Reviewer 1:

We would like to thank you for your comments as we appreciate the time dedicated for this review and have made changes to the manuscript to reflect the suggestions. As a note: I noticed that these comments are based on the initial submission before technical edits and not the version posted for discussion, therefore some suggestions were not as clear but were addressed to the best of my ability. Individual comments from the review are bolded with our responses in italics.

This paper presents a comprehensive evaluation of the TROPOMI satellite NO2 product v1.2 for the New York City/Long Island Sound region, where the NO2 TrVC has high spatial and temporal heterogeneity. The NO2 TrVC measurements from both airborne and ground-based Pandora spectrometers are used to compare with the TROPOMI NO2 products. While Pandora spectrometers provided continuous long-term measurements, airborne spectrometers provide observations with more spatially representative of the satellite measurements. The effects of the cloud retrieval and a priori profile on the biases in TROPOMI NO2 product are analyzed. The study is interesting and provides a welcome addition to the literatures on the measurements of the NO2 TrVCs from satellite, airborne and ground-based spectrometers. The manuscript is well written and the presentation looks good. I would recommend acceptance for publication after the following comments have been addressed.

Specific comments:

What are the exposure time for each scan of GeoTASO and GCAS during the flight and corresponding distance the airplanes flied? I did not find this information in Sect. 2.2.

Integration times for GeoTASO are fixed at 250 ms and GCAS integration times span 225 ms to 750 ms (most the time at 225 ms). I added this information to Table 2. I also added the median aircraft speeds in the text. The ground speed of the HU-25 at altitude averaged 215 m/s, therefore each exposure was ~ 53 m in distance along track. The King Air is slower than the HU-25 with an average ground speed 123 m/s. With the range of integration times, this would mean that one exposure ranged from 30-90 m. Multiple images are co-added to create a pixel size close to 250 m.

The definitions of the tropospheric column seem to be different for satellite (L141-142), airborne (L229-230), and ground-based measurements. In other wards, different 'tropopause altitudes' are used to derive the TrVCs of NO2. Considering that NO2 concentrations in the upper troposphere near the tropopause may be sufficiently large, could these differences in the definition affect the comparisons among the three data sets? How is the airborne stratospheric columns of NO2 retrieved (L268-269)?

Starting with the last question, as another review ask about this this as well. We clarified the airborne spectrometer stratospheric estimation. The airborne stratospheric component is estimated using a stratospheric NO_2 climatology developed using the PRATMO (PRather ATmospheric MOdel) Photochemical Box Model (Prather, 1992; McLinden et al., 2000; Nowlan et al., 2016). The PRATMO column is bias corrected daily using

TROPOMI NO_2 stratospheric vertical columns by calculating the average offset between the two datasets over the LISTOS domain for each day (ranging from $5x10^{13}$ to $6x10^{14}$ molecules cm^{-2}). This is stated in Section 2.3. I also avoided calling it stratospheric 'retrieved' as the actual stratospheric vertical column component is estimated from outside data (TROPOMI+PRATMO) and not directly retrieved but rather that signal is removed when doing the differential slant column to tropospheric vertical column conversion. I added text to Sect. 2.3 to more clearly state this conversion:

> *"Differential slant columns are converted to below aircraft vertical columns (assumed as the tropospheric vertical column, TrVC) by subtracting the estimated stratospheric slant column (climatology bias corrected daily with TROPOMI multiplied by the stratospheric AMF), adding the estimated reference slant column amount (from Pandora) and dividing by the tropospheric air mass factor, similar to Eq. 1 in Judd et al. (2019) or Eq 4. in Nowlan et al. (2018)."*

For each dataset comparisons, we aimed to keep the stratospheric component compatible between the reference and the evaluated measurements. These details are found within the manuscript.

1. *Pandora v. Aircraft: The estimated aircraft stratospheric column is subtracted from Pandora. The uncertainty in that assumed value is in both datasets.*
2. *Pandora v. TROPOMI: For these comparisons, the stratospheric column from the TROPOMI product is subtracted from Pandora. Therefore, that assumption is made in both datasets.*
3. *Aircraft v. TROPOMI: Stratospheric columns retrieved from TROPOMI is part of the estimated airborne column and the largest errors would likely be associated with times furthest from the TROPOMI overpass time as the slope change throughout the day is estimated from the climatology created from PRATMO. So, for airborne/TROPOMI comparisons during the time of the TROPOMI overpass time are mostly comparable. They do use different definitions for the 'tropopause' however, if there were a significant feature making a difference then we would expect to see day-by-day clustering in the comparisons, which we do not. I did go back and calculate what our 'a priori' profile is between the aircraft and the TROPOMI tropopause pressure and that value is less than $2x10^{14}$ molecules cm^{-2} (median is $1.6x10^{14}$ molecules cm^{-2}). We expect any impact to be minimal and would not affect the conclusions made about these comparisons.*

L78-79: In addition to the gradient-smoothing effect, the aerosol-shielding effect may also make a contribution to the uncertainties in the validation of satellite products by ground-based spectrometer, particularly in high-aerosol-load areas (e.g., Ma et al., 2013;Jin et al., 2016). How about the typical aerosol levels over the investigated region? Can the aerosol shielding effect be large enough to affect the comparison of Pandora with TROPOMI and airborne spectrometer measurements?

