Review of "Underestimation of Column NO2 Amounts from the OMI Satellite Compared to Diurnally Varying Ground-Based Retrievals from Multiple Pandora Spectrometer Instruments" by Jay Herman et al

This paper presents a comparison between OMI NASA v3.1 and PANDORA total NO2 VCD, showing a clear under-estimation of the OMI data at 7 long-term sites and 6 campaign-based sites. The results at most of the sites are presented and discussed and few arguments for the general under-estimation result are mentioned. Although the paper is interesting and fulfill the scope of AMT, there is a lack of reference to literature (previous similar studies and scientific proof/reference of why such differences at the different sites). Sensitivity tests or further comparisons on OMI pixel sizes (edge and center of the swath, different position of the pixels, GB time-selection) could be done to help justifying the proposed conclusion. I recommend the publication after the suggested revisions.

General comments:

The paper is short and easy to read, but it lack some "proof" of the proposed explanation of the OMI under-estimation (argument 1= "Because of the local inhomogeneity of NO2 emissions, the large OMI FOV is the most likely factor when comparing OMI TCNO2 to retrievals from the small PANDORA effective FOV", line 20 and argument 2= "OMI estimated air mass factor, surface reflectivity, and the OMI 24x13 km2 FOV (field of view) are three factors that can cause OMI to underestimate TCNO2", line 18).
Some sensitivity tests on the how much the choices made for the OMI pixel selection (FOV distance d<5km for an SZA<70, line 165) and PANDORA selection ("daily data matched to the OMI overpass times ±6miutes", line 87) affect the results would support argument 1 (or at least give an uncertainty range). Additional comparison (or at least further comments on other OMI retrievals, such as DOMINO (Boersma et al., 2011) or QA4ECV (Boersma et al., 2018)), would support argument 2.
Moreover, a lot of (redundant) figures are given (daily and monthly panels in Fig 3, 4, 5 and 9) could be simplified by plotting the mean and the variability – or a scatter plot of OMI vs PANDORA as often done in validation papers – while e.g., number of comparison points or impact of the Lowess(f) monthly running averages is not mentioned/discussed. How much this exercise results would change with a simple mean or median of the daily comparisons? This would allow putting an uncertainty number on the 1.8 and 1.7 PAN/OMI mentioned in page 11.
Consider adding a section or table with the different PANDORA site description, that would help the reader understanding the general differences among the stations (partially already described in the text, but bot for all sites – coordinated of the sites is also missing). This would be a good reference for future studies using these PANDORA data.
Please clarify how some justifying arguments are obtained (add references or explain not shown results). E.g. :
1) P14, line 278 "The relatively moderate TCNO2 values (0.4 to 0.8 DU) are probably a testament to the effectiveness of catalytic converters mandatory on all US automobiles in such a high traffic area" → add reference!;
2) P17, line 290 "The highest amount of TCNO2 recorded during 2018 was about 5DU on 13 July 2018 from 11:20 and 12:30 EST (a time with very light winds (1 km/hr) and moderate temperature (25°C)" → is the meteorology present at each site or only here? Could you shown some correlations? Or is this just a specificity of that time period?

Specific comments and Technical corrections

- Line 6: "14 sites" but only 13 are presented – 7 sites in table 1 and 6 in table 2. Same comment for line 13 "Eight additional sites… "

- Line 9 and 11: why mention sites in Northern Hemisphere and Southern hemisphere if this is never mentioned again in the manuscript? Same comment for line 16 “weekly or monthly average basis”: weekly comparisons are never mentioned again.
- Line 18 – 19 and 19-22: see general comment, these 2 arguments are not discussed a lot in the paper.
- line 87 – 89: the explanation on how the comparison is done is mixed between this line and lines 165. Consider adding a paragraph grouping all the comparison selection choices (cloud free pixels? What is done with the row anomaly? Why is a 6 minutes time-selection selected for the PANDORA? What type of filtering is done for PANDORA ? (cf mention of impact of clouds in line 130), …
  To my knowledge, the way the selection is done could have an impact on the results (size of OMI pixels, pixels covering the station or not, averaging the ground-based data (mean or median value?), …), and this is only poorly/not discussed. What is the impact of “the Lowess(f) monthly running averages” choice?
- Line 117: change “.” to “:”. Same for line 137 giving the link to the data: introduce it in a sentence (e.g., Data can be found here: …). Moreover, a table with coordinates and multiple names of the PANDORA stations would be helpful – “waterflow” overpass is e.g. found in the OMI link, but not on the PANDORA link.
- Lines 142-147: give references and refer to this when discussing daily and monthly evolution of fig 7 and 8.
- Line 172-174: why only give an illustration of O3 comparison for Busan? Also in table A1, there are quite some differences in the percent difference from station to station (from 0 in Baltimore to 5.6 in Mauna Loa). How is the PAN/OMI here? Is the largest difference for in O3 also at the same stations than the largest differences for NO2? Is it in stations where we expect most of the NO2 in the stratosphere (Mauna Loa)? How is the NO2 tropo/strato ratio (seen by the satellite?)
  Comment on table A1! (How to explain O3 differences of 2.5 to 2.8% at stations close to surface level?) if not here, at least in the Appendix.
- Figures 3, 4 and 5: in the monthly averages, there is often peaks not seen in OMI (shortly discussed for some stations (lines 179-180 for Busan), but not for all of them. Regularly, there is also a divergent behavior of the monthly average at the edges of the time-series (e.g., end 2016 for Mauna Loa, in 2017 for NASA HQ, end of 2017 for Waterflow, end of 2017 to early 2018 for Boulder) or OMI columns at the end of the time-series as high as PANDORA (eg Buenos aires, NASA HQ). Is this real of is this related to the “Lowess(f) monthly running averages”?
- Lines 195-196: “The calibration of the Mauna Loa PANDORA will be reviewed as part of a general data quality assurance program that is starting with the most recently deployed PANDORA instruments “ - do you mean that the PANDORA data might be off?
- Lines 209-211: there is some repetition with previous paragraphs.
- Tables 1 and 2: add coordinates of the stations and measurement time-periods. How is the “average“ among the stations performed? Mean? Median? Does it have a large effect? Consider giving the correlations. Comment on Seoul PAN = 1.2 (more than double of all the other sites!) New York value is missing.
- Line 220: give references of the Discover-AQ campaigns and discuss some of the outcomes (several PADORA on close locations; airborne flights; …) Refer also to other studies dealing

