Response to Reviewer #2

We thank Reviewer #2 for reviewing the manuscript and for their helpful comments. We feel that the changes suggested have resulted in improvements in the quality of the study. All reviewer comments are in italics and the author's responses are in standard font.

Johnson et al. present a detailed statistical analysis of FNR observations with two different OMI products, and with TROPOMI. FNR, NO2 and HCHO satellite retrievals are validated against airborne measurements over the New-York area during summer 2018. It is well demonstrated in the paper that the noise of the HCHO satellite retrievals is the limiting factor for the FNR observations since the individual HCHO columns need to be averaged at poorer time and space resolution than NO2. The precision of the OMI HCHO observations does not allow for daily FNR observations at OMI native resolution. The OMI QA4ECV HCHO product is found to perform better than the OMI NASA HCHO product. TROPOMI offers an important improvement in the spatial and temporal resolution of HCHO and NO2 tropospheric columns, allowing for daily FNRs retrievals at TROPOMI native spatial resolution. The results are further improved by averaging TROPOMI observations on a larger spatial grid. Both NO2 and HCHO satellite products suffer from bias compared to aircraft observations. This study identifies an important positive bias over rural regions (lowest columns) for both species, and for OMI and TROPOMI products. However, the positive bias found for the TROPOMI products is reduced compared to OMI thanks to the better spatial resolution and lower noise. It is also demonstrated that the bias of the FNR satellite observations is much lower than the respective NO2 and HCHO biases. This is an important result that would deserve more discussion in the paper. The paper is well written, albeit a bit long and too detailed. The scientific approach is solid, however some points should be tested or clarified. I recommend publication in AMT after some revisions.

General comments

One concern is the small number of days that are available for the validation. Here the field campaign covers only a few days (8 days collocated with OMI, 12 days with TROPOMI). The statistical results are not always significant, especially for OMI. Studies on longer time period could improve the observed correlations, that are poor for HCHO.

We agree with the review that more flight days during the campaign would have been ideal. However, the LISTOS field 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 the 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,

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.

• *The paper could be improved by providing an information about the spatial and temporal resolution that might provide useful FNR observations with OMI (ex. monthly averaged data). How many observations are needed at minimum to reduce the noise at the level of the TROPOMI daily observations?*

We agree with the reviewer that this is an interesting aspect of applying satellite data to derive FNR data for O_3 production sensitivity analysis. See a very similar response to Reviewer #1 for their comment about spatiotemporal resolution needed for trend studies. In response to both reviewers, we added an entire section (Sect. 3.4.2) in the updated manuscript which describes the capabilities of OMI and TROPOMI to observe spatial and temporal variability of FNRs during LISTOS.

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.".

• *It would be interesting to know if the HCHO observations with aircraft instruments are also noisier than the NO2 observations, and therefore also the limiting factor of suborbital FNR observations.*

It is expected that HCHO retrievals will be nosier compared to NO_2. There are two primarily reasons for this: 1) optical depths for HCHO peak in the UV range (<380 nm) at the same wavelengths coinciding with large Rayleigh scattering and optical depths of ozone leading to a weak/noisy signal, and 2) the stronger NO_2 optical depths in the visible wavelength range (400-500 nm), where there are higher signal-to-noise ratios, permits retrievals with less noise. Nowlan et al. (2018) derived the precision of the GCAS/GeoTASO airborne remote-sensing systems used for NO_2 and HCHO retrievals in this study. Nowlan et al. (2018) quantified precisions of 1.0×10^{15} molecules cm^{-2} and 1.9×10^{16} molecules cm^{-2} at a fine spatial resolution of 250 m × 500 m for NO_2 and HCHO, respectively. Averaging these precision values to the spatial resolution of 0.05° × 0.05° improves these precision levels to 6.4×10^{13} molecules cm^{-2} and 1.2×10^{15} molecules cm^{-2} for NO_2 and HCHO, respectively. The campaign-averaged column NO_2 and HCHO abundances from GCAS/GeoTASO at 0.05° × 0.05° were 6.6×10^{15} molecules cm^{-2} and 1.5×10^{16} molecules cm^{-2}, respectively. Comparing the precision values of Nowlan et al. (2018) to the mean abundances during LISTOS at the same spatial resolution results in mean precision levels of 1% and 8% for NO_2 and HCHO, respectively. Overall, the HCHO airborne data for is expected to have a factor of 5-10 more noise compared to NO_2. Text describing this have been added to Sect. 3.2.6 of the updated manuscript.

