Souri and co-authors present an analysis of lockdown-induced changes in NO2, HCHO, and O3 over Europe based on a data assimilation approach involving TROPOMI measurements (NO2, HCHO) and the WRF-CMAQ model. An advantage of the approach is that it explicitly accounts for meteorological influences and so forth in assessing the causes of AQ changes during the COVID period.

The paper topic is suitable for ACP and will make a useful contribution to the literature. There are some methodological and science comments and questions that I feel should be addressed before publication; these are listed below. In a number of cases things are described in a confusing way and need to be clarified. Finally, in many places the writing can be clearer or is overly wordy. For example, many of the paragraphs are about a page long and cover multiple topics, which really does not help communication. Once these issues are addressed I recommend publication in ACP.

We thank you taking the time to provide such detailed feedback.

==========

Science comments. Numbering refers to line numbers.

==========

1. The abstract is very long (almost 500 words!) which partly defeats the point of an abstract. I suggest reducing it by approximately half.

We have shortened the abstract.

56. "Earth's atmosphere has exponentially become more polluted during previous decades". Too vague / sweeping. What do you mean by "previous decades"? Some parts of the world have become significantly less polluted (for PM and ozone) over the last 2-3 decades. Elsewhere I do not think the word "exponentially" is necessarily accurate.

We agree that we wrote this part a bit too general. We have changed it to:

"Earth's atmosphere has substantially become more polluted since the industrial era in comparison to its original environmental condition [Li and Lin, 2015], thus any abrupt hiatus in anthropogenic (man-made) emissions …"

We simply want to emphasize on the magnitude of anthropogenic emissions being so large, which is essentially the main motivation of this study.

78. "whereas the concentrations of several secondarily formed compounds such as ozone increased due to emissions and/or meteorology". This is not universally true; e.g., see Bekbulat et al., https://doi.org/10.1016/j.scitotenv.2020.144693.

Thanks. We added the paper and wrote: "whereas the concentrations of several secondarily formed compounds such as ozone behaved in non-linear ways due to emissions and/or meteorology."

767. For reports, please list references in a way that readers can readily access them, e.g. with a doi or persistent URL.

Added.

138-139. RMSE already has a definition; what you are reporting here is not the RMSE.

We changed it to "*dispersion*" in the text and the table. We defined the dispersion ("as half of the 68% interpercentile").

148-150. "There are challenges…" Unclear what you mean here. Please rewrite for clarity.

We made a new paragraph and rewrote this part:

"Directly incorporating these numbers into an inversion model is challenging, mainly because of spatiotemporal variability in the satellite errors. Ideally, the relationship between errors and retrieval inputs (e.g., albedo, scene radiance, profiles, etc.) would be used as an additional cost function in the inversion, commonly known as variational bias correction [e.g., Auligné et al., 2007]. In the absence of such relationships, we use the biases reported in the validation studies. "

153. "we uniformly scale up NO2 pixels by 25% based on the low bias determined by Verhoelst et al. [2021] while considering the potential reduction in the bias through the use of higher spatial resolution trace gas a priori profiles." Not clear what is meant here by "while considering". How was this considered? Do you mean it was considered by choosing 25% rather than the median value of 37% reported by Verhoelst? Or is something different being implied here? Please clarify.

Please see the next comment.

153. The choice of a 25% bias correction for NO2 seems a bit arbitrary. As I understand it, the argument being presented is as follows: "the bias was reported previously to vary from -23 to -51%, with a median of -37%. But the use of higher-resolution shape factors here should reduce the bias. So, we use +25%." I agree that higher-resolution shape factors will reduce the bias, but there is no quantification of that effect here, so the 25% value seems to be pulled out of a hat. There is also the fact that the TROPOMI bias was shown previously to vary between rural and polluted environments, but this is not accounted for here. Overall, there needs to be either a more rigorous justification for the bias correction being employed, and/or some sensitivity analysis to quantify the degree to which this assumption affects the results.

Thanks for your detailed feedback. We added the justification:

In the case of NO_2, we uniformly scale up the satellite tropospheric columns by 25%. This bias estimate is derived by first assuming a 37% low bias in the columns over polluted regions as reported by Verhoelst et al. [2021]. In turn, this low bias can be mitigated somewhat by the application of high spatial resolution profiles in the air mass factor calculation, such as the ones used in this study. Table 1 summarizes the results from several TROPOMI validation studies at specific locations that calculated NO_2 using model profiles with higher spatial resolution than the operational TROPOMI ($1^o \times 1^o$) profiles (see Table 1 columns "Modification" and "Modified Bias"). In these studies, modified columns show increases ranging from 0 - 25%. Based on these results, we assume a low bias of 37% can be mitigated by ~12% through the use of high spatial resolution profiles, for a resulting total low bias of 25%. This bias is likely not valid over pristine areas, where validation studies show lower biases in TROPOMI NO_2 [Verhoelst et al., 2021, Wang et al., 2020, Zhao et al., 2020]; nonetheless, we previously observed in Souri et al. [2020a] that the low signal-to-noise ratios of those column amounts resulted in small changes in the top-down emissions.

