Answer to RC2:

Reviewer comments are given in black and author answers are in blue. Changes in the revised manuscript are marked in red.

Pinardi et al present inter comparison of tropospheric NO_2 columns between satellite (OMI and GOME-2A) and the ground-based (direct sun and MAX-DOAS) measurements at 39 locations over a period of 2007-2018. The authors take 3 different approaches for selecting the satellite data and intercomparison: 1) all pixels within 50 km of the ground-based site; 2) only pixels smaller than 40 km and encompassing the ground-based site; 3) account for horizontal spatial heterogeneity using dilution correction derived from OMI (2005) resampled data on 0.025_ x 0.025_ with and without ground-based data filtering over 75th percentile. The authors presented a good literature review of the prior validation work considering spatial heterogeneity in tropospheric NO_2 field. They discussed in detail uncertainties in the satellite and ground-based tropospheric NO_2 column retrievals. The authors concluded that satellites underestimated NO_2 tropospheric columns at most locations with the largest effect over the urban locations. The comparison improved if pixel size was limited and encompassed the site location. The best agreements (expressed as slopes and correlation coefficients of the linear regression analysis) were achieved from data filtering of the largest columns and applying dilution correction.
The paper is well written, addresses a very important question of satellite NO_2 tropospheric column quality and is within the scope of AMT.

Major comments:
I recommend the authors consider some reorganization of the paper. Based on the previous studies and the knowledge of the local sources it seems that the "base" case for the validation should be the smallest pixels encompassing the site locations and with the consideration of the measurement direction and horizontal extent within the pixels. After this comparison is done the authors can address the question of differences in pixel size and significantly reduced statistics by expending to include satellite data within 50 km of the site, demonstrating that this approach (as expected) does not improve the comparison even with the larger sample size. Then the authors can introduce the dilution correction method, which potentially increases the sample size and accounts for the heterogeneity. While this is a very promising technique especially if this can be applied to sub pixel heterogeneity, it is premature to call the dilution correction results "validation" due to correctly listed limitations. There are some filter selections and classifications that need better explanation, since a somewhat different selection criterion can potentially lead to a different conclusion.

Answer: *We thank the reviewer for his comments and for his suggestion to reorganize the manuscript structure. Our new baseline is now considering small pixels encompassing the site location. Since measurement pointing direction and horizontal extension are not provided for all ground-based data sets, both parameters have not been taken into account in the new baseline. The manuscript structure is reorganized as follows: Sect. 4 and 5 have been adapted for the new baseline with updates of figures 2 to 6. Figures 12 and 13 of Sect. 7.2 are shifted to a new Sect. 5.3 and become figures 7 and 8, while figures 7 to 11 become figures 9 to 12 in Sect. 6. Figures 14 to 16 do not change, and Sect. 8 becomes Sect. 7. In the revised manuscript, green is used for text moves, while red is used for other corrections/additions.*

Regarding the measurement pointing direction and horizontal extension, a sensitivity test has been performed on Xianghe data, for which both parameters are available. Comparison with OMI has been performed by selecting all daily pixels that intersect the MAX-DOAS field of view. An average of these pixels

is then computed, normalized by the segment of the field-of-view crossed by each specific pixel. This comparison only shows a small increase in the number of available coincident days compared to results with only pixels over the station (from 279 to 288 coincident). The comparison results are also not significantly changed, as can be seen in Fig.1 below.

The following sentences were added in the paper, at P. 15, lines 30-34:

As the pointing direction and horizontal sensitivity length are not reported for all ground-based instruments, our baseline approach is to consider only pixels encompassing the station location. However, a sensitivity test has been performed at the Xianghe station (where both parameters are provided in the data files) by selecting all pixels crossing the MAX-DOAS line of sight. Comparison results were found to be close to those from the baseline case, with only 10 additional coincident days.

