Referee report for amt-2021-39
Henk Eskes (Referee)

Referee comment on "An improved TROPOMI tropospheric NO2 research product over Europe" by Song Liu et al., Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2021-39-RC1, 2021

In their paper Song Liu et al. describe a major update of the DLR scientific retrieval of NO2 based on TROPOMI observations. The paper is generally well written and the results are presented in a transparent way. The figures give a good overview of the impacts of the innovations, consisting of the application of DSTREAM, GE-LER, ROCINN v2, OCRA and POLYPHEMUS. The paper contains a complete and balanced set of references. In general, the retrieval of NO2 is characterised by large uncertainties, mainly related to the air-mass factor, and therefore it is useful to develop alternative scientific retrievals to be compared with the operational product, with independent approaches for the AMF and stratosphere. The DLR retrieval differs from the operational one in several major ways, which makes it an interesting product to compare with. To conclude, I am in favour of publication, but I have several questions and comments, provided below, which I would like to see answered before the paper is accepted.

General remarks:

The reader may find the title and abstract confusing. It suggests that the authors are improving the operational TROPOMI retrieval product, while actually the paper discusses and upgrade of the retrieval product implemented at DLR. This should be made more clear at the top of the abstract and in the title. I suggest the authors make clear which retrieval product they refer to, for instance by calling it the DLR TROPOMI NO2 European product. Maybe good to add the word "scientific" as opposed to the "operational" retrieval.

The performance of the POLYPHEMUS model is not easy to judge. The resolution of this model is not very high, 0.2x0.3 degree, while TROPOMI has a resolution of 0.05 degree. The CAMS models for instance run at 0.1x0.1 degree. This resolution is in between the TM5-MP and TROPOMI. Would higher resolution produce a major further improvement? I would like to see a comparison between POLYPHEMUS and TROPOMI as extra figure. The
POLYPHEMUS model seems to produce very low free tropospheric NO2 concentrations. Is this expected to impact the retrieval?

It would be of interest to extend the comparisons with the operational TROPOMI product and also list differences with other (regional) retrievals like the TROPOMI-POMINO approach for Asia. Maybe in the form of a table listing the choices for albedo, cloud, stratosphere, a-priori for several retrievals. For the validation, Table 4, it would be nice if also the operational retrieval could be added (now there is only the comparison with the old DLR retrieval in brackets).

The word "improved" is used many times in the paper, not only for the final NO2 result, but also for the retrieval aspects like cloud parameters, albedo, a-priori. To my opinion one should be a bit careful with this term. One can claim something is improved if there is a better match with independent observations, or if obvious biases are removed. This is not always clear. The main conclusion of the paper is an increase of the NO2 columns which better agrees with MAXDOAS (smaller bias).

The uncertainty estimates for the individual aspects of the retrieval: cloud fraction/pressure, albedo, profiles could be discussed in more detail. It seems the authors make use of numbers from De Smedt et al, instead of deriving typical uncertainties for GE-LER, ROCINN, POLYPHEMUS. It would be good to have some rough estimates of uncertainty reductions for the individual terms in the new versus old retrieval.

In our experience with the operational product the impact of the free troposphere is not negligible. This may be discussed in more detail.

The paper mentions that the datasets are available upon request. A brief description of the product would be useful, e.g. are the input fields like GE-LER and OCRA/ROCINN cloud parameters included? Are averaging kernels included?

And finally, it would be nice if the authors can provide some recommendations for the future development of the (operational and scientific) TROPOMI NO2 retrievals in the conclusions section!

Detailed remarks:

Title: Improved compared to what? The retrieval and reference should be clearly defined.
Uncertainty strat column $= 3.5 \times 10^{14}$ "for polluted conditions". This is a bit strange: the stratosphere does not have polluted conditions.

Decrease from 55 to 34 %: what is the 55% reference?

At the end of section 6 it is shown that the comparison with MAXDOAS is affected significantly by the differences in the sensitivity profiles for MAXDOAS compared to TROPOMI. With kernel smoothing the remaining difference is about -20%. It would be interesting to mention this.

There is no reference to the data assimilation approach. It would be useful to explain a bit more this alternative approach.

Improvements compared to what? To the previous DLR algorithm or compared to the operational algorithm, or both? Even at the end of the introduction it is still somewhat unclear which two retrievals are compared.

An intensity offset correction is used, while this is not done in the operational retrieval. Would be good to have a brief discussion of the impact of this intensity offset term. How do the slant columns compare with the operational algorithm?

Sec 2.2 STREAM: How does STREAM distinguish the stratospheric background from a free tropospheric background? Please add some discussion on the free troposphere, which is supposed to be included in the tropospheric NO2 column.

"average bias of $1e13$ molec/cm2 with respect to the ground-based zenith-scattered light differential optical absorption spectroscopy (ZSL-DOAS) measurements" This is a very small number. Please provide the uncertainty range on this comparison. STREAM produces a somewhat larger column than the assimilation approach because it does not distinguish stratosphere from free troposphere.

Sec 3.1, description of DSTREAM: I am wondering if there is any interference between DSTREAM and the destriping? The directional part removes east-west biases. Does the destriping do something similar? Is destriping done before STREAM is calculated, or after?

