Review of "Evaluating satellite limb measurements of nitrous oxide" by Saunders et al

The paper under review fulfils the first mandatory step in creating a long-term dataset of nitrous oxide by comparing five space-born satellite measurements of nitrous oxide among themselves and to balloon-born measurements. It naturally fits into the broader community trend toward homogenizing and merging all available remotely sensed data. A clear added value of the paper is the double-checking of the recommendations of instrument teams from user point of view (HIRDLS dataset turns out to be more useful as the instrument team suggests); also inter-comparison of two versions of MLS is useful.

The paper is logically structured and well written. It fits perfectly within the scope of AMT. I recommend publication, provided that the following minor comments are addressed.

General comments:

- Narration would benefit from clear explanation of the choice of satellite instruments. I suppose satellite instruments were reduced to post-2000 limb-viewers only, but then why TES and ILAS-II are not included. I am not saying they must be included, and I suspect they are not good candidates for upcoming merging, but a rationale of the choice should be explained.
- Difference between first and second type of analysis (inter-comparison and validation) is not clearly explained: what is the "test dataset"? was the analysis performed by taking each of 5 satellite datasets as "test dataset" and only SMR-results were chosen to be shown? If not, the whole inter-comparison part looks rather like validation against SMR.

Typos / text precisions:

- l.42 : these records
- l.49: I would add https://acp.copernicus.org/articles/17/11177/2017/acp-17-11177-2017.pdf somewhere at the line 49 and explicitly mention the use of N2O there for positioning the subtropical transport barriers.
- l.55: better say SPARC CCMVal and give abbreviation
- l.67: are those all available nitrous oxide datasets?
- l.73: … balloon measurements, *all 5 satellite instruments being compared to a same reference,* so that …
- l.74: … balloon measurements *taken for reference* are described …
- l.76: … and conclusionS …
- l.84: in ACE-FTS description, v5, v5.2 and v5.3 are used and "current versions" is used for both v5 to v5.3, please rewrite accurately.
- l.112: it is 1000 profiles per day, not 1300 (Fischer 2008, section 8.2)
- l.119: … and *in nominal mode* typically measured
- l.120: 6 – 68km (not 70, Fischer 2008, section 3.2)
- l.125: .. averaging kernelS row ..
- l.126: … uncertainties constitute the total error budget ... – what does it mean?
- l.127: All MIPAS-IMK data – levels or whole profiles?
- L.129-130: The useful vertical range is from 6 to 60 km: it is said for MIPAS ESA, should earlier be also given for MIPAS IMK.
- l.130: It is taken – reformulate.
- l.138-139: it not uniformed, it is log-uniformed => uniform *log*-pressure grid.
- l.147: uniform *log*-pressure grid.
- l.147-148: corresponds to 1-sigma precision of the data – really? Please give the reference.
- l.150: less than *or equal to*, greater than *or equal to*.
- l.159: The profiles extend from about 12-70 km – reformulate.
- l.161: which corresponds to the error due to thermal noise and all interfering smoothing errors – give the reference (it is probably Urban et al 2005, but better to be clear here).

- l.163: which ranges from about 3-6 km – reformulate.
- l.204: should be explained how the dataset given on pressure grid was transformed to altitude grid: I presume you used temperature profiles from MLS, but it must be clearly stated.
- l.219: In the comparisons shown below, SMR is used as the tested dataset – this statement is misleading. It reads as you announce an inter-comparison of 5 datasets, the reader expects pairwise inter-comparisons, instead you compare 4 datasets to the fifth one (SMR) alone. Reformulate.
- l.224-232: the usefulness of both type of comparisons must be explained (which one is useful when drift will be calculated, which one when the merging will be done using one instrument as reference dataset etc).
- l.251: variabilities agree - please explain.
- l.287-288: these instruments are the outliers – please explain.
- l.333: This is due to the deviation in the N2O profile – deviation from what?
- l.343: Overall, SMR appears to perform very well compared to the other instruments; this is another reason that it was chosen as the tested instrument to highlight earlier in this section. – This statement is misleading: where the same study (i.e. "test instrument" assigned to the four other instruments) being performed within the project?
- l.365: where the abbreviation "hr" for "hour" introduced?
- l.399-400: the 1 April 2003 Kiruna flight was separated from the other Kiruna flights because no polar vortex air was sampled – please explain how it was determined.
- l.442: all the way from 9-33 km, ranging from 2-25%. – is it from 9 to 33, from 2 to 25? Reformulate
- l.450-451: It is the only satellite dataset that does not fully capture the decrease in N2O caused by descent within the polar vortex – I find the statement too strong for what is shown on the graphs.
- l.461-462: no polar vortex air was sampled during the 2003 flight – same as l.399-400

Comments to figures:

- fig.10: why this figure is shown for version 5 only, if the obvious candidate for merging is the version 6?
- caption of Fig. 17: "Summary of N2O comparisons in the tropics" => "Summary of ballon-borne vs satellite comparisons of N2O in the tropics"
- caption to Fig.2 and subsequent captions: "each other instrument" => "each of the other instruments".