We have responded to each comment below. Our replies are in blue, and the revised manuscript text is written in bold.

Response to review 1

Qu et al. have studied the impact of top-down NO_x emission estimates derived from two OMI NO_2 satellite data sets (NASA SP v3 and DOMINO v2) on NO_2 and O3 simulations with the GEOS-Chem model. Previous work already showed (e.g. Verstraeten et al. [2015], studies by Miyazaki et al.) that O3 in the troposphere is generally better understood when NO_x emissions are derived from satellite NO_2 data than when taken from emission inventories.

Here, Qu et al. find substantial differences in the agreement of NO_2 and O3 simulations against independent measurements depending on whether data set NASA or data set DOMINO is used. This was to be expected given that it is well-known that the NASA and DOMINO datasets have considerable differences. A useful aspect of the study is that the authors now quantify the consequences of these differences, which is relevant because satellite data is increasingly used to improve model understanding of atmospheric composition.

What is disappointing however is that we do not learn much new. Simulations with the NASA emissions compare better to some metrics, and worse to others, but the authors do not explain why. This makes the manuscript a technical document, where it is left to the reader to figure out what emissions could work best for his/her particular purpose, without actual guidance on why that would be. The authors should do more to investigate why using one dataset leads to better agreement e.g. for surface O3 at remote sites, and the other for polluted sites. Aspects of spatial resolution, temporal representativeness, and vertical sensitivity should be taken into account when providing this guidance to potential users.

We appreciate the comments from the reviewer. We have modified the title, abstract, and the details in the manuscript accordingly to address these concerns. The title is now changed to "**Impacts of** global NO_x inversions **on NO_2 and ozone simulations**."

"**Abstract.** Tropospheric **NO_2 and** ozone simulations have large uncertainties, but their biases, seasonality and trends can be improved with **NO_2 assimilations**. We perform global top-down estimates of monthly NO_x emissions using two OMI NO_2 retrievals (NASAv3 and DOMINOv2) from 2005 to 2016 through a hybrid 4D-Var / mass balance inversion. **Discrepancy in NO_2 retrieval products is a major source of uncertainties in the top-down NO_x emission estimates.** The 12-year averages of regional NO_x budgets from the NASA posterior emissions are 37% to 53% smaller than the DOMINO posterior**. Consequently, the DOMINO posterior surface NO_2 simulations greatly reduced the negative biases in China (by 15%) and the US (by 22%)** compared to surface NO_2 measurements. **Posterior NO_x emissions show consistent trend over China, US, India, and Mexico constrained by the two retrievals. Emission trends are less robust over South America, Australia, Western Europe and Africa, where the two retrievals show less consistency.** NO_2 trends have more consistent decreases (by 26%) with the measurements (by 32%) in the US from 2006 to 2016 when using the NASA posterior. **The performance of posterior ozone simulations has spatial heterogeneities from region to region. On a global scale, ozone simulations using NASA-based emissions alleviates the**

double peak in the prior simulation of global ozone seasonality. The higher abundances of NO_2 from the DOMINO posterior increase the global background ozone concentrations and therefore reduce the negative biases more than the NASA posterior in the GEOS-Chem v12 simulations at remote sites. Compared to surface ozone measurements, posterior simulations have more consistent magnitude and interannual variations than the prior estimates, but the performance from the NASA-based and DOMINO-based emissions varies across ozone metrics. The current hard-constraints on NO_x diurnal variations and limited availability of remote sensing data hinder improvement of ozone diurnal variations from the assimilation, and therefore have mixed performance on improving different ozone metrics. Additional improvements in posterior NO_2 and ozone simulations require more precise and consistent NO_2 retrieval products, more accurate diurnal variations of NO_x and VOC emissions, and improved simulations of ozone chemistry and depositions."

From a data user perspective, this work quantifies how differences in NO_2 retrieval products propagate to the downstream estimates in top-down NO_x emissions and ozone simulations. The discrepancy found in this study is larger than uncertainties caused by data assimilation methods (4D-Var versus Kalman Filter) and chemical transport models [Koukouli et al., 2020], and is therefore a unique contribution of this work. Detailed investigation of the origin of differences in the NASA and KNMI NO_2 retrieval products goes beyond scope of this study. We do note however "**The GEOS-Chem NO_2 SCDs converted using scattering weight from the NASA product are larger than the SCDs calculated using the DOMINO scattering weight and the same GEOS-Chem VCDs (See Fig. S2). These can be explained by the use of different surface albedo and cloud product in the two retrievals."** (Added in Section 3)

Another criticism is that the chain of technicalities is very long and that the experiments are set-up in a sub-optimal manner (for example comparing 2.5 degree simulations of surface NO_2 to surface stations that are representative for much smaller domains).

Comparing NO_2 simulations at 2.5° with in-situ measurements is sub-optimal, but this is the highest resolution we can perform global 4D-Var assimilation using this model.

A major concern I have is with the lack of detail and clarity on how the adjoint incorporates the information from the satellite retrievals. From the manuscript I first suspected that monthly mean column NO_2 data was simply used to estimate the emissions, suggesting that the highly variable and non-linear vertical sensitivities of the retrievals have not been used to interface the model with the satellite data. There are various studies pointing out how crucial it is to account for the vertical sensitivity of the NO_2 retrievals, e.g. Miyazaki et al. [2017], Boersma et al. [2016] to name a few. Then I read the supplementary material and there the impression was given that at least the a priori profile shapes are made consistent between the NASA and DOMINO retrievals, but it remains unclear to what extent this has harmonized the data, and to what extent vertical sensitivities between the two datasets are still fundamentally different.