During Pandora+airborne spectrometer comparisons, over 90% of the coincidences have an AOT at 532 nm < 0.3 (measured by the co-located airborne lidar, HALO), two coincidences are above 0.5 with a max of 0.7. In the supplement, Figure S1 shows the comparison colored by AOT. We added text with the details about AOT during these coincidences to give readers a gauge on aerosol loading. We have also discussed aerosol impacts during outlier coincidences as possible causes during individual cases, though did not find strong evidence that they were a regular impact.

We do not have regular AOT measurements to coincide with the Pandora sites for Pandora/TROPOMI coincidences, therefore we rely on the Pandora algorithm to filter out scenes that are extremely aerosol polluted. In this work, we only use direct-sun measurements, so most of the signal measured by the spectrometer in clear scenes is from the direct solar beam. In the presence of clouds or heavy aerosols, scattered light can become a small fraction of the light observed by the Pandora direct-sun measurement and in the case of heavy aerosol loading (~AOT>1), these measurements are flagged as lower quality (like a cloud). This work only includes high quality measurements, therefore, cases of extreme aerosols are inherently filtered out.

In scenes with lighter aerosol loading, there is the potential for aerosol impacts to the TROPOMI retrieval. In Sect 4.1 we mention that scattering from aerosols are assumed as indirectly sensed through the cloud retrieval, though it is not explicitly accounted for in the TROPOMI retrieval. During aircraft coincidences with TROPOMI, AOT at 532nm measured by HALO has a mean of 0.22 and a standard deviation of 0.15. This detail been added to Sec. 4.1 for additional context on aerosol loading. In future work, we plan to use aerosol profile measurements from HALO to estimate the sensitivity to aerosol loading in this region, which would enable us to more explicitly answer this question.

For now, we have added the following text to the paper in Sect 7. that references sources promote that aerosol impacts should be included in future investigations

> *"One component not explicitly explored in this work, that should be in the future, is the potential impact of aerosols on the TROPOMI retrieval and whether their indirect accounting through the cloud retrieval accurately reflects the impacts within the radiative transfer calculations for the air mass factor calculation (e.g., Leitão et al., 2010; Ma et al., 2013; Jin et al., 2016)."*

Technical issues:

L21-23: please rephrase the first sentence in the Abstract. It should be stated that

the measurements were made or the measurement data were collected. Better to

describe more clearly which coincided with the early measurements from the Sentinel-

5P TROPOMI instrument?

I reworded the first sentence to say: ' Airborne and ground-based Pandora spectrometer NO_2 column measurements were collected during the 2018 Long Island Sound Tropospheric Ozone Study (LISTOS) in the New York City/Long Island Sound region which coincided with early observations from the Sentinel-5P TROPOMI instrument.'

L37: change 'biggest' to 'largest'.

Change made as requested.

L124: the words 'to be' can be deleted.

I changed the phrase: 'to span late June through..'

L153: how is the qa_value defined?

QA_values are defined within the TROPOMI product file. This is not a value that I am defining. Information on how it is defined is located in the references in this part of the discussion (particularly the product user's manual; Eskes et al., 2019).

L163: what does the dynamic range of NO2 refers to?

The dynamic range is referring to the range of NO_2 columns observed from day to day. This can vary day to day from very clean (less than $1x10^{15}$) to very polluted (up to $100x10^{15}$). The main point in this discussion is that the peak in the annual average in that area is $12x10^{15}$, but day to day variations can be quite a bit more or less polluted than that.

L74: please check the phrase 'through June 30'. Did Hu25 fly only one day?

Table 3 has a summary of the flights and 'through June 30th was referring to all flights prior to and on June 30th. I changed the phrasing to say 'GeoTASO was flown on the NASA LaRC HU-25 Falcon during the three June flight days…'

L187: please give the pressure altitude in hPa.

Instead, I removed the word pressure and refer to it as aircraft indicated altitude, as I am referring to the aircraft being set to fly at an altitude of 28,000 ft according to its altimeter.

L790; 'This is the first work that airborne spectrometer measurement dataset has been used to . . .'?

Changed the sentence to say 'This is the first work that uses an airborne spectrometer dataset to evaluate the TROPOMI tropospheric NO_2 product.'

Figure 2: please add x10^(15) to the labels of both x-axes and y-axes.

This label is on both axes.

References

Ma, J. Z., Beirle, S., Jin, J. L., Shaiganfar, R., Yan, P., and Wagner, T.: Tropospheric NO2 vertical column densities over Beijing: results of the first three years of groundbased MAX-DOAS measurements (2008-2011) and satellite validation, Atmos. Chem. Phys., 13, 1547-1567, 10.5194/acp-13-1547-2013, 2013.

Jin, J., Ma, J., Lin, W., Zhao, H., Shaiganfar, R., Beirle, S., and Wagner, T.: MAXDOAS measurements and satellite validation of tropospheric NO2 and SO2 vertical column densities at a rural site of North China, Atmospheric Environment, 133, 12-25, http://dx.doi.org/10.1016/j.atmosenv.2016.03.031, 2016.