The reason why we resort to using precision statistics from Nowlan et al. (2018), and not the LISTOS data set, is that airborne flight tracks during LISTOS were focused on the source region of NYC, and surrounding areas, which did not allow us to define a "clean" region for both NO_2 and HCHO. A caveat to using precision statistics from Nowlan et al. (2018) is that the observations were obtained at different locations/times and under different atmospheric and viewing geometry conditions which could results in different signals. However, we feel that the large difference in noise derived in the manner explained above is sufficient to assume that the airborne HCHO retrievals are nosier compared to NO_2.

• *In Table 2, I recommend adding a line providing the mean value +- the standard deviation of FNR, HCHO and NO2 for the aircraft, NASA OMI, QA4ECV OMI, and TROPOMI (0.15° and 0.05°).*

This information has been added to Table 2 of the updated manuscript.

• *I recommend more tests on the selection of the data, that is currently at the edge of the statistical significance (see later).*

One interesting result of the paper is that the errors in NO2 and HCHO columns tend to offset in the FNR observations. There might be good reasons for this, such as error cancellation. It is

therefore important to use HCHO and NO2 products that have been retrieved with algorithms and auxiliary data as consistent as possible. This is an important message for the future TEMPO product.

In response to this comment and Reviewer #1, the manuscript has been re-written in a way that it is clear that the systematic/median biases of HCHO and NO_2 retrievals tend to cancel out in FNR calculations. However, the uncertainty in HCHO and NO_2 retrievals when compared to airborne observations do not cancel out. This is clear as the unresolved error/RMSE values for FNRs are still large. Furthermore, biases for HCHO and NO_2 retrievals from NASA OMI and TROPOMI are not correlated with R^2 values <0.05. This is now described in detail in the updated manuscript.

Below, in response to your next comment, and in the updated manuscript, we discuss in detail about potential reasons why median errors tend to cancel out while unresolved errors do not.

• *It would be good to discuss further what type of error might cancel out, or at least might reduce, when using NO2 and HCHO retrieved using consistent algorithms to derive FNR (surface albedo, cloud products, a priori profiles).*

In direct response to this comment a discussion section (Sect. 3.4.3) of the updated manuscript has been added. Furthermore, two additional sections have been added (Sect. 3.4.1 and 3.4.2) to discuss the spatial and temporal capabilities of OMI and TROPOMI as well as relative errors of these satellite retrievals. 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.".

• *I recommend adding a table providing a quick look at the auxiliary data used in the AMF calculations for the NASA, QA4ECV and TROPOMI products, and TEMPO.*

This has been added as Supplemental Table S1 in the updated manuscript.

• *Discuss the different FNR biases with the level of consistency between NO2 and HCHO AMF settings.*

Please see the discussion above, and new Sect. 3.4.3 in the updated manuscript, which addresses this comment.

The low HCHO correlations are also partly due to lower spatial variability of the HCHO distribution compared to NO2, also in the airborne measurements, over the time and domain of the study.

We agree with the reviewer that the spatial variability of HCHO is lower compared to NO_2 during the study. However, we feel that the low correlation of the satellite/airborne tropospheric HCHO data is primarily due to the inability of the satellites to capture the spatial variability of observed HCHO.

Selection of data:

• *Filter row anomaly both for HCHO and NO2 products.*

The pixels impacted by the row anomaly were removed using data quality flags in both OMI HCHO and NO_2.