To clarify, we added the following sentences to Section 2.2:

"**We converted GEOS-Chem NO_2 VCD to SCD using scattering weight (NASA product) and averaging kernel (DOMINO and QA4ECV product) from the OMI retrievals and then compare GEOS-Chem SCD with SCD retrieved from OMI. A cost function is defined as the observation error weighted differences between simulated and retrieved NO_2 SCD, plus the prior error weighted departure of the emission scaling factors from the prior estimates. We minimize the cost function using the quasi-Newton L-BFGS-B gradient-based optimization technique [Byrd et al., 1995; Zhu et al., 1994], in which the gradient of the cost function with respect to the control parameter is calculated using the adjoint method. Details of the** assimilation of NO_2 slant column densities (SCDs), **how vertical sensitivities of satellite retrievals are accounted for**, and the hybrid 4D-Var / mass balance inversion of NO_x emissions **are described in** Qu et al. [2017]."

More detailed technicalities have been described in our previous publications cited in the manuscript and are therefore not the focus of this manuscript. The focus here is to apply this method for global NO_x inversion, evaluate the impact of different retrieval products on top-down emission estimates, and how the changes in NO_x emissions affect ozone simulations. Therefore, we did not repeat all the technical details that can be found in the cited publications. Please see our detailed responses below for all the concerns raised by the reviewer.

Specific comments

P2, L42-43: the formation depends not only on the local NO_x and VOC concentrations, but also on the radiative regime in which these occur.

Changed to "Ozone formation and trends depend nonlinearly on the local relative abundances of NO_x and VOCs **and the radiative regime in which these occur**."

P2, L65: different → differ

Modified as suggested.

P3, L72: import →importer

Modified as suggested.

P3, L78-81: Zhang et al. [2008] and Verstraeten et al. [2015] already showed that through optimizing NO_x emissions in China, the simulated O3 over the Pacific and over the western US indeed improved.

We changed this sentence to:

"Optimization of NO_x emissions in the upwind regions can improve remote ozone simulations in downwind regions after transport of intercontinental pollution plumes from the free troposphere to the surface **[Zhang et al., 2008; Verstraeten et al., 2015]**."

Section 2.1 It is unclear in this manuscript how the adjoint accounts for (a) vertical sensitivity of the satellite retrievals, and (b) the diurnal cycle of NO_x emissions. These aspects are important enough to describe in the manuscript, for (a) useful information is provided in the supplement but it is not clear whether the replacing of the a priori profiles by GEOS-Chem prior profiles was also applied in the research to estimate the emissions. The authors should clarify this in section 2, and also briefly quantify to what extent the differences in prior simulations have been minimized by this approach.

Many of these aspects have been described in details in a previous publication cited in Section 2.2 (Qu et al. 2017), so we do not repeat the same information in this manuscript. To clarify, we added a brief summary of our inversion in Section 2.2, see our response above.

For the reviewer's information, The comparison of SCDs ($VCD_{GC}AMF_{GC} - SCD_{OMI}$) is theoretically equivalent to comparisons of VCDs ($VCD_{GC} - \frac{SCD_{OMI}}{AMF_{GC}}$). These have been described in Qu et al. [2017], pasted below:

"In all of our simulations, we calculate the air mass factor (AMF) for GEOS-Chem simulated NO_2 columns (AMF_{GC}) following Equations 1 to Equation 4 in Bucsela et al. [2013]. Here, AMF_{GC} is expressed as the ratio of the sum of slant sub-columns in the troposphere (S) to the sum of vertical sub-columns in the troposphere (V):

$$AMF_{GC}(i,j) = \frac{S}{V}$$

where

$$S = \sum_{l\ in\ the\ troposphere} MR(i,j,l)(P(i,j,l) - P(i,j,l+1))SCW_{OMI}(i,j,l)$$

$$V = \sum_{l\ in\ the\ troposphere} MR(i,j,l)(P(i,j,l) - P(i,j,l+1))$$

Here, MR is the mixing ratio of NO_2, P is the pressure at the center of the GEOS-Chem grid, SCW_{OMI} is the scattering weight linearly interpolated from the OMI product to GEOS-Chem grid using the scattering weight pressure from the Level 2 product and pressure at the center of each model grid cell, with application of temperature correction following Equation 4 of Buscela et al. [2013]. AMF_{GC} is then used for conversion of GEOS-Chem NO_2 vertical column densities to SCDs, which are directly comparable to SCDs retrieved from OMI,

$$SCD_{GC}(i,j) = AMF_{GC}(i,j) \sum_{l\ in\ the\ troposphere} (c(i,j,l) \times h(i,j,l))$$

where c is simulated NO_2 concentration [molecules cm^{-3}] and h is the height of the box."

We added the following sentence to the first paragraph of Section 3:

"The cost function has reduced by 6% - 29% in the monthly inversion."