We also elaborated the error characterization:

We assume the errors of observations originate from two main sources: i) the precision error provided with the data ($e_{precision}$) and ii) a fixed error estimated from comparisons to in-situ measurements (e_{const}). Mathematically, the final error is:

$$e_O^2 = e_{const}^2 + \frac{1}{n^2}\sum_{i=1}^{n} e_{precision,i}^2 \quad (1)$$

where n is the number of samples for a given grid and e_{const} equals to 1.1×10^{15} molec/cm^2 ($<6\times10^{15}$ molec/cm^2) in clean regions and 3.5×10^{15} molec/cm^2 ($>=6\times10^{15}$ molec/cm^2) in moderately to highly polluted regions based on the wide ranges reported in Verhoelst et al. [2021] ($3\text{-}14\times10^{15}$ molec/cm^2 for moderately to highly polluted region).

183-187. A similar comment applies to HCHO. I appreciate that the authors pay close attention to uncertainty and bias in the satellite data. But in the end the employed corrections are chosen a bit haphazardly from the range of reported biases. How can this choice be better justified, or if it is necessarily a little arbitrary, how can the impact of that assumption be quantified?

Those numbers are estimated based on the Figure 3 in Vigouroux et al. [2020]. Do we think this is the best way of accounting for the errors? No. In fact, our former studies exclusively focusing on AQ campaigns validated the measurements with respect to independent measurements; similar to what we currently showed in terms of MODIS /AERONET AOD.

There are two ways to improve the error characterization: i) the number of FTIR, Pandora, and MAX-DOAS observations would grow and they become permanent stations and publicly available (this is not a complaint; it is an utopia every quantitative study dealing with satellite is dreaming of) and/or ii) a more systematic approach would be taken to parametrize the errors in the validation studies; this can be done by establishing a relationship between errors and the inputs used for the retrievals. Say, one might find varying biases by changing cloud fractions such that you can reproduce the bias given a defined cloud fraction. Ultimately, a variational bias correction method [Auligné et al., 2007, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.56] can be adopted to engage those relationships (along with their uncertainty) in the inversion framework.

We modified this part: "We assume the constant term of errors (e_{const}) to be equal to 4% of HCHO total columns based on Vigouroux et al. [2020]."

156. "we set the RMSE to 1.1x10^15 molec/cm2 in clear regions and 3.5x10^15 in moderately to highly polluted regions." This is confusing because at this point in the text we don't know what is meant by "set the RMSE". We learn later that these values will be used to populate the error covariance matrices for the inversion; please clarify that here so the reader understands what is happening.

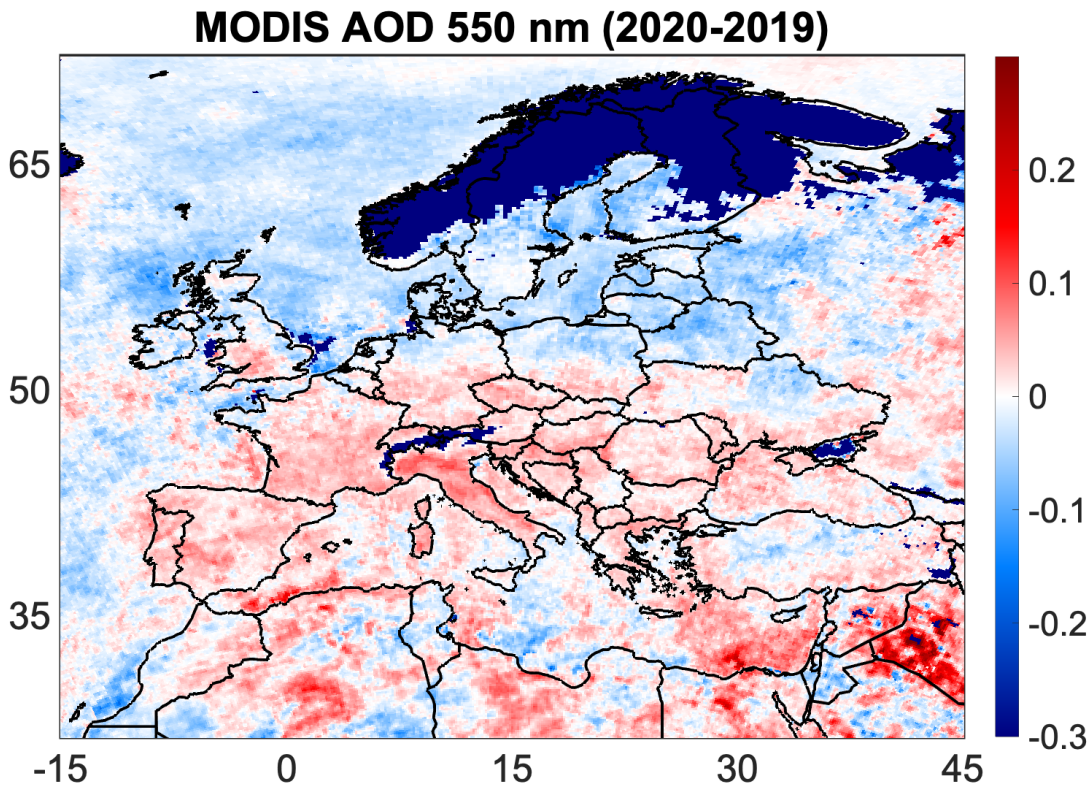
Thanks, we changed it. Please see our earlier comment.

