The paper studies the impact of typical (relatively coarse) spatial resolutions of past, present and planned satellite missions on tropospheric NO_2 retrievals over areas characterized by a strong spatiotemporal variability in the NO_2 field. High resolution airborne GEOTASO NO_2 VCDs as well as TEMPO, TROPOMI and OMI VCDs, simulated from the GEOTASO observations, are compared with coincident observations at 10 PANDORA sites at the western shore of Lake Michigan and over the Los Angeles Basin. By combination of the different data sets, the work provides an interesting insight into the spatiotemporal tropospheric NO_2 variability over polluted areas. The impact of mismatched spatial representation has been quantified. The results and discussions are valuable in the assessment of validation strategies for the future generation of air pollution satellites. The scientific content of the paper fits well within the scope of AMT, and the manuscript is well-written and generally well-structured. However, some revisions (detailed below) should need to be conducted in the paper before publication.

General comments

The first two sections are missing some essential information that are valuable to the reader to better interpret the results and the geophysical parameters. I suggest the following:

-Please add a "Campaign" section. Some information about the campaigns is scattered in the manuscript, but a clear campaign section shortly discussing the geophysical sites, number of flights, time and duration of flights, SZA change during flights, environmental conditions, e.g. cloud fraction, etc. would improve interpretation. Information on the exact dates are for example provided for the first time on p. 5, L.9 in the context of a discussion on BRDF derived albedo.

-Section 2.1 is unbalanced when compared to 2.2 and 2.3 and not structured well. Maybe make a separation between instrument and retrieval description. I also suggest to move parts to campaign section (e.g. P.4,L.13 to L.18).

-Discussion of rasters on p. 5, L.33 should better be moved to another section (new sub-section under section 3 "Results") as it shows actual results, PANDORA locations, etc. Moreover some details are lacking again, e.g. "morning flight", is this at 6 AM or 10 AM or…? This is important for interpretation of for example traffic plumes.

-Introduction: please provide a more detailed overview of currently existing UV-VIS mapping instruments for completeness of the literature overview.

P.2, L.35: Scanning at different azimuth angles can improve the representativeness of the ground-based data when compared to satellite retrievals. Even though the focus is on direct-sun observations in this work, the potential of multi-azimuth scanning to cope with the representativeness problem should be discussed in the introduction and/or conclusion. Especially as you mention "Best practices for satellite validation strategies" in the abstract.

P.4, L.3: Please explain why this small window for NO_2 retrievals is selected and not a window which is better comparable with OMI, TROPOMI NO_2 retrievals. Moreover, in Nowlan et al. (2016) the fitting window 420-465 nm was used. Please properly refer to where the DOAS retrieval settings can be found or provide them in the manuscript.

It would be interesting for Section 3.2 to differentiate between PANDORA stations with a heterogeneous and rather homogeneous NO_2 distribution around the station (semi-background stations). This could be done based on the spatial distribution around the station observed by GEOTASO. The impact of the mismatched spatial representation, as reported in section 3.2, is expected to decrease in case of a more homogeneous distribution. An effort in this direction is done by differentiating based on the magnitude of PANDORA TropVCs and assuming that high NO_2 values can be associated with localized features and thus strong heterogeneity (P.17, L.11). This is true, but low PANDORA TropVCs do not necessarily mean that the NO_2 field is semi-background and that there are no fine-scale plumes around the site. It could be depending on the viewing geometry of the direct-sun measurement which is missing a plume or plumes present around the station.

Conclusion: Related to the mismatched spatial representation reported in 3.2, please provide as well suggestions to solve this for future operational validation of satellite data such as TROPOMI and TEMPO, based on ground-based stations. As airborne data is collected on campaign basis, it will not always be available. Do you consider more viewing angles (thus not only direct-sun observations) to add as an additional constraint? Maybe using AQ model data around the stations, providing knowledge on the emission sources and direction of the plumes? This could help to assess if the viewing direction is hitting a (localized) plume or not.

Minor comments

P.2, L.44: Replace “the” by “a” unique perspective. Mobile-DOAS measurements for example can provide as well a unique insight in the spatial variability around a ground-based station.

P.4, L.9: The overlapping GEOTASO retrievals also allow an interesting way to compare coinciding VCDs and assess the GEOTASO product quality (even if we know that the NO2 field is changing). Has this been done?

P.4, L.25: I have a bit hard time to interpret this. You are discussing larger retrieval uncertainty due to less signal (low albedo, large SZA)? The multi-linear regression is applied on which data? Details are lacking on time of flights (or SZA) to properly interpret this. Moreover, I assume flights took place with SZA smaller than 60°, so the uncertainty related to SZA should be smaller as reported?

P.5, L.26: “simplified from..” → not sure if it is needed to cite this reference as it is a commonly used equation and appears in earlier publications.

P.5, L.30: Please provide some more info on the reference used. Do you average spectra over a certain period to reduce noise? Do you use a different reference per spatial pixel in order to reduce striping effects? Is the instrument stable enough to use a single reference for the whole campaign period?

P.15, L.35: To improve readability, please repeat again explicitly which exact data sets (which days) you are comparing here.

P.15, L.30: Could you see any consistency with traffic peak times (or diurnal photochemistry) when looking at TropVC values in the 4 grids acquired on the same day?

P.15, L.47: True for ground-based vs satellite retrievals. Maybe highlight here again the advantage of airborne measurements, able to fully cover satellite pixels at high resolution.

Figure 9 and 11: Please provide the fit parameters and correlation as well in the plot or legend.

Figure 11: Pandora min and max data during the overpass +- 5 minutes can sometimes show large variations and maybe too large to be fully attributed to temporal variations. Can you shortly explain this? Outliers? I assume you lack good statistics to use 2 x st.dev. or 10-90 percentile for the whiskers.

In general for Section 3.3: Please compare your results as well with other studies that have done efforts to compare OMI with ground-based measurements, e.g. with MAX-DOAS and assess if your findings are consistent.

Section 3.3: Did you make use of the OMI averaging kernels to smooth the PANDORA VCDs in order to take into account differences in sensitivity?

Technical corrections

P.5, L.25: Remove “the”

P.7, L.36: Formulation is confusing. Maybe mention that these are the DSCD precision and accuracy or provide a typical value for the AMF, e.g. “assuming an AMF….”

P.18, L.29: Please change “city-to-regional spatial scale” to “regional spatial scale”

P.18, L.31: Please remove “very”. A priori profiles and surface reflectances can be retrieved at much higher resolutions.