For (b), some info is given but only late in the game (P7: The diurnal variations of NO_x emission are constrained to be those of the prior emissions), and we do not learn what the diurnal cycle is in the first place. Please revise section 2 thoroughly with this in mind.

We added the following sentence to Section 2.1:

"The diurnal variation of NO_x emissions is derived from EDGAR hourly variations (http://wiki.seas.harvard.edu/geos-chem/index.php/Scale_factors_for_anthropogenic_emissions#Diurnal_Variation) and is not optimized in the inversion."

Then I have other questions:

- how does the adjoint approach account for other relevant aspects of data assimilation?

Details of our 4D-Var inversion are in Qu et al. [2017]. In brief, we define a cost function as described in Section 3 of Qu et al. [2017]. Then, "We minimize the cost function using the quasi-Newton L-BFGS-B gradient-based optimization technique [Byrd et al., 1995; Zhuetal., 1994], in which the gradient of the cost function $J(\boldsymbol{\sigma})$ with respect to the control parameter $\boldsymbol{\sigma}$ is calculated using the adjoint method. The adjoint model is driven by a forcing term, which is the error weighted difference between predicted and simulated NO_2 slant columns. Inversions are considered to have converged when the cost function decreases by less than 1% in three consecutive iterations."

- how is the OMI data averaged spatially to the grid of GEOS-Chem, and how are superobservation errors incorporated?

We did not average OMI data or use super-observations. Instead, we assimilate each OMI retrieval separately and compare it with GEOS-Chem simulations at the corresponding hour, with corresponding averaging kernel applied. Please see Section 3 in Qu et al. [2017] for more details, which state:

"Slant column densities from OMI at each observation time and site are used to constrain monthly anthropogenic NO_x emissions. The observation error covariance matrix, $\mathbf{S_{obs}}$, is assumed to be diagonal. Absolute uncertainties of these diagonal values are read from NASA $OMNO_2$ L2 products for each individual OMI observation. On average, the tropospheric slant column uncertainty of OMI is estimated to be $\sim 0.7 \times 10^{15}$ molecules cm^{-2} [Boersma et al., 2008; Castellanos and Boersma, 2012]. To reduce the influence of observations below the OMI detection limit, which mainly occur in remote locations, we conservatively assume an absolute uncertainty of 1.0×10^{15}molecules cm^{-2}, and we add this value to $\mathbf{S}_{obs}$."

- did the authors only use the mostly cloud-free OMI retrievals?

Yes, only retrievals with cloud fraction less than 0.2 are used. This has been stated in section 2.2 of this manuscript:

"We screen all OMI NO_2 retrievals using data quality flags and by the criteria of positive tropospheric column, cloud fraction < 0.2, solar zenith angle < 75°, and viewing zenith angle < 65°."

Section 2.2: OMI is suffering from the so-called row anomaly, which was absent until mid-2007, and then became gradually more important. How did the authors ensure that the growing impact of the row anomaly did not unduly affect their trends in NO_x emissions?

The OMI data affected by row anomaly are filtered out using the quality flag. We added the following sentences to section 2.2:

"We excluded all retrievals that are affected by row anomaly."

We have tested the differences between annual mean OMI NO_2 column densities without data filling after excluding pixels affected by row anomaly and when filling missing data by linearly interpolating column densities from adjacent years in Qu et al. [2017]; we found the filling to impact annual mean SCDs by less than 10% for all regions shown in Figure 8 of Qu et al. [2017]. Differences in these two SCDs for all studied years are less than 1% in mainland China.

Another approach to mitigate inconsistent sampling of the data is to follow Duncan et al. [2013] and consider the trend in NO_2 columns from only rows 10 to 23 of the NASA standard product, which are unaffected by the row anomaly throughout the period. These are shown in the grey lines in Figure 8 of Qu et al. [2017]. Please also note that even though we are using the same rows each year, this doesn't necessarily mean that the number of observations is the same after screening according to our other filtering criteria, nor does it mean the same geographical locations are observed throughout the period. The correlation of this dataset with OMI data from the standard NASA product in all rows is >0.75 in most regions.

Though we recognize the benefits of using a consistent number of observations to analyze the trend of NO_2 columns alone, this is not necessarily the case for a Bayesian inversion of NO_x emissions. The inversion is forced by the residual model error summed over all available observations; fewer observations in some years or locations will thus naturally result in greater dependence on the prior emissions. If we exclude observations to maintain consistency in the rows used, emissions in many grid cells do not get updated due to lack of observations (see Fig. R1). This would lead to spatial trends in posterior emissions that could have been avoided if using all available observations (after data screening).

We think the two approaches to invert NO_x emissions, maintaining consistency in rows used or not, both have their pros and cons. Since the goal of this work is to derive top-down emissions, which would benefit from broader observation coverage (in the example of January 2006 below, we would not be able to get posterior emissions for regions covered in white if eliminating those rows affected by row anomaly throughout) and the trend of NO_2 columns between these two

does not differ much, we chose to use all observations available after data selection.