199-202. Does this mean that you only use the dark blue product? Please clarify.

Yes, we clarified it in the text.

The TROPOMI retrievals do not account for aerosols in the scattering weights. Yet I presume that aerosol loadings over Europe changed between the COVID and reference period. To what degree does this bias the retrieval differences and therefore the NO2 and HCHO comparisons between these periods?

This is an important question but not really a big issue here. Based on a very recent study from our group studying the aerosol impact on AMFs using the OMI aerosol product, Jung et al. [2019] observed negligible differences in the amount of HCHO VCD (<10%) over Europe in March-May after accounting for the impacts (including AOD, SSA, and aerosol height). The same thing applies to NO2 [Cooper et al., 2019]. The MODIS deep blue channel indicates that changes in AOD at 550 nm between 2020 and 2019 (in March-May) are <0.15:



We speculate that the aerosol complications are more relevant for biomass burning areas, regions with large AOD like China, and retrievals that are extremely sensitive to even small amount of aerosols such as CO2 and CH4.

To account for the reviewer's comment:

"Aerosol effects on the scattering weights are not taken into consideration. Based on radiative transfer calculations and satellite-based aerosol products, Jung Y. et al [2019] and Cooper et al. [2019] observed small changes (<10%) in AMFs with and without considering the aerosol impacts in Europe in springtime. This tendency likely results from a low aerosol optical depth."

236. "We nudge moisture, wind and temperature fields toward the reanalysis data used only outside of the PBL layer." Wording is unclear, as is the reason for doing this. Please clarify.

We included the purpose:

"To minimize the deviation of the model from the reanalysis data, we turn on the grid nudging option with respect to wind, moisture, and temperature only outside of the PBL region. The inclusion of this option outside of the PBL is because we do not want the coarse reanalysis data washes out the relatively high-resolution dynamics".

239. "Extensive model evaluations based upon surface observations show a striking correspondence". The model temperature bias appears to be 50% smaller in 2020 than 2019 (0.8 vs 1.2 degrees). Does this have any impact on the model interpretation of changes between years? For example, assessing changes in anthropogenic VOCs relies on distinguishing changes in biogenic emissions which depend exponentially on temperature.

The absolute difference only matters which is only 0.3º (-0.9 vs -1.2) on average. We were able to derive changes in anthropogenic VOCs (due to the chemical feedback of NOx on VOC) only in March where biogenic emissions are so low such that the difference in the temperature bias is negligible. This is because the average temperature and solar radiation are so low to have any significant impact on biogenic emissions. Please see the new Figure 4 including the biogenic fraction.

239. PBL height is a major factor for model performance in simulating AQ-relevant species. How well does the simulation capture measured PBL depths over your domain?

Thanks for your comment. It is indeed important. Reliable PBL observations are rare and are usually estimated based on signal processing methods applied on LiDAR attenuated backscatter values. While we did have this type of information during NASA-funded campaigns on aircrafts and over the ground, we really could not find any publicly available data over this region.

242. Please state the time resolution at which you are optimizing emissions. I guess there is a single 3-month mean value being derived for each grid cell but unless I missed it I don't think this is stated anywhere.

We had mentioned it in the inversion part, but now we elaborated to:

"The inversion window is monthly meaning we have three separate correction factors in months of March, April, and May"

250. Please state how the Jacobian is calculated. Is there a finite difference run for every model grid cell, each tracer, and each iteration?

