General comments:

The research paper by Drosoglou et al. presented an investigation of ground-based NO_2 measurements by comparing with satellite datasets. The authors use MAX-DOAS NO_2 data collected from late March 2015 to mid-March 2016 at Guangzhou, China to compare with three different satellite NO_2 datasets, and conclude that all three satellite datasets underestimated tropospheric NO_2 concentrations. The authors also investigated effects of some coincident criteria, such as cloud fraction, distance, and averaging time. In general, the paper gives useful and important information about differences in ground-based and satellite NO_2 measurements at a specific site (Guangzhou). However, the MAX-DOAS dataset used in this study is too short (and seasonal biased) to draw further solid conclusions (such as local tropospheric NO_2 seasonal pattern), or to perform solid satellite validation work.

This paper should be published with major revisions been done.

Specific comments:

Page 4, line11. I can not agree with the author that "the MAX-DOAS observations of tropospheric NO_2 were quite sufficient to be used for the satellite data validation". The total coincident measurements were only 22 for OMI, 13 for GOME-2A, and 28 for GOME-2B (found in Table 2). I suggest this paper should not be presented as a satellite validation work, which will need some relative long-term measurements and more work (at least more than one year, without large data gap). I suggest author rephrase relevant parts in the paper, and be concerned with the use of the term -- "satellite validation".

Page 4, line 12. Here it is mentioned the lowest measurement was made with 2°, but why it is not used? On page 6, line 16, it is mentioned that NO_2 was retrieved from 15° and 30° measurements.

Page 6, line 18. Please specify which residual is referred to here. Spectral fitting residual, or maybe NO_2 dSCDs residual, or something else?

Page 7, line 13-14. Can you find some evidence to support this idea? If we use simple geometry to estimate a coarse horizontal sampling distance of MAX-DOAS instrument (e.g., 15° elevation, 1 km NO_2 layer height, as mentioned in your paper), you may find the MAX-DOAS is sampling air mass very close to the site. This makes me feel puzzled, because why GOME-2B (with the largest footprint) could "luckily" measured the days have NO_2 smoothly distributed in a larger area. Using the reference criteria group (Table 2), you have 23 coincident measurements with GOME-2B, which doubled the numbers for GOME-2A and OMI. So, maybe this good agreement is not simply due to "luck".

Page 7, line 19. Since the error bars from both MAX-DOAS and satellite data are not small, can you include error weighted fitting? I am wondering if this could improve the results (or show more insightful detials).

Page 8, line 23, Table 2. I found by tightening CF criteria, you have improved r for both OMI and GOME-2A, but (slightly) decreasing trend for GOME-2B. I do not think this is due to over strengthening the criteria, since with CF ≤ 0.2, you still have 15 coincident measurements for GOME-2B, which is about twice the number of coincident measurements from OMI and GOME-2A with CF ≤ 0.2. Do you have any comments on this?

Table 2. Another point is the CF tightening lowered the bias for the OMI and GOME-2B, but (slightly) increased the bias for GOME-2A. Any comments? With small sample size, the changing of bias and/or r should be carefully quantified, before any meaningful conclusion can be made. So, I suggest, for example, maybe calculate the confidence interval for r and bias.

Table 2. The CF criteria part is nicely visualised in Figure 9. Maybe you could perform the same visualisation for the radius and averaging time criteria. Few important information might be overlooked in such large and busy table. For example, why tightening radius criteria led a large jump in GOME-2B mean bias (from -1.8e15 to 4.6e15 molec/cm^2)? Any comments?

Page 11, line 5-8. Please provide more evidence (number) to support this idea. Maybe box-and-whisker plots of satellite and ground-based NO_2 can show/support that GOME-2B really sampled low NO_2 loading condition. It is hard to conclude this by only looking at Figure 6.

General comments:

The sample size in this study might be not good enough for satellite validation; maybe it is to sparse to draw a solid conclusion on the effects of the criteria studied here. However, even if you found the sample size is too small to provide critical results by the end, the process of drawing this conclusion is still valuable. I suggest author spend more time in tightening the conclusions made in the manuscript, and I think this could be a valuable paper.

Technical corrections:

Page 4, line 23. The I_0 effect. The subscript should be zero, not “o”. Please also correct the ones used in Table 1.

Figure 4. dAMFs is not defined in anywhere. Give unit for the x-axis.

Figure 5. Apparently, Guangzhou is not the only NO_2 hotspot on the figures. So, maybe adding location symbol for Guangzhou and a map scale will help the reader to find the city. Please do not use letter x for a multiplication sign.

Figure 9. Give unit (percentage sign) for the top panel y-axis.

Page 12, line 15. Check all your references. Use consistent abbreviations for journals.

Page 15, line 2. A comma is missing after “Atmospheric Research”. And please check if the format of page numbers is correct.

Page 16, line 29. This paper has been published on AMT; please do not cite the AMTD version.