• *The lower bound limits for HCHO and NO2 appear to be strict, compared to the reported standard deviations of the bias. For HCHO, the bias std ranges from 9 to 5e15 molec.cm-2, while the lower limit has been set to -8e15. For NO2, bias std is about 4e15, while the lower limit has been set to -1e15 molec.cm-2. There is a possibility that a significant part of the negative values has been filtered out while it actually belongs to the normal distribution. The effect could be an artificial increase of the mean background values. Please test a lower bound limit for the data selection.*

The lower and upper bounds were based on suggested limits used in recent OMI and TROPOMI validation studies (e.g., Zhu et al., 2020) and personal communication with OMI NO_2 retrieval team. However, to test whether the lower limit of HCHO and OMI impacted the high bias in background concentrations retrieved in OMI and TROPOMI, we reduced the lower bound of HCHO and NO_2 to -5.0×10^{16} molecules cm^{-2} -1.0×10^{16} molecules cm^{-2}, respectively. The statistical comparison during LISTOS was not impacted by reducing this lower limit.

• *At the spatio-temporal resolution of the study, OMI retrievals are clearly at their detection limit. Please consider testing a lower grid resolution (0.2°) for OMI.*

As explained in Sect. 2.6 of the original manuscript, now Sect. 2.5 of the updated version, we apply a point oversampling technique when spatially averaging the retrievals. When averaging OMI data to the 0.15° × 0.15° spatial resolution (standard radius of 0.075°), we employed a radius twice the

standard size equal to 0.15°. This helps avoid issues due to the fact that the 0.15° × 0.15° grids are near the native spatial resolution of OMI at nadir.

• To increase the number of collocations, I would suggest testing a larger temporal window of 3h for the airborne retrievals.

Increasing the temporal threshold to lengths greater than 1 hour will increase temporal data representativity error. We have already adopted a longer temporal threshold compared to other satellite validation studies using GCAS/GeoTASO observations during LISTOS (e.g., Judd et al., 2020). This issue is particularly true for NO_2 near the surface where it's lifetime can be minutes to hours. However, to test how increasing the temporal colocation threshold to 3 hours would impact the statistics we conducted this sensitivity test for NASA OMI. As expected by the reviewer, this increased satellite/airborne colocations by ~70%. However, it degraded the statistical evaluation of the satellite retrievals especially for NO_2 where median biases, bias standard deviations, correlation, and RMSE were noticeable worse compared to using a temporal colocation threshold of 1 hour. Given that correlation statistics are mostly significant to a 95% confidence interval using the limited number of colocations (see our response above) using the threshold of 1 hour, and we already use a temporal threshold longer than others evaluating satellites with GCAS/GeoTASO observations, we kept our statistical analysis using the threshold of 1 hour for our updated manuscript.

It would be good to better stress the specificities of this paper compared to the recent paper of Souri et al., 2022, which also compares OMI and TROPOMI NO2, HCHO and FNR errors over the US. (Characterization of Errors in Satellite-based HCHO / NO2 Tropospheric Column Ratios with Respect to Chemistry, Column to PBL Translation, Spatial Representation, and Retrieval Uncertainties)

The recent paper by Souri et al. (2022a) assessed the major error components of retrieving FNRs using satellite data. The primary driver of uncertainty in satellite-derived FNRs identified in this study was from systematic bias and unresolved error of the NO_2 and HCHO retrievals themselves (HCHO retrieval uncertainty being the main issue). Souri et al. (2022a) estimated TROPOMI biases using stationary point-source observation data (MAX-DOAS) and OMI errors using airborne in situ data. This study builds off these findings to better characterize OMI and TROPOMI FNR retrieval error using a unique validation data set (i.e., GCAS and GeoTASO) providing coincident NO_2 and HCHO information obtained during the LISTOS field campaign. This particular data set has not yet been used to assess HCHO, NO_2, and resulting FNR retrieval errors. As explained in the response above to the reviewer comment about the choice of using LISTOS campaign data, this unique data set provides information about the performance of coincident HCHO and NO_2 retrieval from 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 emphasized in the updated manuscript.