Each iteration, uniformly to all grid cells (20%. Perturbation to each grid cell). One for NOx and one for VOC. Perturbing individual VOCs is an overkill because the problem is under constrained. Perturbing each individual grid cell while keeping others constant will require 369979 forward simulations, each takes around ten days using 250 cpus. Given the computation resources we have here, that would take around 5068 years (369979 * 2. (nox and voc) * 10 days / 365 days / 2 (two runs at same time with 500 cpus). We obviously need the adjoint of the model to expand the number of state vectors which is not updated for the newer versions of CMAQ.

"and K_i $(= K(\mathbf{x}_i))$ is the Jacobian matrix calculated explicitly from the model using the finite difference method by perturbing separately NO_x and VOC emissions by 20%. The perturbations are applied for each iteration."

252. "In terms of the prior errors, we use the numbers reported in Souri et al. [2020a]." Since this is an important aspect of the inversion please briefly summarize here.

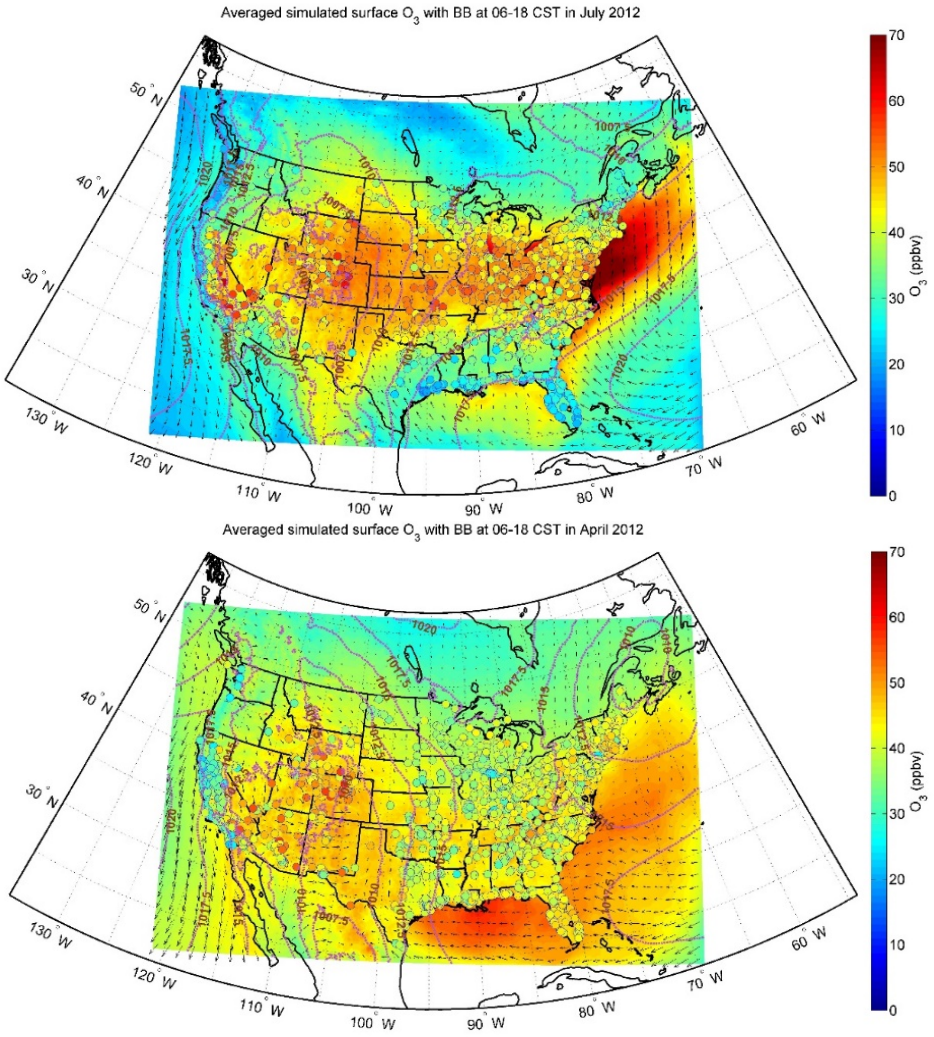
Sure: "The prior errors in anthropogenic NOx and VOCs emissions are set to 50% and 150%, respectively. In terms of the biogenic emissions, the errors are set to 200% for both NOx and VOCs."

257. "here we iterate Eq 1 3 times." How do you know this is sufficient? As you know the emission-concentration relationship for NOx in particular is highly non-linear. Do you employ a test for convergence?

This number is based on two factors: first, we used the same number in our previous study [Souri et al., 2020a] over East Asia for which we found satisfactory results against satellite/observations indicating a reasonable convergence. Non-linear chemistry was a problem in that region due to large oxidation capacity and emission rates. Second, we defined certain time and computational resources for this project. The analytical inversion is extremely time (and space) consuming at a continental regional-scale.