Why is this latitude weighting introduced? Even though the diurnal cycle effect is
smaller at the equator, I would assume it could still be modelled with STREAM? For instance, average slopes with viewing angle could be accommodated as a function of latitude also near the equator.

I273: Can the difference with the model (3.5e14) be considered a true uncertainty estimate? Or is it a lower limit, e.g. because of the finite model resolution?

I 290: Does the GE-LER approach also provide an uncertainty estimate?

I290: How sensitive is the GE-LER to L1B calibration errors? How has the GE-LER been validated, e.g has it been compared with MODIS-based BRDF results? Please add some more info for the reader to judge the performance of the GE-LER approach.

I315: "The surface LER values from GE_LER are lower than the climatological OMI values by 0.03 on average" Is this a statement for February or August? It seems that the average differences in August are smaller than in February (by looking at the figures).

Fig. 7: Please comment on the very low albedo values over the Mediterranean compared to the OMI LER.

Sec 4.2: Please provide the definition of the POLYPHEMUS European domain (lat-lon domain boundaries).

I339: "European TNO-MACC". Please specify the version and reference year.

I349: "The profile shape from POLYPHEMUS/DLR agree better with the MAX-DOAS measurements". The profiling capabilities of the MAXDOAS instruments are limited and often also quite uncertain. Part of this profile shape is a-priori defined. So, does it make sense to compare these profiles?

Fig.10: It would be nice to also show a direct comparison of POLYPHEMUS tropospheric NO2 column against TROPOMI (using the averaging kernels), e.g. also for February and August. Would it be possible to add these plots?

Sec. 4.2: Has POLYPHEMUS been compared to the CAMS regional modelling results? If so, please provide a brief summary of these comparisons. Please provide more detail on how
the model has been validated in general, e.g. against observations of NO2.

Sec. 4.2: Fig.9 indicates that POLYPHEMUS has basically very low concentrations of NO2 in the free troposphere. Even though the free tropospheric column is small, this may have quite a significant impact on the AMF, and may in part be a reason why we see only red colors in Fig. 10. With larger free tropospheric NOx concentrations I would expect more blue colors especially in the more remote areas away from the main pollution hotspots. Please comment on this.

Fig.11: "comparisons with version 1.x " Please be more explicit. Do you show cf(2.1)-cf(1.x) or cf(1.x)-cf(2.1)? What is x in "1.x" in this case?

Fig.12: Why do you show a normalised CF instead of a mean CF?

Is Fig.12 consistent with Figs. 11 and 13 and the text? Fig.12 indicates smaller CF at the right side of the orbit in v2.1, while e.g. line 389 mentions an increase of the cloud fraction?

Fig.13: Again: do you show cf(2.1)-cf(1.x) or cf(1.x)-cf(2.1)?

Fig.13: Is there any evidence that the new v2.1 cloud pressures are better than the old ones? Please discuss this in more detail.

How can I understand the reduction in NO2 over the Po Valley? The cloud pressure does not seem to change much here.

l407: Could you also provide the average % difference in the AMF in winter and summer for CAL vs CBR?

Fig. 16: Please be explicit what is shown: CRB-CAL or CAL-CRB

l419: Please also provide the % increase in winter and summer.

Figures 8, 10, 16 provide the contributions to Fig. 18. However, it would be good to also
have the numbers for the contributions of the various terms (new STREAM, new profiles, GE-LER, CAL) to the increase. Could such numbers be provided for February and August?

427: I think it is good to refer to van Geffen 2020 here, who discuss this in detail. The 4.5e14 is similar/close to the 5.2e15 estimated for the operational retrieval.

433: The table in the paper of De Smedt is mentioned. Are these numbers used without any modification? Are these consistent with estimates for e.g. the GE-LER and OCRA/ROCINN CAL uncertainties? I suggest to include the relevant numbers in the text! Some more discussion on the uncertainties related to GE-LER, OCRA and ROCINN would be very relevant. For clear sky the AMF uncertainty is estimated as 20%

469: "which is mostly explained by the relatively low sensitivity of spaceborne measurements near the surface, the aerosol shielding effect, and the gradient smoothing effect." This is not so clear. The retrieval accounts for the lower sensitivity at the surface, and with the new GE-LER these uncertainties are hopefully reduced. The aerosol shielding effect was discussed as implicitly accounted for via the cloud retrieval. So it is not clear if a bias should remain due to these effects.

Table 4: It would be great if also numbers for the operational product could be included.

Figure 23: Nice to see that also the operational product is included!

Fig. 24: Nice to see this plot! The sensitivity profiles of MAXDOAS and satellite are very different, so it is good to demonstrate the impact on the comparison.

500: "see Sec 5.2" ? Section 5.2 discusses the uncertainties and does not discuss the kernels.

Conclusions section: It would be relevant to comment on the differences between the new DLR retrieval (and inputs) and other NO2 retrieval approaches (operational, NASA, POMINO, ECCC, BEHR, Bremen) and discuss possible recommendations following from this comparison of retrieval methods.

Conclusions section: The bias in the updated retrieval against MAXDOAS is reduced from 55 to 34%, but is still substantial. Do the authors have an opinion what are the main retrieval aspects causing this difference? How can this gap between surface and satellite observations be closed? Something is said about this in section 6, but it would be
interesting to discuss it again in the conclusions, and perhaps including recommendations for a way forward.