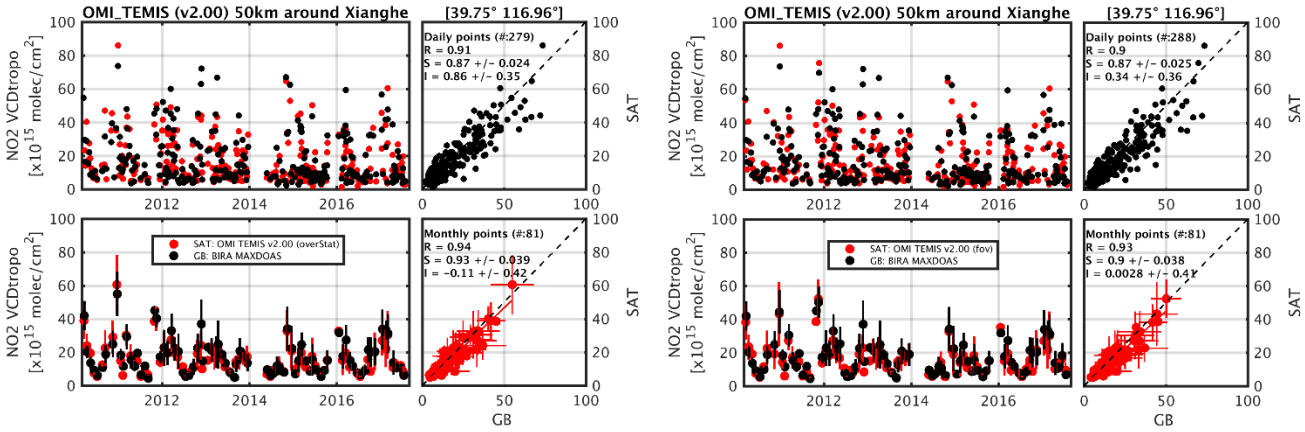

Figure 1. Daily and monthly comparisons at Xianghe when considering only pixels over the station (left), and pixels that crosses the MAX-DOAS field-of-view (right).

Minor comments:

P2, l34: Pandora provides operationally only total columns of NO_2 and O_3 from the direct sun measurements *--> Indeed SO2 and HCHO are not "operational" products yet. We reformulated to:*
"In particular, the recently developed Pandora instrument (SciGlob, http://www.sciglob.com/) operationally provides direct sun measurements of O_3 and NO_2, and SO_2 and HCHO in a scientific mode.."

P3, l20: "direct sun measurements also match better the horizontal resolution of satellite observations". DS during the summer months or near tropics at 13:30 local time will not provide a representative horizontal resolution; *-->DS, by its remote sensing nature is anyway closer to the satellite measure than an in-situ measurement, but indeed, it does not "match" the satellite horizontal resolution. We reformulated to "**Remote sensing** measurements also match…."*

P14, l29: I would recommend: "Equipped with a 2-axis positioner, direct sun-capable DOAS instruments measure non-scattered photons. Such measurements are equally sensitive to both tropospheric and stratospheric absorptions (Figure 1b). They have a very small uncertainty in AMF, and can provide accurate total column measurements with a minimum of a-priori assumptions." *-->done*

P14, l35: I would recommend: Direct sun observations are routinely available from the Pandora spectrometer instruments. A standardized Pandora network has been set-up by NASA (Herman et al., 2009, Tzortziou et al., 2014, Pandora project, https://pandora.gsfc.nasa.gov/) and expanded by ESA and LuftBlick to form the PGN (Pandonia Global Network, https://www.pandonia-globalnetwork.org/). *-->done*

P15, l27: how was the cloud radiance fraction selected? *-->the cloud radiance fraction values are coming from the satellite retrieval, and the pixels are filtered for CRF<50%.*

P16, l9: Do you mean to say: "On this basis, in addition to the daily comparisons at each station, corresponding monthly averages were also compared." If not, please elaborate why do you think daily data are accurate enough considering spatial and temporal variability and averaging? *--> Correct, this is what it meant. The sentence has been revised as suggested.*

P18, l4: I would recommend rephrasing: Due to different deployment strategies, the direct sun measuring instruments (especially Pandoras) were located closer… *-->done*

P18, l6: I would recommend rephrasing: The MAX-DOAS ensemble of stations measured NO_2 ~~total~~ tropospheric columns in the 2 to 20 x 10^{15} molecule/cm_2 range… *-->done*

