Response to Reviewer #1

We thank Reviewer #1 for reviewing the manuscript and for their helpful comments. We agree with the majority of the reviewer's feedback and feel that these comments have led to an improvement in the quality of the manuscript. All reviewer comments are in italics and the author's responses are in standard font.

This study evaluates OMI and TROPOMI retrievals of NO2, HCHO and FNR using aircraft measurements during the LISTOS campaign. The manuscript is well-written, and it is a good for for AMT. See my comments below.

- *The authors focus on statistical results of the comparison, especially the mean biases. But I don't think that mean biases could tell much about the uncertainties of TROPOMI and OMI. The standard deviation of the mean biases is large, which made me wonder if the overestimates of underestimates are broadly consistent. If not, presenting the mean biases here may not help understand the performance of satellite retrievals. For example, how well do these retrievals capture the spatial and temporal variability of NO2, HCHO, and FNR? And how the errors in satellite retrievals affect the interpretations of the ozone sensitivity?*

In the updated manuscript we now include root mean squared error (RMSE) statistics to help demonstrate the uncertainty (this term is used throughout the updated manuscript to describe all unresolved errors beyond systematic biases such as random errors and relative biases) in both OMI and TROPOMI. We have also deemphasized the discussion about median bias in the updated manuscript in order to allow for more equal focus on systematic biases and uncertainty/unresolved errors in the retrievals.

In response to a comment by Reviewer #2 below, we address the capability of OMI and TROPOMI to capture the spatial variability of NO_2, HCHO, and FNRs observed. As for the temporal variability, low earth orbit (LEO) satellites obtain, at best, a single snapshot per day, so we don't get much temporal information from these spaceborne systems. What we can demonstrate is the capability of the satellites to capture the inter-daily magnitude variability of NO_2, HCHO, and FNRs observed by airborne spectrometers. To demonstrate this, we calculated daily mean tropospheric column quantities of NO_2, HCHO, and FNRs from both satellites and airborne data for the entire LISTOS domain, and within 0.35 degrees of the NYC city center (identified as the emission source region), to calculate daily correlation statistics. The following text was added to Sect. 3.4.2 of the updated manuscript to summarize this evaluation and results "Given the limited spatiotemporal data coverage provided by the LISTOS campaign, a robust understanding of the temporal capabilities of OMI and TROPOMI to retrieve FNRs is not possible. LEO satellites obtain, at best, a single snapshot of both HCHO and NO_2 each day, so one could only hope to obtain daily variability of FNRs from these spaceborne systems. To determine whether OMI and TROPOMI could capture the variability of the daily mean tropospheric column quantities of NO_2, HCHO, and FNRs over the entire LISTOS domain from airborne data, we compared these daily mean values from NASA OMI, QA4ECV OMI, and TROPOMI to the airborne observations. For NASA OMI, daily

correlation (R^2) values were 0.85 (p = 0.001), 0.58 (p = 0.03), and 0.26 (p = 0.20) for NO_2, HCHO, and FNRs, respectively. For QA4ECV OMI, daily correlation values were 0.85 (p = 0.001), 0.80 (p = 0.002), and 0.47 (p = 0.06) for NO_2, HCHO, and FNRs, respectively. For TROPOMI, daily correlation values were 0.92 (p = <0.001), 0.85 (p = <0.001), and 0.41 (p = 0.03) for NO_2, HCHO, and FNRs, respectively. All daily correlation statistics for HCHO and NO_2 were significant to a 95% confidence interval and suggest that both OMI and TROPOMI can capture the overall inter-daily magnitudes of FNR indicator species. However, only TROPOMI could observe the daily variability of domain-wide FNRs within a 95% confidence interval. This suggests that unresolved errors in either HCHO or NO_2 retrievals (the analysis from this study suggests uncertainty in HCHO are driving FNR bias variability) from OMI, using both the NASA and QA4ECV algorithms, are too large to confidently capture the inter-daily variability in FNRs.