Detailed comments

Abstract

Line25: "high spatiotemporal coverage": please provide numbers, such as the native resolution of OMI and TROPOMI. I would rephrase "OMI and TROPOMI are capable of providing NO2 and HCHO daily global observation at native resolution of respectively … and …". However, satellite observations are known to be affected by noise and biases, that limit the precision of FNR.

In order to shorten the abstract, in response to Reviewer #1, we have removed much of this discussion as we provided these details in the main body of the text.

Line 25: "…, yet a recent study suggested ….". This sentence is rather vague. Which study?

This statement has been removed from the abstract.

Line 30: Please specify the covered period.

This has been added to the abstract.

Line 32: Please be clearer in the abstract with the term "suborbital". This is not obvious for a general reader.

This has been replaced with "aircraft-based".

Line 49: Place replace large by larger biases.

Corrected.

Introduction

Line 95: please add the 2 following references: Wang et al, 2022; Harkey et al., 2015.

These references don't use both satellite HCHO and NO_2 to study ozone production sensitivities so would not be appropriate to cite here.

Line 100: the choice of references seems weird. It might be good to add references for NO2 and HCHO L2 products of each sensor, and not only for studies using both species together. The SCIAMACHY instrument is missing in the list.

This sentence has been updated to read "Multiple past and current space-based spectrometers have the capability to retrieve simultaneous NO_2 and HCHO tropospheric columns to calculate FNRs for studying O_3 production sensitivity regimes including…" in order to emphasize the purpose of this statement. The purpose of this statement is to identify spaceborne systems which have been used for studying O_3 production sensitivity regimes and is why we chose the specific references. As requested by the reviewer we have added SCIAMACHY to this sentence.

Methods

Line 163: The OMI rows affected by the row anomaly should be filtered out in the HCHO product such as in the NO2 product. The reference sector method does not correct for the row anomaly, but for the stripes between the valid rows. Please rephrase (and check that the HCHO data are filtered correctly).

The pixels impacted by the row anomaly were removed using data quality flags in both OMI HCHO and NO_2. The text in the updated manuscript is now clearer in this section: "The row anomaly in NO_2 and HCHO retrievals was avoided in this study using data quality flags to filter out rows/pixels flagged by the row anomaly detection algorithm."

Line 204: Please explain what you mean by "iterative fitting algorithm" and "simultaneous fitting". To me, a DOAS fit is an iterative fit (least-squared fit).

We thank the reviewer for identifying our misinterpretation of the literature. Given this statement was not necessary for the study, it has been removed in the updated manuscript.

Line 215: The QA4ECV fitting window is 328.5-359 nm, such as TROPOMI. For all HCHO products, please double check the retrieval intervals that are mentioned in the paper. Most of the recent retrievals use a fitting window larger than 328.5-346 nm.

We thank the reviewer for identifying this error in the text. The fitting window ranges have been corrected for each sensor/algorithm.

Line 261: Please explicit the term SWs.

This sentence has been removed as described above.

Results

Line 444: "Tropospheric columns NO2 concentrations", "tropospheric columns NO2 retrievals". Could be simplified to "Tropospheric NO2 columns" and homogenized throughout the paper.

These phrases for tropospheric column NO_2 and HCHO have been simplified as suggested by the reviewer.

Line 465: It should be emphasized here that TROPOMI offset for low columns is lower than OMI at the resolution of 0.05.

The following sentence was added to the updated manuscript: "TROPOMI at its near native spatial resolution has the least high bias of background tropospheric NO_2 columns demonstrated by the lower y-axis intercept compared to all OMI and TROPOMI data products at the coarser spatial resolution".

Line 491: add a reference to Verhoelst et al. 2021.

Added.

To our knowledge, the cited references do not report a high bias of NO2 for background values. But the studies were made with the previous version of the TROPOMI NO2 product. This should be clarified here.