Only row 10-23 All observations after filtering

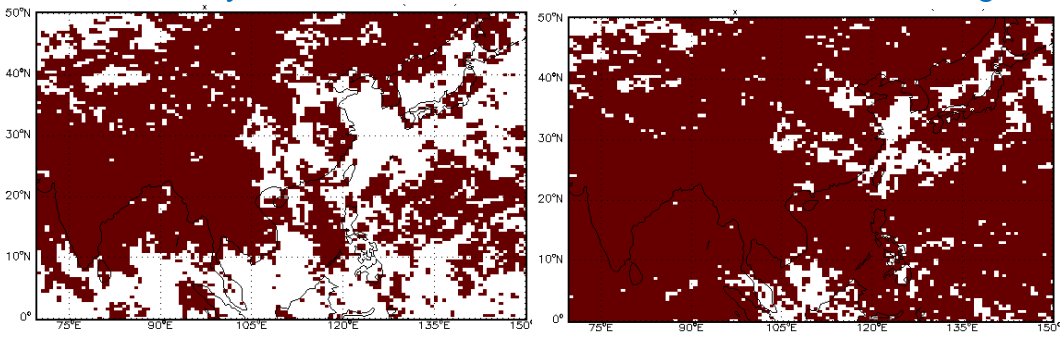


Figure R1. Data coverage in January, 2006, using only rows 10 to 23 (left) and all rows (right), where, red color stands for grid cells that have at least one observation during the month.

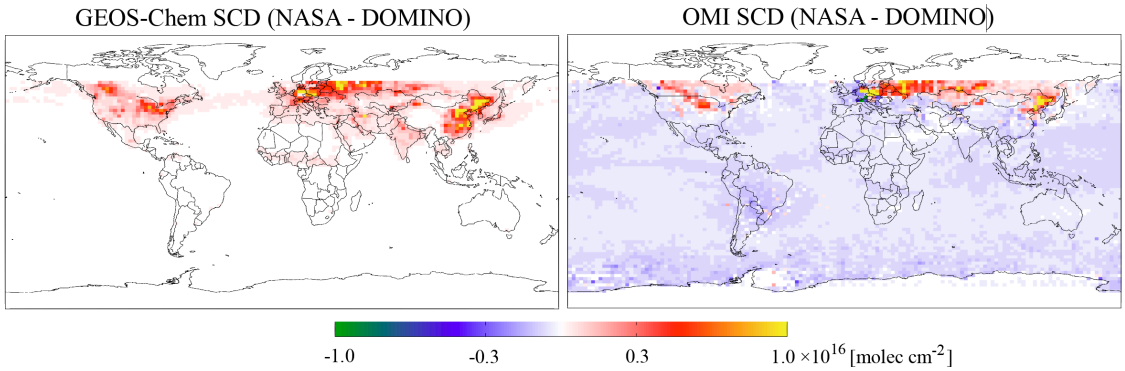
Section 2.3: it remains unclear what type of surface station was used for the GEOSChem surface evaluation. Using urban background and regional stations seems appropriate to evaluate the large GC grid cells, but urban street stations should be excluded.

We checked the monitoring site lists and a document defining the site category (http://www.bjmemc.com.cn/xgzs_getOneInfo.action?infoID=1661). None of the sites included in this study was listed as roadway sites. We added the following sentence to Section 2.3:

"The city monitoring sites included in the analysis represent either urban background or the averaged pollutant concentrations over the city."

P5, L152-154: what explains the OMI-driven differences between the posterior NO_x emissions, differences in tropospheric slant columns or in the AMFs? Presumably the latter, but since the a priori profile differences have been "minimized", the differences must be in the assumptions on surface albedo and clouds. It would be best if the authors could shed more light on how the scattering weights or averaging kernels are different between the OMI NO_2 retrievals. Please clarify.

We added a new Figure S2 to the supporting information:



“

Figure S2. Differences in tropospheric NO_2 SCDs between the NASA and the DOMINO products in January 2010. The differences in GEOS-Chem SCDs (left figure) are calculated by converting the same GEOS-Chem VCD using scattering weight and averaging kernel from the two products. In the right figure, AMFs provided by the two products are applied to their corresponding VCDs to calculate the differences in SCDs. ”

We also added the following sentences to the cited paragraph:

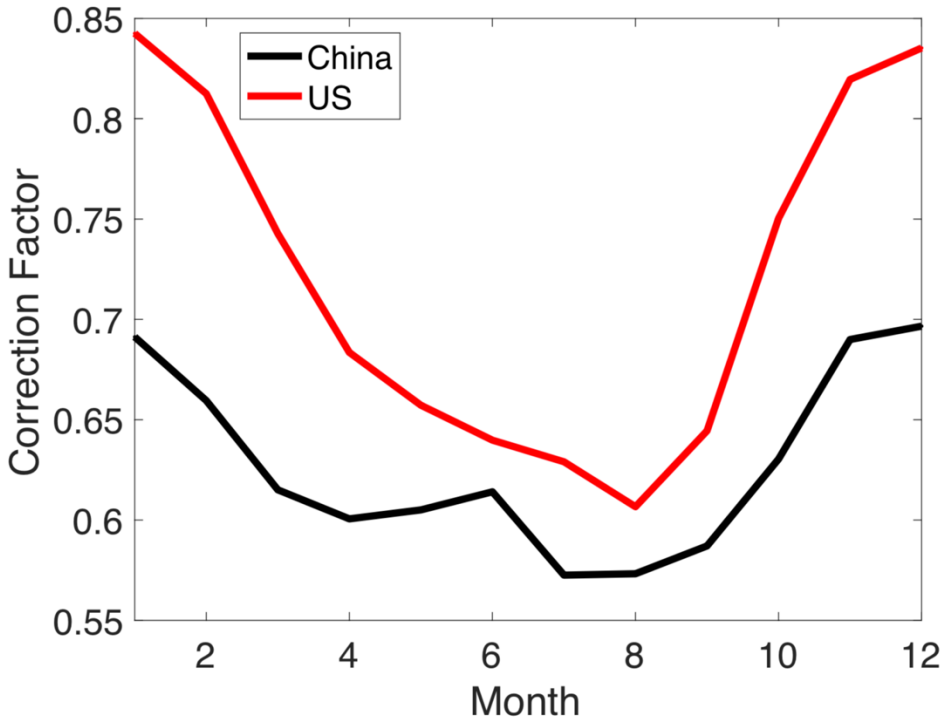
“The GEOS-Chem NO_2 SCDs converted using scattering weight from the NASA product are larger than the SCDs calculated using the DOMINO scattering weight and the same GEOS-Chem VCDs (See Fig. S2). These can be explained by the use of different surface albedo and cloud product in the two retrievals. The retrieved NO_2 SCDs from the NASA product are mostly smaller than the DOMINO retrieval except for some regions between 40°N – 60°N in January 2010. The smaller magnitude in OMI SCD and the larger magnitude in GEOS-Chem SCD using the NASA scattering weight lead to smaller magnitude of posterior NO_x emissions than inversions from the DOMINO product.”

P6, L173-174: the statement that “NO_2 column simulations at 2◦ × 2.5◦ in this study are likely to be underestimated and lead to high biases of posterior NO_x emissions to match satellite NO_2column concentrations” needs more evidence. The hypothesis that instant dilution leads to too much OH (by Valin et al. [2011]) may be valid for isolated NO_x sources in otherwise pristine areas, but instant dilution of NO_x emissions situated in high-background NO_2 regions such as the eastern US or western Europe is probably of less concern.

We removed the cited sentences.

P6, L193: what is the magnitude of the correction factors over China and the US? How do they vary by season?

We added the following figure in the SI:



"Figure S3. Seasonal variation of the NO_2 correction factors in China (black) and the US (red) calculated following Lamsal et al. [2008]. "

We added the following sentences to the cited paragraph:

"The correction factors are generally higher in the US than in China, but have similar seasonality (see Fig. S3)."

P7, L195-199: this part is rather inconclusive. The GEOS-Chem simulations have been corrected for resolution (an increase) and surface measurements have been corrected down for molybdenum interference, and still GEOS-Chem with posterior emissions is biased low by 20%-50%. What explains the persistent low bias?

We added the following sentences to this paragraph:

"These remaining negative biases reflect the unrepresentativeness of 0.1° pseudo measurements for real point measurements for resolution bias correction, comparison of NO_2 concentrations averaged over 2°×2.5° simulation to limited measurements, the underestimates of NO_2 retrievals using coarse resolution a priori, and the inability of data assimilation to increase emissions at grid cell where NO_2 retrievals are below the detection limit of OMI."

Also, we do not expect the posterior simulations to be completely unbiased given the potential biases from model and satellite retrieval.

P7, L224-225: OMI measurements frequently miss the high values of NO_2 column densities that occur before or after its overpassing time. OMI was never designed to measure NO_2 before or after its overpass time, so to say that OMI misses these high values is misleading. Please rephrase.

We changed the sentence to:

"The daily NO_2 column densities from OMI are also underestimated compared to the diurnally varying ground-based retrievals [Herman et al., 2019]."

P7, L226: twice-per-day constraints on NO_x emissions have been achieved in earlier studies based on SCIAMACHY + OMI (Boersma et al. [2008], GOME-2 + OMI [Lin et al., 2011], including via sophisticated assimilation schemes [Miyazaki et al., 2017].

We changed the sentence to "**Assimilating NO_2 observations from instruments overpassing at different times of the day [e.g., Boersma et al., 2008; Lin et al., 2010; Miyazaki et al., 2017] and using** hourly constraints from the geostationary satellite data (e.g., Geo-stationary Environmental Monitoring Spectrometer (GEMS), Tropospheric Emissions: Monitoring of Pollution (TEMPO) [Zoogman et al., 2017] and Sentinel-4) have the potential to improve simulations of ozone diurnal variations and different ozone metrics, **although the ratio of NO_2 column densities from satellites that overpass in the morning and afternoon are generally lower than the same ratio from surface measurements [Penn and Holloway, 2020]."**

P8, L237: the June peak in NO_2 over China can be easily traced back to crop residu burning in that month – e.g. Stavrakou et al. [2016].

We added the following sentence:

"The June peak in China has been explained by the crop residual burning [Stavrakou et al., 2016]."

P8, L238-240: can you explain more why the DOMINO product would be more sensitive to soil NO_x emissions? It's not because of the different a priori profiles assumed in the NASA and DOMINO retrievals?