The same analysis was conducted for NASA and QA4ECV OMI except just for retrievals near the large anthropogenic source regions in NYC (within 0.35 degrees of the city center) where relative errors due to satellite retrievals for FNR calculations were the lowest (see Fig. 6). Daily correlation (R^2) values for FNR retrievals near the source region of NYC for NASA OMI (0.13; p-value = 0.39) were reduced compared to domain-wide means and QA4ECV OMI (0.66; p-value = 0.01) correlations were improved near the source region of NYC. Indicator species correlation values from NASA OMI were degraded compared to the domain-wide analysis suggesting that this satellite product may not be able to capture inter-daily variability of FNRs even in large source regions. However, this analysis suggests that QA4ECV OMI data has the capability to retrieve daily variability of FNRs in large emission regions such as NYC to a statistically significant level. Overall, TROPOMI retrievals at both fine and coarse spatial resolutions evaluated in this study are able to capture daily variability of tropospheric FNRs over the entire domain and emission source regions better compared to OMI products.”.

To gather a more complete picture of the extent to which each satellite retrieval product lose spatial information (variance) compared to airborne data, we follow a recent algorithm named SpaTial Representation Error EstimaTor (STREET) (Souri, 2022) using NASA OMI and TROPOMI retrieval data. This method creates semivariograms determining the changes in spatial variability with distance for a defined variable (for this case HCHO and NO_2 trace gas columns). The following description and results were added to Sect. 3.4.2 of the updated manuscript “To understand the extent to which OMI and TROPOMI retrieval products lose spatial information (variance) compared to airborne data during the LISTOS campaign, we applied the algorithm named SpaTial Representation Error EstimaTor (STREET) (Souri, 2022) using NASA OMI and TROPOMI retrieval data. This method creates semivariograms determining the changes in spatial variability with distance for a defined variable (for this case we used tropospheric column HCHO and NO_2). The maximum variance at which the modeled semivariogram levels off is defined as a sill and data sets with larger sill values possesses richer spatial information. Figure S10 shows semivariograms, and the fitted stable Gaussian function described in Souri et al. (2022a), applied to TROPOMI and NASA OMI compared to airborne NO_2 columns. Concerning the comparison of TROPOMI and airborne data at 0.05° × 0.05° resolution, we observe airborne semivariogram as high as 20×10^{15}

molecules cm^{-2}, a factor of two larger than what TROPOMI achieves. At a ~20 km length scale, TROPOMI can only observe ~40% of the airborne spatial variance, indicating that the spatial representation error in TROPOMI is ~60% at this scale. Similarly, NASA OMI fails to recreate >50% of the maximum variance observed in airborne data at 0.15° × 0.15° resolution. At ~20 km length scale, the spatial loss of OMI is >70%.

Figure S10 depicts the semivariograms and fitted exponential curves applied to TROPOMI and airborne HCHO columns. Immediately evident is that both semivariograms level off at longer distances compared to the analysis of NO_2. This stems from the fact that HCHO columns tend to be spatially more homogeneous in the region of the LISTOS domain. For most length scales, TROPOMI can relatively well replicate the spatial variance observed in airborne data (~70%), which is explainable by the fact that HCHO concentrations are not highly heterogeneous in this region. We do not present the semivariogram for NASA OMI HCHO columns as the underlying unresolved biases in OMI are very large, introducing artifacts that cannot be solely attributable to unresolved spatial scales. Overall, TROPOMI and OMI capture spatial variance of NO_2 similarly, TROPOMI performs slightly better; however, OMI is unable to capture the spatial variability of observed HCHO due to unresolved biases in this retrieval product. Since TROPOMI is able to capture the observed HCHO variability to a sufficient degree, combing these two facts suggest that TROPOMI has better capability to retrieve FNR spatial variability compared to OMI products.".