We agree with the reviewer and the sentence has been correct to read: "The results here suggest that OMI, and to a lesser extent TROPOMI, tropospheric column NO_2 retrievals errors have a magnitude dependence and tend to have some high bias in rural/background regions and a low bias in moderately to highly polluted regions which agrees with past validation studies (e.g., Zhao et

al., 2020; Lamsal et al., 2021; Verhoelst et al., 2021).". OMI has been shown to have a high bias in clean regions (e.g., Lamsal et al., 2021) which are larger compared to TROPOMI (Zhao et al., 2021). Many studies show that OMI and TROPOMI NO_2 data compare well to stations located in clean/background sites; however, this study applying airborne remote-sensing data is better able to retrieve clean and polluted regions in the same location on the same day compared to previous validation sites.

Line 495. The comparison of TROPOMI NO2 Bias at 0.05 and 0.15° also clearly shows the spatial resolution effect on the background values (from negative to positive and similar to OMI NMB). Please mention this resolution effect.

The following sentence has been added: "Finally, TROPOMI NO_2 data averaged to the coarser spatial resolution of OMI has a similar campaign-averaged high median bias as both OMI retrieval algorithms; however, displayed RMSE values nearly twice as small as NASA and QA4ECV OMI, further emphasizing the importance of spatial resolution for retrieving tropospheric NO_2 columns.".

Table2: Please add one line with the mean FNR, NO2 and HCHO columns and their standard deviations.

This information has been added to Table 2.

Figure 3: Please test different data selection as suggested in the general comments.

We tested the lower limit of NO_2 and HCHO values, increased qa_values for TROPOMI HCHO, and increased temporal colocation threshold as suggested by the reviewer (described above). Decreasing the lower limit of NO_2 and HCHO values from OMI and TROPOMI and increasing the qa_value for filtering TROPOMI HCHO had no impact on the statistical results of the study and were kept the same in the updated manuscript due to selecting these values based on satellite data user's manuals, past validation studies, and personal communication with algorithm teams. Increasing the temporal colocation threshold to 3 hours increased the number of colocations for statistical evaluation; however, also increased spatial representation errors which degraded the statistics of the satellite retrievals (especially for NO_2 which has a shorter atmospheric lifetime compared to HCHO). These suggestions were good for testing the robustness of our satellite evaluation methods; however, for the reasons above were not included in the updated manuscript.

Line 517: The results are not so much in agreement with the study of Vigouroux, who reported indeed a high bias for the lowest columns, but for columns lower than 2.5e15 molec.cm-2. The TROPOMI bias ranges from 0 to negative values for columns larger than 5e15 molec.cm-2.

See our response to the similar reviewer comment below.

Line 527: Please also compare the bias standard deviation between OMI and TROPOMI.

More emphasis on discussing bias variability and uncertainty using RMSE statistics for all retrievals has been added to the updated manuscript.

Line 530: In De Smedt 2021, it is reported that the OMI HCHO offset is larger than for TROPOMI. But the reported bias are all negative for columns larger than 5e15 molec.cm-2. The conclusions of this study are therefore not completely in agreement with De Smedt et al. or with Vigouroux et al..

In order to provide a more quantitative comparison with the recent validation studies of OMI and TROPOMI HCHO (Vigouroux et al., 2020; De Smedt et al., 2021), we separated our collocated satellite/airborne data points using clean ($<5.0\times10^{15}$ molecules cm^{-2}) and polluted ($\geq8.0\times10^{15}$ molecules cm^{-2}). We chose a slightly higher threshold for separating clean HCHO columns to optimize the number of colocations for statistics and to be as similar as possible to Vigouroux et al. (2020). We also added a highly polluted threshold ($>16.0\times10^{15}$ molecules cm^{-2}) to further emphasize our results. The table below summarizes the median bias ± bias standard deviation and NMB results for NASA OMI, QA4ECV OMI, and TROPOMI at coarser/fine spatial resolution for the different HCHO column magnitudes.