As the reasons are not entirely clear, we changed the cited sentence to:

"The peak of the DOMINO posterior NO_x emissions in the United States and Mexico shifted earlier in the year to June and July compared to the prior and NASA posterior emissions, similar to the results from Miyazaki et al. [2017]. **The peak in DOMINO posterior emissions corresponds to the time of high soil NO_x emissions, which are reported to be underestimated in high-temperature agricultural systems in the bottom-up inventory [Oikawa et al., 2015; Miyazaki et al., 2017]."**

P8, L243-244: please see my previous comment. The authors seem to know something very interesting here, but they don't show it. Is there any evidence that one retrieval would be more

sensitive to NO_x sources than the other? That would be extremely relevant to know more about. Since the satellite measurements are identical for the NASA and OMI retrievals, it must have to do with AMF differences , see e.g. Lorente et al. [2017]. But what drives the apparent difference in sensitivity – albedo, cloud fraction, cloud pressure?

The two retrievals have the same spectrum but the retrieved tropospheric SCDs are not exactly the same (for instance, the two products use different stratosphere-troposphere separation), see our previous response and the new figure S2. All of the factors the reviewer mentioned here are different between the two products. It is hard for us to pinpoint which of the albedo, cloud fraction, or cloud pressure drives the sensitivity without running the radiative transfer model and performing the retrieval ourselves, which is beyond the scope of this study.

We changed the cited sentence to:

"**These retrieval products have similar number of observations and spatial distributions of observation densities after the filtering.** The different seasonal variations in the posterior NO_x emissions may reflect the **AMF structural uncertainties when the retrieved NO_2 column densities use different ancillary data [Lorente et al., 2017]. For instance, the GEOS-Chem NO_2 SCDs converted using the scattering weight from the NASA product have larger seasonal variations than the SCDs converted using the DOMINO averaging kernel in the US, reflecting the different seasonal variations of vertical sensitivities from the two retrievals.**"

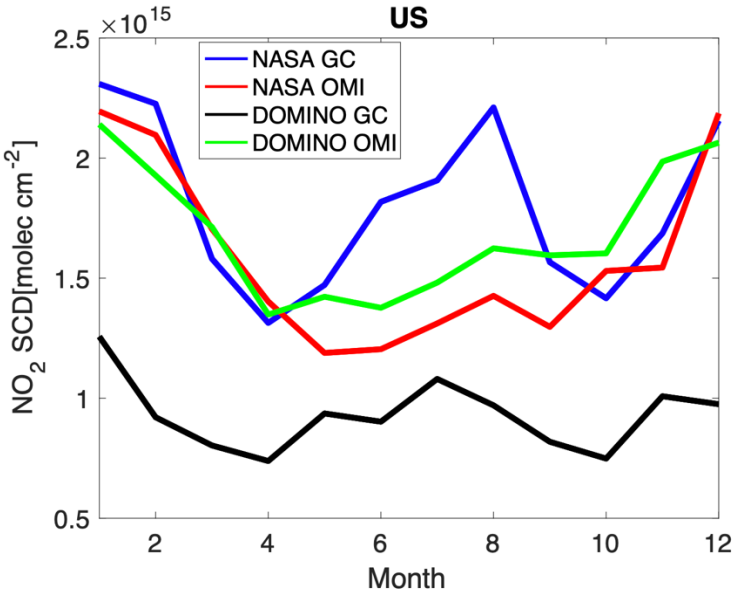


Figure R2. Seasonal variations of OMI NO_2 SCDs from NASA (red) and DOMINO (green) retrievals, and the GEOS-Chem simulated NO_2 SCDs using scattering weight from the NASA (blue) and the DOMINO (black) products.

P8, L246-256: Figure 5 – the daytime O3 simulations in China all seem strongly low biased relative to the observations. The other ozone metrics in China and all in the US match much better. Why is this?

Thanks for pointing this out. There was a bug in processing daytime ozone in China, which is fixed now. Please see the revised Figure 5 below.

