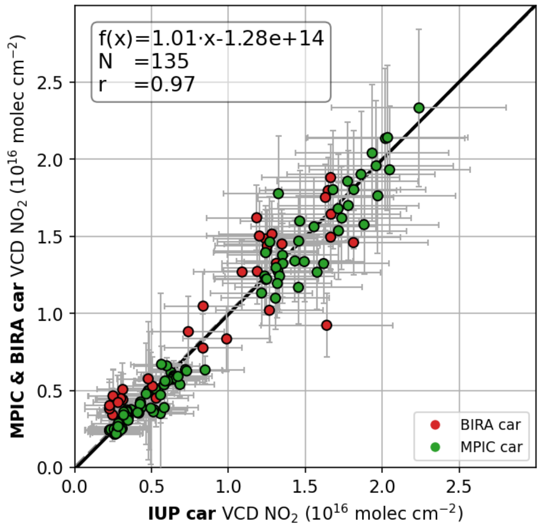

Figure 1: Scatter plot between collocated car DOAS measurements (± 5 min time window) of MPIC and BIRA car DOAS data versus IUP car DOAS tropospheric NO2 VCDs averaged within 200 m x 200 m grid boxes and 5 min time intervals. The data points from the MPIC and BIRA car DOAS instrument are color coded in green, respectively red. The thick solid black line represents the orthogonal distance regression. Error bars represent the error in the tropospheric NO2 VCD retrieval, averaged within the 200 m x 200m grid boxes and 5 min time intervals.

Small corrections:

Page 7, Line 148: Change "are retrieved" to "were retrieved"

done

Page 8, Line 179: I think the acronym DLER has not been introduced before.

Yes, this is right, we added the explanation here and deleted it later in the text.

Page 12, Equation 12: Why is the VDC_ref,trop not simply added to dSCD/AMF_trop? Why do we need to multiply VCD_ref,trop with AMF_ref,trop / AMF_trop? Please explain.

The dSCDs are slant colums retrieved relative to the reference measurement and the AMF of the actual measurement AMF_{trop} and the reference measurement $AMF_{ref,trop}$ are not the same, so adding simply the $VCD_{ref,trop}$ to dSCD/ AMF_{trop} would introduce an additional uncertainty. We added a comment in the manuscript.

"Since the AMF of the actual measurement (AMF_{trop}) and of the reference background measurement ($AMF_{trop,\ ref}$) are usually not the same, simply adding the VCD_{trop}, ref would introduce additional uncertainties."

Page 29, line 661: I think the recommended filter criterion of 0.5 applies to the cloud radiance fraction, not to cloud fraction.

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.

On page 5 line 127 we have already written "The cloud radiance fraction retrieved in the TROPOMI NO2 spectral window (cloud_radiance_fraction_nitrogendioxide_window_crb) for S5P overpass times, was on average 0.21 ± 0.10 with a maximum of 0.48 and thus for all measurements below the recommended filter criterion of 0.5."

We have changed the text in line 661 accordingly.

Reference:

van Geffen, J., Eskes, H., Compernolle, S., Pinardi, G., Verhoelst, T., Lambert, J.-C., Sneep, M., 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.