Statistical evaluation of NASA OMI, QA4ECV, and TROPOMI retrievals of tropospheric column HCHO. Statistics presented are median bias ± bias standard deviation and NMB (%).

NASA OMI (0.15° × 0.15°)				QA4ECV (0.15° × 0.15°)			
	Clean	Polluted	Highly Polluted		Clean	Polluted	Highly Polluted
Bias	2.8±6.2	4.6±7.9	-2.3±9.2	Bias	2.7±7.3	2.1±8.7	-3.8±7.4
NMB	75.1	30.3	-8.9	NMB	72.1	13.7	-14.6
TROPOMI (0.15° × 0.15°)				TROPOMI (0.05° × 0.05°)			
	Clean	Polluted	Highly Polluted		Clean	Polluted	Highly Polluted
Bias	3.1±1.4	1.8±4.4	-2.2±4.8	Bias	2.4±2.3	1.3±6.5	-2.7±7.0
NMB	78.1	12.5	-8.7	NMB	60.9	8.5	-10.1

While the positive tropospheric HCHO column biases derived in our study are higher compared to the recent studies of Vigouroux et al. (2020) and De Smedt et al. (2021), the magnitude dependance is similar. We show here that clean/background satellite HCHO columns are larger than observations for all satellite products and transition to a low bias in highly polluted regions. Text describing this, and the table above was added as Table S3, was additional text was added to the updated manuscript in Sect. 3.2.3.

Line 594: I agree with the reasons for the poor HCHO correlation. Please add that they are also partly due to the low HCHO variability over the studied time and domain. A full year study would result in larger correlations.

We agree with the reviewer that the low correlations between the satellite and observed HCHO is primarily driven by the spatial variability in this study. This differs from many recent studies which use stationary point-source observations (Vigouroux et al., 2020; De Smedt et al., 2021) which primarily capture temporal variability in column HCHO retrievals. This may suggest that temporal

variability in HCHO is easier to retrieve from space compared to spatial variations which rely on input geophysical and a priori data sets (e.g., surface albedo, aerosol, a priori profiles, clouds) to accurately capture entire scenes variability of the specific variable. This has been expanded on in the updated manuscript.

High pollution case study: The added value of this section is not clear. As the paper is already long and detailed, I would suggest removing this section. If not removed, I then suggest to discuss the causes of higher NO2 columns and lower HCHO columns, such as surface temperature.

We agree with the reviewer and this section has been removed.

Common a priori sensitivity tests:

- It is not clear why the WRF-CMAQ profiles need to be scaled for the NASA OMI datasets, but not for the TROPOMI datasets.

The primary reason for differences between OMI and TROPOMI retrievals using the WRF-CMAQ a priori profiles is the difference between the shape factors derived from WRF-CMAQ and the a priori information used in OMI (GMI) and TROPOMI (TM5). The exact comparison of the shape factors produced by WRF-CMAQ, GMI, and TM5 is inhibited by the fact that the standard retrieval products of tropospheric NO_2 from OMI and TROPOMI do not provide this a priori profile information. Therefore, this hypothesis could not be tested in this study. The impact of higher spatial resolution model simulations when used as a priori information was described in the original version of the manuscript and compared to other studies seeing the same results.

- Figure 6: please explain in the legend what is the NASA OMI (scaled).

The figure caption explains this in the original manuscript. We have slightly updated it to now read: “The OMI FNR retrievals calculated with the scaled WRF-CMAQ profiles are identified as “scaled” in the figure panel titles.”.

- Comparing Table 2 and Table 4, I can only see an improvement for TROPOMI at 0.05° resolution. The added value of this section is not clear, given the uncertainties in the WRF-CMAQ profiles.

It should be noted when comparing Table 2 and Table 3 in the updated manuscript that various aspects of OMI retrievals were also improved when using the WRF-CMAQ shape factors for AMF calculations. This section has been rewritten to better emphasize aspects of the retrievals that were improved by the higher spatial resolution model a priori profiles. We wanted to include this section of the manuscript as there is currently large interest in the literature to use higher spatial resolution air quality model output to reprocess satellite retrievals. Therefore, these results will be important for others working on this.