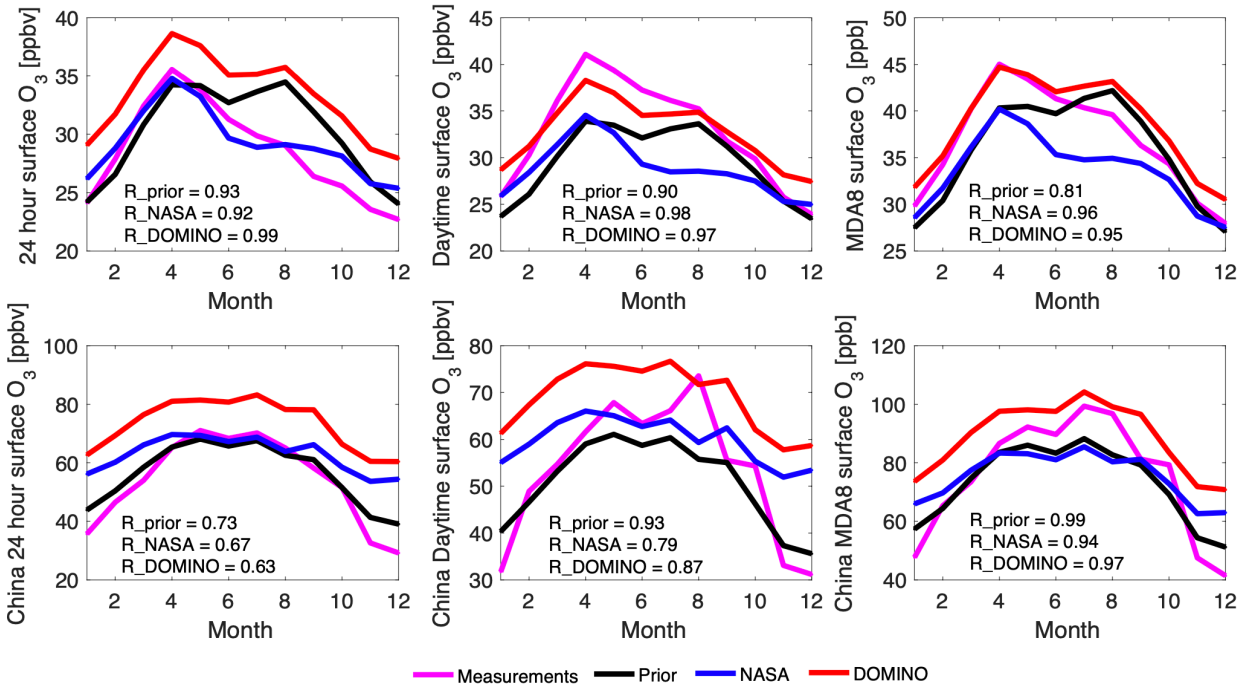


Figure 5. Seasonality of surface ozone concentration at 2 meters in 2010 compared with TOAR (top) and in 2015 compared with CNEMC (bottom). Surface measurements are shown in magenta lines. Simulations are performed using GCv12 with NO_x emissions from CEDS (black line), NASA posterior (blue line) and DOMINO posterior (red line).

P9, L271-272: "also not reflected"?

Changed to "not reflected"

P9, L276: no reduction of NO_x emissions in Europe? This is strange – NO_2 tropopsheric columns are decreasing over Europe, and Miyazaki et al. [2017] showed reductions in for NO_x emissions. Overall, Figure 6 looks very odd to me. DOMINO NO_2 columns are 40% higher than NASA, but the NO_x emissions inferred from DOMINO are more than 40% higher than the emissions inferred with NASA (L278-281). Also, Miyazaki et al. [2017] (Figure 9) still find reductions in NO_x emissions over Europe between 2005 and 2014 based on the same DOMINO data, so how can you find increases? Please clarify.

We do not expect the relative differences in the direct comparison of NO_2 column densities from the two OMI products to have similar magnitude with the differences in their posterior emissions. As shown in the newly added Figure S2, the adjustment in NO_x emissions are determined not only by the differences in NO_2 SCDs from OMI retrievals but also by the GEOS-Chem SCDs after applying scattering weight / averaging kernel (equivalent to converting OMI SCD to VCD using a new AMF based on GEOS-Chem profile and compare with GEOS-Chem

simulated VCD). The smaller magnitude in OMI SCD and the larger magnitude in GEOS-Chem SCD using the NASA scattering weight leads to even smaller magnitude of posterior NO_x emissions than the posterior constrained by the DOMINO product.

As for the posterior emissions in Europe, the result from Miyazaki et al. (screenshot shown in the left panel of Figure R3) shows large fluctuations around 0 throughout 2005 and 2014, and it is hard to say there is a decreasing trend from their Figure 9. The relative change from 2005 to 2014 in this study, shown in the right panel of Figure R2, is also negative (-1.3%), consistent with results in Miyazaki et al. [2017]. The slight upward fluctuation of posterior NO_x emissions in this study happened after 2014, which is not included in the time range of Miyazaki et al. [2017].

We changed the cited sentence to:

"In Western Europe and Africa, posterior NO_x emissions fluctuate and do not have a significant consistent trend from the two inversions."

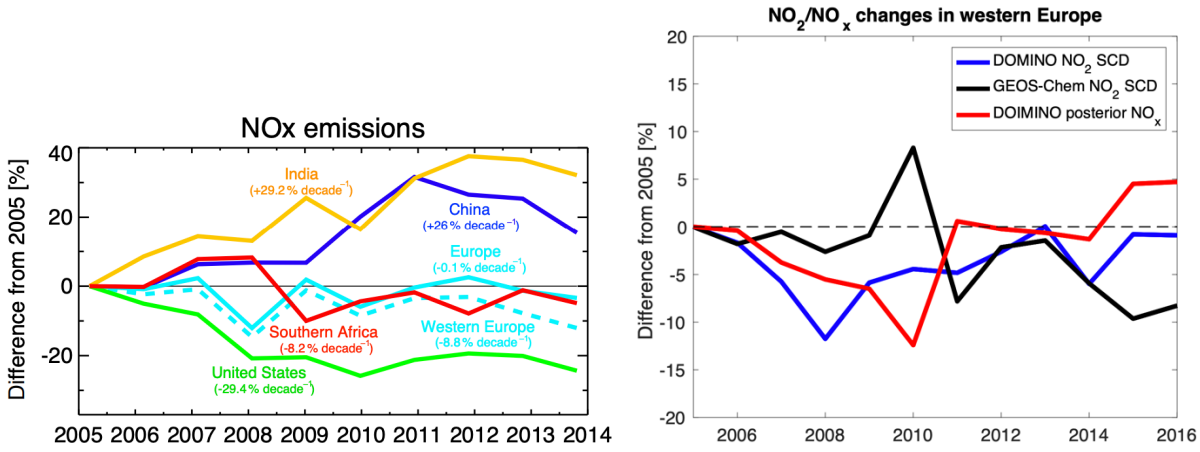


Figure R3. Relative changes in NO_x from Miyazaki et al. [2017] (left) and this study (right).