As for the impact of satellite retrieval errors on the interpretation of O_3 sensitivity, the recent study by Souri et al. (2022a) shows that satellite retrievals errors, in particular the unresolved bias in HCHO products, is the largest source of uncertainty in using satellite FNRs to investigate O_3 sensitivity. Here we propagate the errors calculated from NASA OMI, QA4ECV OMI, and TROPOMI to FNR calculations during LISTOS using Eq. (15) from Souri et al. (2022a) and created maps of relative error shown in a new Fig. 6. The following text has been added as Sect. 3.4.1 of the updated manuscript "There are numerous sources of error when using satellite retrievals of tropospheric column HCHO and NO_2 for investigating surface-level or PBL O_3 production sensitivity regimes. The primary uncertainty sources are using indicator species to infer the complex chemistry driving O_3 production and destruction, horizontal spatial representation error, uncertainty in converting tropospheric columns to PBL and surface-level values, and satellite retrieval unresolved biases (Souri et al., 2022a). As for the impact of satellite retrieval errors on the interpretation of O_3 sensitivity, the recent study by Souri et al. (2022a) shows that satellite retrievals errors, in particular the unresolved error in HCHO products, are the largest source of uncertainty in using satellite FNRs to investigate O_3 sensitivity. Here we propagate the uncertainty (RMSE) calculated from NASA OMI, QA4ECV OMI, and TROPOMI to FNR calculations during LISTOS 2018 using Eq. (15) from Souri et al. (2022a) and created maps of the relative error (see Fig. 6). From this figure it can be seen that satellite retrieval errors in HCHO and NO_2 contribute significantly to satellite-derived FNR relative errors. In the largest NO_x emission source regions of NYC, where combined column abundances of HCHO and NO_2 are largest, is where the lowest relative errors of FNRs occur. For TROPOMI, which has the smallest values of uncertainty/RMSE compared to both NASA and QA4ECV OMI algorithms for HCHO and NO_2, relative errors are as

low as ~40%. Away from the emission region of NYC these relative error values reach as high as ~80%. Similar patterns of relative error in FNRs from NASA and QA4ECV OMI retrievals are derived; however, the lowest relative error values over NYC are ~50% and reach values up to 100%. The largest relative errors are seen outside the source region of NYC in QA4ECV OMI retrievals due to having the largest uncertainty in HCHO and lower column abundances of this species in the rural regions of the domain. In addition to the fact that the less noisy retrievals from TROPOMI result in lower relative errors in FNR data, Fig. 6 further demonstrates the larger uncertainty in OMI as the relative error patterns are more heterogeneous. The spatial averaging of TROPOMI data results in the lowest relative errors of all four satellite products discussed in this study. TROPOMI at the coarser (0.15° × 0.15°) spatial resolution had relative errors as low as 35% and only increase to ~60% outside of the source location of NYC.".

- *It's also not clear to me how the statistical results drawn from a single field campaign can be generalized to other regions or other time periods. I'd strongly recommend the authors go beyond the statistical comparison, and have a more thorough discussions about the sources of uncertainties, and the associated errors, and whether their conclusions can be generalized.*

We agree with the review that more flight days during the campaign would be ideal. But this campaign provided a unique opportunity to use airborne remote-sensing observations of tropospheric column NO_2 and HCHO to validate both OMI and TROPOMI coincidently (the overlap of both spaceborne sensors is novel). Also, the airborne sensors allowed for evaluation of OMI and TROPOMI over large areas which equates to having hundreds of ground-based systems for validation. While having long-term observations for robust validation of satellite sensors is ideal, this case study is unique in that it provides information about the performance of both OMI and TROPOMI over variable emission source regions (urban to rural) and scenes with differing physical characteristics (e.g., surface albedo, tropospheric compositions, clouds, etc.). This is now emphasized in Sect. 2.3 of the updated manuscript. Furthermore, to provide the reader information about the statistical significance of the satellite/airborne data comparison correlation values in Table 2 which are statistically significant to the 95% confidence level are identified in the updated manuscript.

Finally, in response to Reviewer #2, in addition to this comment, Sect. 3.4.3 of the updated manuscripts now discusses potential sources of systematic bias and uncertainty in OMI and TROPOMI HCHO and NO_2 retrievals and how they impact satellite-derived FNR products. The text for Sect. 3.4.3 is as follows: "As demonstrated in this study, median biases of OMI and TROPOMI HCHO and NO_2 retrievals tend to cancel out when calculating tropospheric column FNRs. Figures S4 and S5 show that the median bias spatial distribution of all satellite HCHO and NO_2 retrievals are similar with a small low median bias in column abundances near the source region of NYC and high biases in the background regions. Table S1 shows that AMF calculations from NASA OMI, QA4ECV OMI, and TROPOMI use many of the same input data sets for geophysical variables (e.g., surface albedo, cloud fraction, cloud radiance, etc.) resulting in