Expected FNR information from TEMPO:

- What is the expected signal ratio of Tempo compared to TROPOMI for NO2 and HCHO? Can we expect an improvement of the HCHO noise?

In order to shorten the paper and provide more focus on the major results/conclusions, we have decided to remove this section of the paper. The authors feel that the TEMPO section does not fit well with the rest of the study and would be more appropriate in a different publication.

- It would be interesting to show the diurnal variation of NO2 and HCHO from the TEMPO simulations.

See our comment above about the removal of the synthetic TEMPO data section.

- Line 777 and figure 7b and 7c. Not clear if retrieved OMI and TROPOMI are shown (line 777) or only synthetic TEMPO data averaged at the different spatial resolutions. It should be possible to show real data for OMI and TROPOMI in 2020.

See our comment above about the removal of the synthetic TEMPO data section.

Conclusion

Line 831: Please comment on the spatial and temporal resolution allowed by the OMI datasets. This is important for trend studies.

We agree with the reviewer that this is an interesting problem for trend studies. However, defining exact temporal and spatial resolutions allowed by OMI or TROPOMI for these types of analysis is not the focus of our current study. In order to address this comment, 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 further address this comment, we added an entire section (Sect. 3.4.2) in the updated manuscript which describes the capabilities of OMI and TROPOMI to observe spatial and temporal variability of FNRs during LISTOS. Furthermore, the following text was added to this section discussing daily- versus monthly-averaging OMI FNR data: "Recent studies have shown that averaging OMI data (especially HCHO retrievals) for longer temporal periods can reduce the noise and uncertainty in this data product. For example, in the recent paper by Souri et al. (2022a), it was shown that unresolved errors in OMI HCHO can be reduced in monthly-averages compared to daily retrievals by ~33% while there was little improvement in uncertainty statistics of NO_2 retrievals from OMI. However, recent studies (e.g., Schroeder et al., 2017) have also shown that for trend studies, monthly-averaging column FNR data can mask FNR temporal gradients that exist within that period. This could hinder the results of trend studies of pollution conditions on O_3 exceedance days, and days of lower pollution, which is a primary purpose of using satellite column FNR data.".

Line 860. The statements made on the new version of the NASA OMI HCHO product appear to be optimistic. The SNR of the retrievals is primarily determined by the SNR of the instrument. Please be more cautious, especially since no publication can support the statements.

We agree with the reviewer that this comment could be viewed as too optimistic without data analysis to support it. We have removed it from the conclusion section.

Line 866-867: This does not seem so clear in the paper that "using the WRF-CMAQ-predicted a priori information, resulted in highly accurate retrievals of FNRs". All L2 products used in the study also results in median biases lower than 0.5 for FNRs.

The reviewer is correct, and this sentence has been updated to reflect that the systematic bias is similar to the operational satellite products. However, other aspects of the statistical analysis of the reprocessed OMI retrievals were improved such as the correlations and RMSE values. This is discussed in more detail in the conclusion section of the updated manuscript.

Line 871-872: This sentence is misleading. The need for accurate shape factors is not only for OMI retrievals. It should be even more important for TROPOMI and TEMPO because of their finer spatial resolution.

This sentence has been updated to include TROPOMI.

References

De Smedt, I., Pinardi, G., Vigouroux, C., Compernolle, S., Bais, A., Benavent, N., Boersma, F., Chan, K.-L., Donner, S., Eichmann, K.-U., Hedelt, P., Hendrick, F., Irie, H., Kumar, V., Lambert, J.-C., Langerock, B., Lerot, C., Liu, C., Loyola, D., Piters, A., Richter, A., Rivera Cárdenas, C., Romahn, F., Ryan, R. G., Sinha, V., Theys, N., Vlietinck, J., Wagner, T., Wang, T., Yu, H., and Van Roozendael, M.: Comparative assessment of TROPOMI and OMI formaldehyde observations and validation against MAX-DOAS network column measurements, Atmos. Chem. Phys., 21, 12561–12593, https://doi.org/10.5194/acp-21-12561-2021, 2021.