P18, l7-8: I am not sure how relevant this statement is to the satellite validation since accuracy of both satellite and MAX-DOAS retrievals are impacted by the clouds. A part of the observed variability in MAX-DOAS measurements is the retrieval error since most MAX-DOAS inversion algorithms assume cloud-free conditions. *--> MAX-DOAS retrievals assume cloud free conditions and a-posteriori cloud filtering techniques can be applied (see e.g. Gielen et al., 2014; Wang et al 2015). However, since such filtering is not applied to all stations, some datasets could still contain partially cloudy scans.*

P19, l7-8. Part of the bias can also be difference in NO_2 molecular absorption cross section temperature used in DOAS analysis. MAX-DOAS is typically analyzed using 298K while direct sun (at least Pandora data) at the profile effective temperature of 254K. Spinei et al., 2014 (https://amt.copernicus.org/articles/7/4299/2014/amt-7-4299-2014.pdf) showed that at polluted sites during hot summer months this could result in 5-10% underestimation in NO_2 total column derived from the direct sun data compared to the true effective temperature. *--> We thank the reviewer for this comment. Although it is not relevant for the discussion of Figure 4 (these three specific stations are not analyzed with cross-sections at the temperature of 254K, as done for Pandora), the following sentences have been added in the discussion of Figure 11 (former Figure 9), in P. 30:*

Potential reasons are (1) the higher uncertainty in determining the true NO_2 column amount in the reference spectrum and (2) the more spatially localized direct-sun measurements, especially at high sun. Moreover, the Pandora DOAS analysis is performed with NO_2 absorption cross section at a temperature corresponding to the effective temperature of 254K, while MAX-DOAS are typically analysed for 298K. Spinei et al. (2014) showed that at polluted sites during hot summer months this could result in 5-10% of underestimation in NO_2 total column derived from the direct sun data compared to the retrieval results at the true effective temperature.

P20, l19: what definition was used for urban and suburban? Is there some specific distance and "source" size used? *--> No clear definition of urban and suburban classification was found when preparing the manuscript. Thus, the classification is based on PIs knowledge of the site, communicated to the author. A clarification has been added in the text in Sect. 5.2, P. 20:*

To illustrate this point, the different stations have been qualitatively classified by the station PIs into urban, sub-urban and background sites (see Tables 2 and 3), based on their location with respect to known pollution sources.

P20, l19: It appears that the "goodness" of the linear correlation, as shown in Fig 5, is almost entirely driven by the highly polluted sites for GOME-2A with MAX-DOAS comparison. If for some reason Yokosuka and Beijing data were removed the conclusion about the correlation "goodness" will be very different. In my opinion, the authors did not convenience the readers that using the urban-suburban classification vs. "source strength combined with the source size" help understand actual correlation between the satellite and ground-based measurements. *--> It is indeed true that the linear regression is driven by the highly polluted sites (this is why we introduced a filtering (75th percentile) to exclude the largest columns from the linear regression, considering that extreme large values at a given site usually correspond to local events not representative of satellite observations). This is the case both for urban sites (with MAX-DOAS at Beijing and Yokosuka and direct-sun at Beijing and Seoul), and for suburban sites like Xianghe. On the other hand, our qualitative urban/suburban classification is related to the locations (and strength) of the sources from the site based on PI's knowledge, and that's why a more quantitative characterization is introduced in Sect. 6.*

P27, l22: While the slope improves, the scatter actually gets worse. Adding fit RMS might be more representative of the actual fit quality. *--> RMS values have been added for the scatter plots in tables 4 and 5.*

P28, l4-5: Pandora is a spectroscopic instrument with a 2-axis positioner, diffusers and neutral density filters to allow for a wide dynamic range measurements (direct sun, moon, and multi-axis). I would recommend changing: This is likely related to the fact that, as already mentioned, direct sun measurements (specifically Pandoras) tend to be located… *--> done*

Another potential reason is also higher uncertainty in determination of the "true" amount in the reference spectrum and much more "localized" measurements (e.g. at high sun) *--> this clarification has been added in P. 31, with also the comment on the different NO_2 cross-section temperature used by the Pandora and MAX-DOAS (see above).*

P30, l13: Why 9th and 91th percentiles were chosen? *--> these limits are the default values in the box-and-whisker plot routine used.*

Fig. 11: please add the color-coding. *-->done*