campaign-averaged AMFs of HCHO, NO_2, and the ratios of these products (AMF FNRs) which are relatively similar across the LISTOS domain (see Fig. S11). For all satellite products, HCHO and NO_2 AMFs have much less variability compared to AMFs derived for airborne data which along with SCD biases may contribute to the median high biases in background HCHO and NO_2 retrievals. A primary reason for the inability of satellites to capture AMF variability over the LISTOS domain is likely the shape factors being used for these calculations having spatial resolutions of 1.0° × 1.0° to even coarser grids (Table S1). Furthermore, while TROPOMI and QA4ECV OMI retrievals used daily model data for shape factor calculations, NASA OMI uses monthly products which will be challenged to capture the large spatiotemporal variability of tropospheric HCHO and NO_2 vertical profiles in urban and rural regions occurring in reality. Finally, coarse geophysical input data sets used in AMF calculations (see Table S1) will not capture the spatial distribution of these variables in reality. Airborne AMF calculations use much higher spatial resolution input data sets (e.g., 500 m surface albedo data (Judd et al., 2020) compared to 0.5° × 0.5° or coarser surface reflectivity products used in OM and TROPOMI) and shape factors are calculated with 12 km × 12 km CMAQ model simulations which both aid in the much larger spatial variability of AMFs not captured in satellite retrievals.

The more interesting aspect found in this study is that unresolved errors in HCHO and NO_2 retrievals don't cancel out in FNR calculations as do the systematic/median biases. While there are some reasons why uncertainty in HCHO and NO_2 retrievals could stem from opposite impacts of geophysical parameters in AMF calculations, such as AMF uncertainties in HCHO and NO_2 having opposite trends with increasing surface reflectance (comparing Fig. 10 from De Smedt et al. (2018) and Fig. 20 from Liu et al. (2021)), these differences are minor and overall AMF calculations for both species in NASA OMI, and QA4ECV OMI, and TROPOMI have similar input data sets. A portion of the uncertainty of HCHO and NO_2 retrievals not canceling out stems from the AMF calculations shown in Fig. S11. In order for HCHO and NO_2 AMFs to have no impact on VCD uncertainty cancelations, AMF FNRs would be a constant or similar value at all locations. However, from Fig. S11 it is shown that AMF FNRs, while having smooth spatial variability, are not a constant value. Therefore, some of the unresolved error residual in the FNR calculations will be due to differences in HCHO and NO_2 AMF calculations. This is emphasized in NASA OMI AMF FNR plots in Fig. S11 where different CTMs, at different spatial resolutions (see Table S1), are used to derive HCHO and NO_2 shape factors leading to noticeable differences in the respective AMF calculations. This likely is one of the reasons that NASA OMI FNRs have the largest uncertainty (highest bias standard deviation and RMSE values) compared to airborne data (see Table 2) of all OMI and TROPOMI satellite products. Finally, the airborne AMFs are more variable compared to satellite products due to the finer-scale shape factors and geophysical parameter input data used in AMF calculations which satellites inherently are not able to capture, contributing to the satellite uncertainty.

The rest of the remaining unresolved error in FNR calculations is likely due to the SCD retrievals from OMI and TROPOMI sensors. As demonstrated in this study the uncertainty in both OMI and TROPOMI retrievals of HCHO is large. The SCD retrievals of HCHO from TROPOMI have been

shown in the past to have less noise compared to OMI due to the higher spatial resolution and at least the same signal-to-noise (De Smedt et al., 2021). The larger uncertainty in OMI retrievals of HCHO compared to TROPOMI directly leads to the higher bias standard deviation and RMSE values for derived FNRs in OMI compared to TROPOMI (see Table 2). This is further emphasized in the spatially-averaged TROPOMI data (at 0.15° × 0.15° to match OMI data) where HCHO and FNR retrievals have a factor of 2-3 lower RMSE compared to NASA OMI and QA4ECV OMI. TROPOMI NO_2 SCDs have also been shown to have less noise compared to OMI retrievals due to the higher spatial resolution and similar signal-to-noise (van Geffen et al., 2020, 2022). This is also shown in Table 2 when averaging TROPOMI data to match the OMI spatial resolution. Overall, HCHO and NO_2 SCD noise contributes to uncertainty in OMI and TROPOMI VCDs and are not cancelled out in FNR calculations; however, the reduced noise in TROPOMI SCD retrievals leads to improved VCDs of HCHO and NO_2 abundances and the ratios of these products.".