275. "faster vertical mixing due to larger sensible fluxes (more diluted columns due to stronger advection in higher altitudes)". This is a little convoluted. Faster vertical mixing by itself wouldn't change the column amount, and faster winds during summer (really?) would only smear the columns.

If we rule out the advection, the vertical mixing does change the amount of column through changing the dry deposition rates (please see the dependency of dry deposition and vertical diffusion in Figure 10). In several places (such as the southeast US), winds are dominantly stronger within 700 hpa in *summer* due to certain synoptic conditions. Some examples based on Souri et al., 2017 are shown below (summer top, spring bottom). Please note that the Bermuda high is shifted westwardly increasing the pressure gradient between the Midwest low and itself ultimately resulting in strong low jet streams.



If the receptor is located at *A* and winds are stronger vertically, the receptor *A* will experience lower concentrations of NO2 columns. We cannot change the location of the receptor at the same time.

To better word this, we changed the sentence to: "faster vertical mixing due to larger sensible fluxes (more diluted columns for a given receptor due having a greater chance of experiencing stronger winds in higher altitudes),…"

277-280. Wording is quite awkward here.

We reworded: "This sequential decline of NO_2 obscures the quantitative interpretation of the satellite observations in two ways: first, as noted by Silvern et al. [2019], the free tropospheric background NO_2, which are highly uncertain, becomes comparable to those located in near-surface, and second, the relatively lower signal-to-noise ratios reduce the amount of information obtained for NO_x emission estimates (discussed later). "

281. "pronounced decreases", please clarify that you mean in 2020 vs 2019.

Added.

282. "In contrast, we see negligible reductions…" actually some of the regions mentioned seem to show a clear March increase.

We changed to increase.

296. "suggests an abrupt hiatus in the ongoing reduced NOx emissions". Unclear if this means the emissions went into hiatus or the reduction went into hiatus.

We changed is to: "In May, the anomaly of the tropospheric NO_2 suggests that the reduction in NO_x emissions abruptly experiences a hiatus in central Europe (box G)."

317. "but also stems from the fact that isoprene reactivity significantly increases by rising temperature [Pusede et al. 2015]." This is a bit oddly worded; I think you simply mean that OH is increasing seasonally along with isoprene emissions.

Thanks, we changed it to: "This signal is not only induced by the inherent temperature-dependency of biogenic isoprene emissions, but also stems from faster isoprene oxidation through higher levels of OH [Pusede et al. 2015]."

335-337. These evaluation statistics should be displayed in SI in a table or figures.

Thanks, we included relevant figures in the SI.

"We observe a large improvement (31-45%) in the bias associated with simulated surface NO_2 using the posterior emissions compared to the surface measurements in many places around Europe with an exception to northeastern Germany where TROPOMI NO_2 observations deviates the model from the measurements (Figs S7, S8, S9 and 10). The improvements in correlation were minimal indicating that the prior location of emissions are well known."