Judd, L. M., Al-Saadi, J. A., Szykman, J. J., Valin, L. C., Janz, S. J., Kowalewski, M. G., Eskes, H. J., Veefkind, J. P., Cede, A., Mueller, M., Gebetsberger, M., Swap, R., Pierce, R. B., Nowlan, C. R., Abad, G. G., Nehrir, A., and Williams, D.: Evaluating Sentinel-5P TROPOMI tropospheric NO_2 column densities with airborne and Pandora spectrometers near New York City and Long Island Sound, Atmos. Meas. Tech., 13, 6113–6140, https://doi.org/10.5194/amt-13-6113-2020, 2020.

Lamsal, L. N., Krotkov, N. A., Vasilkov, A., Marchenko, S., Qin, W., Yang, E.-S., Fasnacht, Z., Joiner, J., Choi, S., Haffner, D., Swartz, W. H., Fisher, B., and Bucsela, E.: Ozone Monitoring Instrument (OMI) Aura nitrogen dioxide standard product version 4.0 with improved surface and cloud treatments, Atmos. Meas. Tech., 14, 455–479, https://doi.org/10.5194/amt-14-455-2021, 2021.

Nowlan, C. R., Liu, X., Janz, S. J., Kowalewski, M. G., Chance, K., Follette-Cook, M. B., Fried, A., González Abad, G., Herman, J. R., Judd, L. M., Kwon, H.-A., Loughner, C. P., Pickering, K. E., Richter, D., Spinei, E., Walega, J., Weibring, P., and Weinheimer, A. J.: Nitrogen dioxide and formaldehyde measurements from the GEOstationary Coastal and Air Pollution Events (GEO-CAPE) Airborne Simulator over Houston, Texas, Atmos. Meas. Tech., 11, 5941–5964, https://doi.org/10.5194/amt-11-5941-2018, 2018.

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.

Vigouroux, C., Langerock, B., Bauer Aquino, C. A., Blumenstock, T., Cheng, Z., De Mazière, M., De Smedt, I., Grutter, M., Hannigan, J. W., Jones, N., Kivi, R., Loyola, D., Lutsch, E., Mahieu, E., Makarova, M., Metzger, J.-M., Morino, I., Murata, I., Nagahama, T., Notholt, J., Ortega, I., Palm, M., Pinardi, G., Röhling, A., Smale, D., Stremme, W., Strong, K., Sussmann, R., Té, Y., van Roozendael, M., Wang, P., and Winkler, H.: TROPOMI–Sentinel-5 Precursor formaldehyde validation using an extensive network of ground-based Fourier-transform infrared stations, Atmos. Meas. Tech., 13, 3751–3767, https://doi.org/10.5194/amt-13-3751-2020, 2020.

Zhao, X., Griffin, D., Fioletov, V., McLinden, C., Cede, A., Tiefengraber, M., Müller, M., Bognar, K., Strong, K., Boersma, F., Eskes, H., Davies, J., Ogyu, A., and Lee, S. C.: Assessment of the quality of TROPOMI high-spatial-resolution NO_2 data products in the Greater Toronto Area, Atmos. Meas. Tech., 13, 2131–2159, https://doi.org/10.5194/amt-13-2131-2020, 2020.

Zhu, L., González Abad, G., Nowlan, C. R., Chan Miller, C., Chance, K., Apel, E. C., DiGangi, J. P., Fried, A., Hanisco, T. F., Hornbrook, R. S., Hu, L., Kaiser, J., Keutsch, F. N., Permar, W., St. Clair, J. M., and Wolfe, G. M.: Validation of satellite formaldehyde (HCHO) retrievals using observations from 12 aircraft campaigns, Atmos. Chem. Phys., 20, 12329–12345, https://doi.org/10.5194/acp-20-12329-2020, 2020.