Specific Comments:

Abstract: The abstract is lengthy. I'd suggest the authors shorten the abstract to include only the core findings of this work. For example, the first paragraph may belong to introduction.

The abstract has been shortened as much as possible.

Line 370: What are the quality flags for? Is this the same quality flag as for TROPOMI? If so, why do you choose different thresholds? Better to include references here.

OMI data user's manuals for NO_2 and HCHO state the in order to use the highest quality data that only pixels with qa_values = 0. For TROPOMI, the individual species data user's manuals for NO_2 and HCHO make recommendations for the qa_values used in this study (0.75 for NO_2 and 0.5 for HCHO) in order to use high quality data. The HCHO data user's manuals for TROPOMI recommends removing data with qa_values < 0.5 which we followed in the original manuscript. We tested whether using the higher qa_value of 0.75 to filter TROPOMI HCHO retrievals would impact the results of this study. When removing TROPOMI HCHO pixels with qa_values < 0.75 the statistics had a very minor change and the results of the study remained consistent. In the updated manuscript we have added the following clarifications in Sect. 2.5: "Satellite retrievals with high quality were filtered for use by removing individual retrievals that did not have quality flags (qa) = 0 for HCHO and NO_2 when applying OMI data. This qa value is suggested in the OMI data user's manuals for the application of the highest quality science data and for the removal of OMI pixels impacted by the row anomaly. For TROPOMI, individual retrievals of NO_2 and HCHO that had qa < 0.75 and qa < 0.5 were removed prior to spatial averaging, respectively, as recommended by the TROPOMI data user manuals for each species.".

Table 2: I'd suggest include an estimate of the error, such as normalized mean standard errors. NMB doesn't tell much about the precision of the retrievals.

We have added root mean squared error statistics throughout the updated manuscript.

Line 530: Maybe you could have a figure of the mean biases of HCHO to show where OMI or TROPOMI HCHO is biased high?

Supplemental Fig. S4 and S5 in the updated manuscript now show the spatial distribution of campaign-averaged OMI and TROPOMI NO_2 and HCHO biases during LISTOS.

Line 600: I'm not sure if we could call this as a 'high pollution' day because ozone was actually low on this day. This could very well be a cold day when the lifetime of NO2 is long, and the the photolysis is low. I'm not sure how much value there is to evaluate FNR on this day. It'd be more interesting to add another day with both high ozone and high NO2.

In response to a comment by Reviewer #2, and to reduce the length and complexity/density of the manuscript, we have removed this section from the updated paper version as it did not add much value to the overall study.

Line 700: It is interesting to see that improved the a priori from CMAQ does not improve the retrieval performance of OMI. The authors attribute this to coarse resolution of OMI. Could this be due to the coarse resolution of cloud and surface albedo data used in the retrieval?