with PANDORA data for validation of NO2, eg., Judd et al., 2019 (https://www.atmos-meas-tech-discuss.net/amt-2019-161/) discussing heterogeneous NO2 situations.

- Lines 245-267: consider re-organizing the paragraphs (order and repetition). Discuss first Fig 6 completely, and then comment on Fig 7. In the comments of figure 6, reference to literature trends is missing (e.g., Duncan et al., 2016; …). It is a pity that only 6 of the 7 long-term stations are shown in Fig6. Move the discussion of the Boulder trend from the figure caption to the main text. Is there an explanation for the 3 classes of mean bias results (1) about -24 to -27% for Boulder, Mauna Loa and NASA HQ; 2) about -37% for Waterflow and 3) about -46% for Buenos Aires and Seoul) ?
- Lines 258-259: consider giving all the correlation coefficients in the tables as suggested.
- Figure 7 and 8: pity that the figures are not presented for the same year (2018), so that we could compare NASA HQ Washington and New York NO2 levels. Moreover, the TCNO2 axis limit is changing from panel to panel, so it is not so easy to see the seasonal behavior.
- Lines 278-279: "The relatively moderate TCNO2 values (0.4 to 0.8 DU) are probably a testament to the effectiveness of catalytic converters mandatory on all US automobiles in such a high traffic area". Is it purely speculative? Is there any correlation with when the regulation measures have been put in place? Give references!
- Line 284: "the pollution levels are quite high, rivaling the pollution levels in Seoul, South Korea." → this is not seen in Tables 1 and 2, and we don't have these kind of plots for Seoul, only Busan (fig 1).
- Line 293: "For both Washington DC (Fig. 7) and New York City (Fig. 8) there is strong day-to-day and month to month variability that depends on the local weather and the amount of automobile traffic in the area" – has the dependence on weather and traffic been tested or is this a guess or literature reference?
- Line 296: "Poor air quality affecting respiratory health would be improperly characterized by both the OMI average values being too low (Fig. 4) and by missing the extreme pollution events that occur frequently in the late afternoon". Also add a comment (with references) that here total columns are being analyzed, while tropospheric columns could be used, which anyway don't reflect systematically the surface concentrations important for air quality.
- Caption of figure 9: "Lowless(0.08)" it is the first time that the "f" is mentioned. Why is it different than in Herman et al., 2018 (e.g., caption of figure 9 "Lowless(0.1)")?
- Line 308-309: "there is a period in March 2018 when OMI TCNO2 slightly exceeded that measured by PANDORA." Where are those pixels? Over the sea? What is their size? What is the wind condition?
- Line 2018-2019: "The OMI underestimate is much larger than error estimates for TCNO2 retrievals for either PANDORA or OMI". Consider adding the error on some of the graphs for illustration!
- Add some discussion in the conclusion about new and upcoming satellites (eg TROPOMI with smaller pixels and geostationary that will be able to see the diurnal variation) and the uncertainties of this study (impact of the NASA product selection for OMI (wrt to DOMINO and QA4ECV) and related to the way the comparison is done (see general comment)).
- Appendix: comment on table A1 O3 results (up to 2.8% also outside mountain conditions)
- Referecences: Boersma et al., 2011 is missing. Add suggested references. Mind the formatting!

Suggested references:

Duncan, B. N., L. N. Lamsal, A. M. Thompson, Y. Yoshida, Z. Lu, D. G. Streets, M. M. Hurwitz, and K. E. Pickering (2016), A space-based, high-resolution view of notable changes in urban NOx pollution around the world (2005–2014), J. Geophys. Res. Atmos., 121, doi:10.1002/2015JD024121.

Judd, L. M., Al-Saadi, J. A., Janz, S. J., Kowalewski, M. G., Pierce, R. B., Szykman, J. J., Valin, L. C., Swap, R., Cede, A., Mueller, M., Tiefengraber, M., Abuhassan, N., and Williams, D.: Evaluating the impact of spatial resolution on tropospheric NO_2 column comparisons within urban areas using high-resolution airborne data, Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2019-161, in review, 2019.

Boersma, K. F., Eskes, H. J., Richter, A., De Smedt, I., Lorente, A., Beirle, S., van Geffen, J. H. G. M., Zara, M., Peters, E., Van Roozendael, M., Wagner, T., Maasakkers, J. D., van der A, R. J., Nightingale, J., De Rudder, A., Irie, H., Pinardi, G., Lambert, J.-C., and Compernolle, S. C.: Improving algorithms and uncertainty estimates for satellite NO_2 retrievals: results from the quality assurance for the essential climate variables (QA4ECV) project, Atmos. Meas. Tech., 11, 6651-6678, https://doi.org/10.5194/amt-11-6651-2018, 2018.