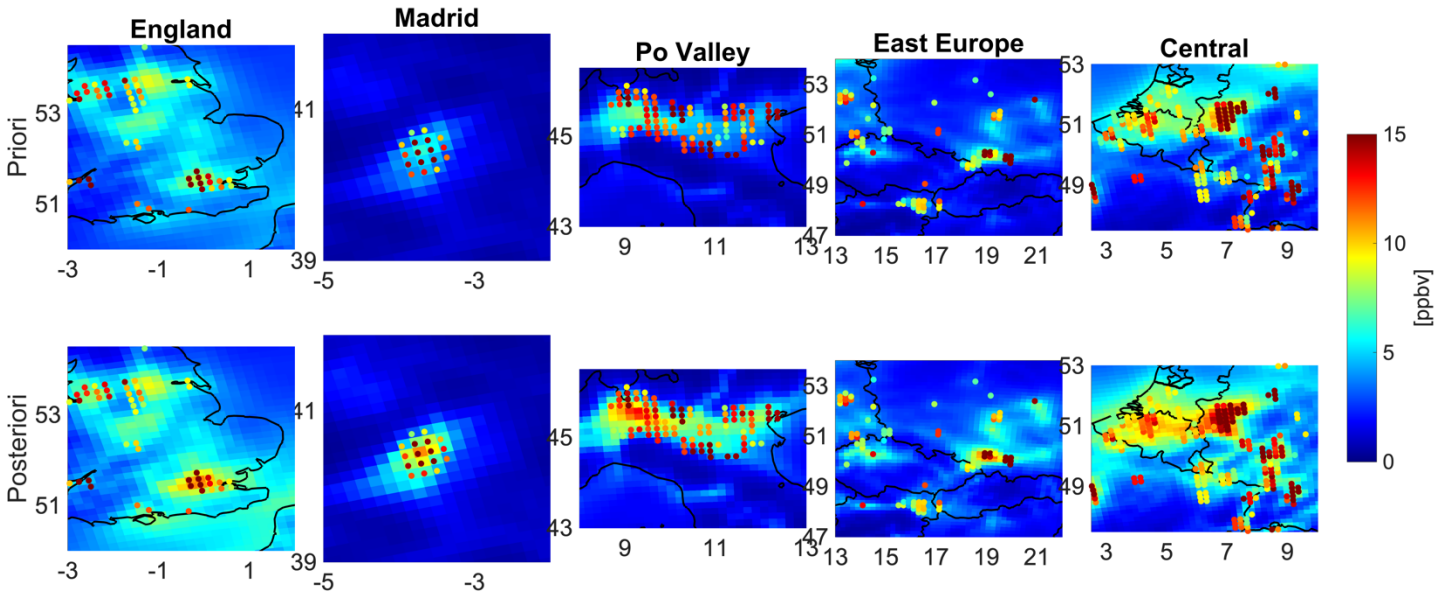


Figure S7. Comparison of daily-averaged surface NO2 observations (circles) against the simulated model in different regions around Europe in March-May 2019 (baseline). The first row uses the prior emissions whereas the second is based on the top-down emissions constrained by the satellite observations through an analytical non-linear inversion.

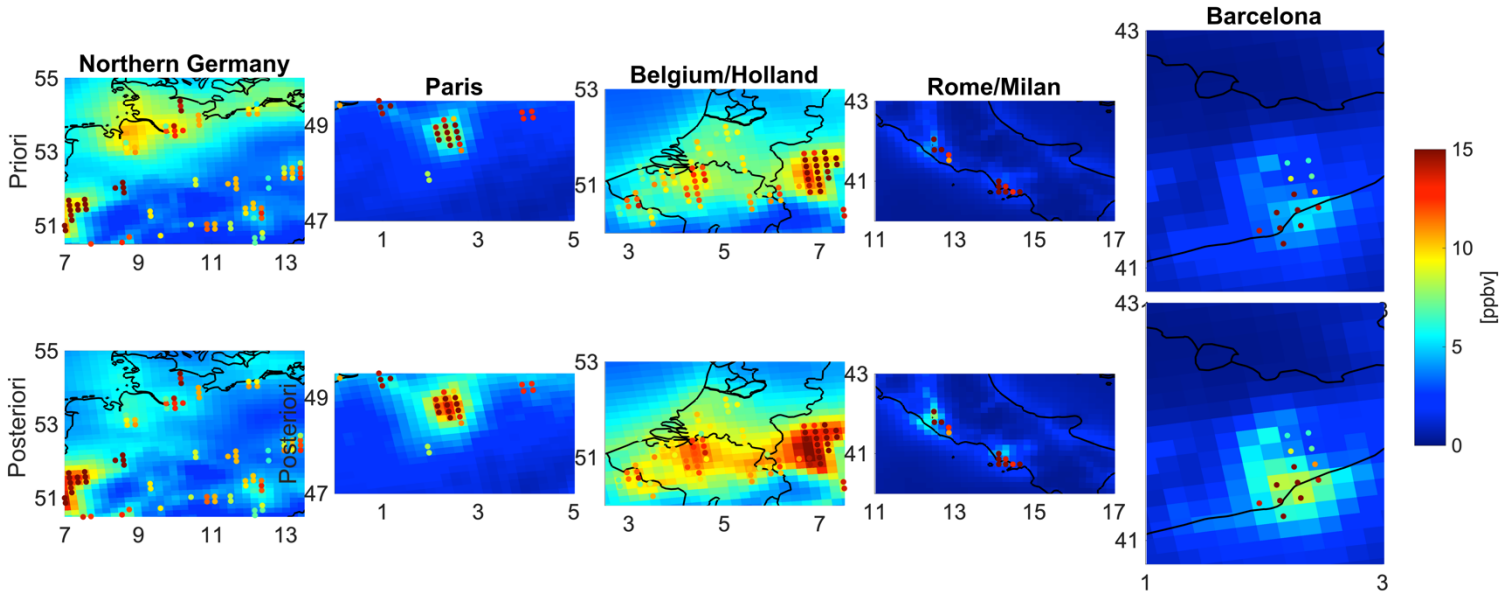


Figure S8. Similar to Figure S7 but in different areas.