We attribute the degradation in OMI retrieval performance when using high spatial resolution CMAQ data for a priori data primarily due to the too steep shape profile in the model data. This was explained in detail in the original manuscript. Overall, for tropospheric NO_2 retrievals from OMI, the shape factor from CMAQ has higher NO_2 concentrations in the PBL and lower values in the free troposphere compared to the a priori profiles used in standard NASA OMI retrievals resulting in lower air mass factors (AMF). As discussed in this manuscript, and in other papers referenced in the manuscript (e.g., Goldberg et al., 2017), the differences in shape factors produced by CMAQ in comparison to the coarser global model output use as a priori information in OMI is the primary reason for the poorer performance in NO_2 retrievals. We compared the "raw" WRF-CMAQ output to airborne observations and confirmed the model biases and "scaled" the higher spatial resolution model data to better replicate observations. While median biases of OMI NO_2 and HCHO retrievals reprocessed with the scaled WRF-CMAQ a priori information was only moderately improved compared to the standard retrievals, the spatial variability of the species was better retrieved compared to observations. The following text updates were added to Sect. 3.3 of the updated manuscript to reflect this point: "Scaled NASA OMI tropospheric column NO_2 and HCHO retrievals had smaller median biases of $-0.3\pm3.9 \times 10^{15}$ molecules cm^{-2} and $4.4\pm7.1 \times 10^{15}$ molecules cm^{-2} and much lower RMSE values of 3.9×10^{15} molecules cm^{-2} and 7.8×10^{15} molecules cm^{-2}, respectively, compared to the retrievals with raw WRF-CMAQ predictions. This result demonstrates the need for accurate shape factors (i.e., vertical distribution of trace gases) to be used as a priori information in NASA OMI retrievals. Finally, the improved accuracy of tropospheric column NO_2 and HCHO retrievals using scaled WRF-CMAQ model predictions resulted in a slightly higher magnitude of FNR median bias (0.5 ± 3.2); however, with lower RMSE values, compared to reprocessed data using raw CMAQ predictions. In comparison to standard NASA OMI products, the reprocessed satellite data using scaled WRF-CMAQ a priori information had similar median biases in FNR values and lower median biases for HCHO ($4.4\pm7.1 \times 10^{15}$

molecules cm^{-2}) and NO_2 ($-0.3\pm3.9 \times 10^{15}$ molecules cm^{-2}). All reprocessed data variables using scaled model simulated shape factors, due to the reduction in uncertainty in retrieve HCHO and NO_2 data, had lower RMSE values, higher correlation (except for FNR), and similar to better linear regression slopes compared standard satellite retrievals.".

A detailed discussion of how the coarse spatial resolution input geophysical data sets (e.g., cloud fraction/radiation, surface reflectance, etc.) used in OMI and TROPOMI, can contribute to systematic bias and uncertainty in the FNR retrievals is now provided in Sect. 3.4 of the updated manuscript.

Line 835: While the low mean biases of FNR is low, the standard deviation is very large. The R sure is also low for FNR. Thus I don't think the errors of HCHO and NO2 could cancel out. The errors in HCHO and NO2 can offset only if the errors are correlated. I'd suggest the authors make a scatter plot for errors of HCHO versus NO2, and see if they are correlated.

The manuscript has been re-written in a way that it is clear that the systematic/median biases cancel out in FNR calculations. However, as identified by the reviewer, the uncertainty in HCHO and NO_2 retrievals do not cancel out in FNR calculations. Biases for HCHO and NO_2 retrievals from NASA OMI and TROPOMI are not correlated with R^2 values <0.05. This is described in detail in the updated manuscript in Sect. 3.4.

References

Goldberg, D. L., Lamsal, L. N., Loughner, C. P., Swartz, W. H., Lu, Z., and Streets, D. G.: A high-resolution and observationally constrained OMI NO_2 satellite retrieval, Atmos. Chem. Phys., 17, 11403–11421, https://doi.org/10.5194/acp-17-11403-2017, 2017.

Souri, A.: ahsouri/STREET: STREET 0.0.2 (0.0.2). Zenodo. https://doi.org/10.5281/zenodo.6993116, 2022.

Souri, A. H., Johnson, M. S., Wolfe, G. M., Crawford, J. H., Fried, A., Wisthaler, A., Brune, W. H., Blake, D. R., Weinheimer, A. J., Verhoelst, T., Compernolle, S., Pinardi, G., Vigouroux, C., Langerock, B., Choi, S., Lamsal, L., Zhu, L., Sun, S., Cohen, R. C., Min, K.-E., Cho, C., Philip, S., Liu, X., and Chance, K.: Characterization of Errors in Satellite-based HCHO / NO_2 Tropospheric Column Ratios with Respect to Chemistry, Column to PBL Translation, Spatial Representation, and Retrieval Uncertainties, Atmos. Chem. Phys. Discuss. [preprint], https://doi.org/10.5194/acp-2022-410, in review, 2022a.

Souri, A. H., Chance, K., Sun, K., Liu, X., and Johnson, M. S.: Dealing with spatial heterogeneity in pointwise-to-gridded- data comparisons, Atmos. Meas. Tech., 15, 41–59, https://doi.org/10.5194/amt-15-41-2022, 2022b.