P10, L295-297: I'm missing an explanation or hypothesis why NO_x emissions from one dataset would do better than the other for different ozone metrics.

We added the following sentences to the cited paragraph:

"The different performance of NO_x emission datasets for different ozone metrics is a consequence of the hard constraint on NO_2 diurnal variations within the assimilation (and the lack of sufficient observations to constrain this). This can lead to better agreement of mean ozone concentration with measurements over particular hours but worse mean concentrations averaged over other hours."

P10, L304 and L315: please clarify how the impact of meteorology and non-NO_x sources on O3 changes was evaluated.

We changed the original sentence on L304 to "The trends of simulated annual MDA8 ozone concentrations are correlated with impacts from meteorology and non-NO_x sources **based on simulations (shown as green lines) that use the same anthropogenic NO_x emissions for all years and simulations that use interannually varied anthropogenic NO_x emissions,** leading to …"

We added the following sentences to the original sentence on L315:

"…as well as meteorology and non-NO_x sources. **The second trend is calculated through simulations that use constant NO_x emissions throughout the studied years. It has similar trend from GCv12 and GCadj as shown in the green lines in Fig. 9. The trend caused by NO_x emissions is obtained by subtracting the second trend from the ozone trend simulated using NO_x emissions at each corresponding year.** The ozone trends…"

L306-307: "The trends of simulated MDA8 ozone are similar when using the NASA and the DOMINO posterior NO_x emissions as inputs" – yes, but please also explain why the magnitude of the NASA-derived MDA8 O3 levels are biased high then.

The blue and red colors in this figure now represent ozone simulations from different models. The differences from NO_2 retrievals are now represented in the error bars. The NASA-derived MDA8 ozone are actually lower than the DOMINO-derived one. We added the following sentences to this paragraph:

"The DOMINO-derived MDA8 ozone concentrations are higher than the NASA-derived ones in all studied regions, represented by the upper and lower limit of the error bars respectively. GCv12 simulated ozone concentrations are smaller than simulations from GC-adj, especially over relatively less polluted regions, consistent with the inclusion of halogen chemistry in GCv12, which depleted ozone."

P11, L332-333: the prior simulated O3 profiles in Figure 10 agree much better with the O3 sondes between 800-400 hPa than the assimilated profiles. I don't understand why that is, since the effect of the updated NO_x emissions should be mostly felt in the lower 2 kms of the atmosphere. Or is this the impact of changes in background O3 in response to changing Asian emissions?

The reviewer must have been mistaken when considering this figure, as it is not true that all nor even most prior simulations (black dotted and black solid lines) agree better with ozone sondes (magenta solid) in Figure 10. In the 800-400 hPa range, the figure shows the GC-adj simulation using the DOMINO posterior NO_x emissions (dashed red) is almost always the closest to the sonde data. More detailed statistics of ozone profiles between 700-900 hPa, where ozone is mainly impacted by Asian emissions (Figure S8), show that the posterior O3 from GC-adj have smaller NMB and NMSE than the prior at 4 of 6 sites.

We added the following sentence to the title of Figure 10:

"**The six sites are over remote regions and are used to evaluate the intercontinental transport of ozone.**"

P13, L394-395: one important difference between this research and the work done by Miyazaki in a number of papers, is that the latter assimilates also other species relevant for NO_x inversions and O3 simulations (e.g. CO, HNO3, SO2). It would be interesting to also discuss to what extent these additional constraints can help explain the "remaining differences between simulated and measured ozone".

We added the following sentence:

"Assimilation of multiple species (e.g, ozone, CO, HNO_3 and SO_2) together with NO_2 may improve posterior ozone simulations, but the performance of posterior simulations may depend on the chemical transport model, as shown in Miyazaki et al. [2020], where the GEOS-Chem adjoint model v35 shows mixed performance in correcting the bias between ozonesonde and posterior simulations between 850-500 hPa at different latitude band."

P13, L398-400: the statement "Both OMI NO_2 retrievals employed in this study use NO_2 vertical shape factors from coarse resolution simulations, and therefore are biased low compared to in-situ measurements [Goldberg et al., 2017]." Brought up the question (again) whether both OMI NO_2 retrievals are at least consistent now in their use of the same coarse-resolution vertical shape factors (i.e. those from GEOS-Chem).

Yes, we converted GEOS-Chem VCD to SCD using scattering weight from these two products for comparison, but mathematically they are equivalent to replacing the shape factor with the same GEOS-Chem one. Please see more detail in our response to previous comment.

P13, L401: "retrievals also have not explicitly accounted for the aerosol optical effects, which are demonstrated to degrade the accuracy of NO_2 column concentrations". This is an overstatement. Only when AOD is very high (>0.5-1.0) there are indications that implicit corrections break down. Even in Liu et al. [2019] accounting for AOD did not solve the low bias in tropospheric NO_2 which was not apparent in the DOMINO scheme without an explicit aerosol correction.

We change this sentence to "which are demonstrated to degrade the accuracy of NO_2 column concentrations **when AOD is very high**".