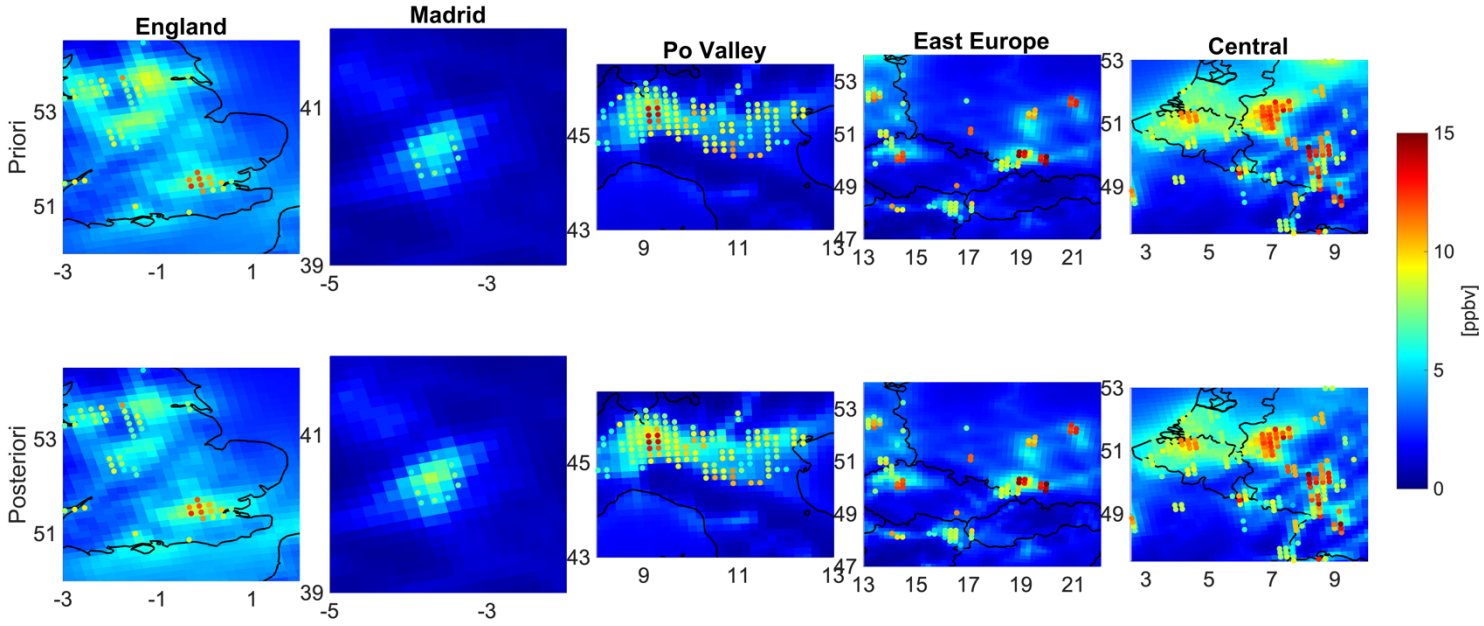


Figure S9. Comparison of daily-averaged surface NO2 observations (circles) against the simulated model in different regions around Europe in March-May 2020 (baseline). The first row uses the prior emissions whereas the second one is based on the top-down emissions constrained by the satellite observations through an analytical non-linear inversion.

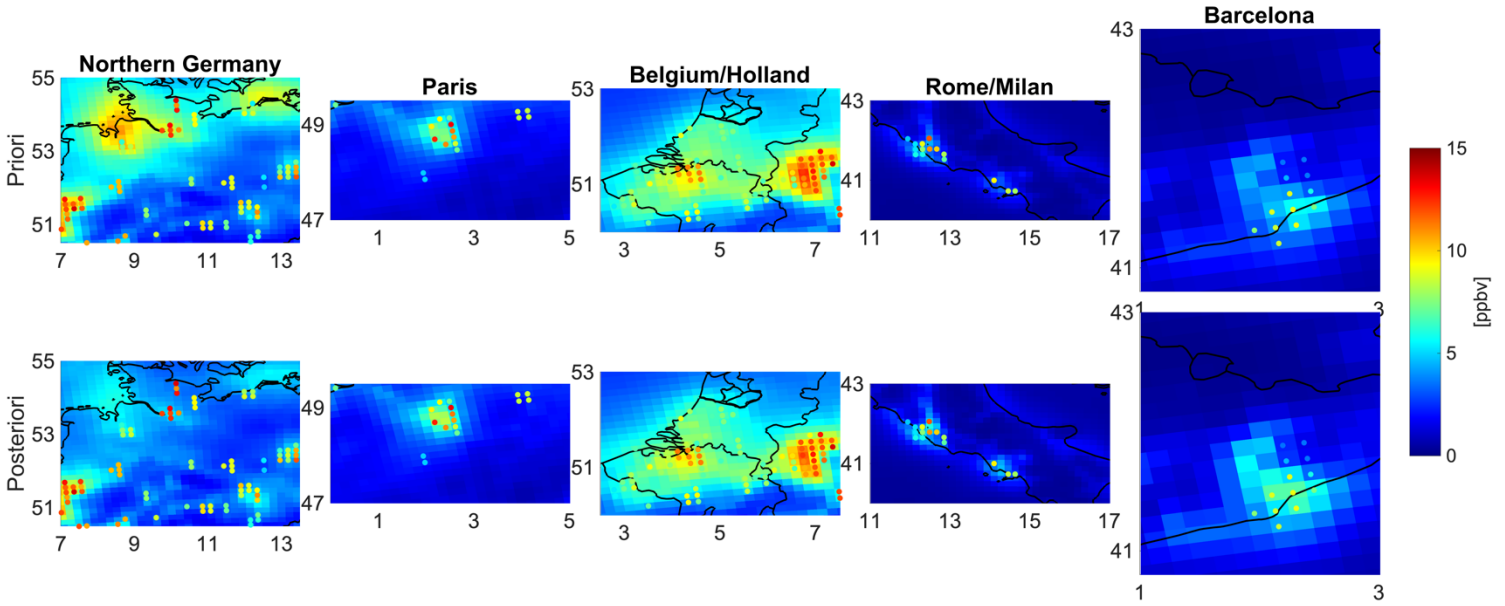


Figure S10. Similar to Figure S9 but for different areas.

341-344. "However, in practical terms, the magnitude of these anomalies is not as drastic as the ratio of observation to model ratio because of the consideration of observational errors and chemical feedback [Souri et al., 2020a], which always leaves some doubt about the practicality of direct mass balance methods." I am unsure what the authors are trying to say here.

We decided to remove this sentence because we had already emphasized this non-linear pattern in more detail in our previous studies.

358-360. The optimization naturally improves the simulation of HCHO with respect to TROPOMI, that is the whole point of the optimization. Does it also improve the simulation with respect to independent observations?

The HCHO observations used for the inversion are only based on TROPOMI. As far as we know, in-situ HCHO measurements are not available (or at least publicly available) during the period. Publicly available MAX-DOAS observations do not cover this period (http://uv-vis.aeronomie.be/groundbased/QA4ECV_MAXDOAS/index.php).

358-378. This paragraph is really unclear; I had to read it multiple times to try and parse what is being argued. It sounds like you're arguing that the chemistry changed the emissions. Please rewrite.

Thanks, we decided to remove this discussion based on Reviewer #2.

395. "Horizontal transport (shown as wind vectors) plays a critical role in explaining the spatial variations in emissions downwind." Why would wind affect the emissions?

We removed the discussion about Figure 6.

397-418. This section is all quite speculative and unconvincing. It does not appear that there is much required information conveyed here, recommend deleting.

We removed it.

409. "This in turn will provide an opportunity for the volume of air to become dispersed". Poor wording. The VOC lifetimes do not affect how a "volume of air is dispersed".

This part has been removed.

422. "Unfortunately we limit the analysis to NO2 due to the lack of routinely measured HCHO observations." The HCHO data are ultimately being used to constrain VOC emissions; so are there VOC measurements that can be used for this purpose?

The HCHO observations used for the inversion are only based on TROPOMI. As far as we know, in-situ HCHO measurements are not available (or at least publicly available) during the period. Publicly available MAX-DOAS observations do not cover this period (http://uv-vis.aeronomie.be/groundbased/QA4ECV_MAXDOAS/index.php). In case the link does not work during the review process:

QA$_4$ECV MAXDOAS reference data sets

FP7-SPACE-2013-1/Project No 607405

Introduction

This webpage provides access to the QA$_4$ECV NO_2 and HCHO MAXDOAS reference data sets. The QA$_4$ECV MAXDOAS product consists of harmonized multi-year time series of NO_2 and HCHO tropospheric vertical column densities at a selection of 10 stations corresponding to urban, sub-urban, and rural conditions. The list of stations is the following:

Station	Lat, Long	Class	Data Source	Time coverage
De Bilt/Cabauw (NL)	52°N, 5°E	Sub-urban	KNMI	03/2011 - 11/2017
Uccle (BE)	50°N, 4°E	Urban	BIRA	04/2011 - 06/2015
OHP (FR)	44°N, 5.5°E	Rural	BIRA	02/2005 - 12/2016
Xianghe (CHN)	39°N, 117°E	Sub-urban	BIRA	04/2010 - 01/2017
Bujumbura (BU)	3°S, 29°E	Sub-urban	BIRA	01/2014 - 12/2016
Bremen (DE)	53°N, 9°E	Urban	IUP-UB	02/2005 - 12/2016
Nairobi (KEN)	1°S, 37°E	Rural / Urban	IUP-UB	01/2004 - 11/2014
Athens (GR)	38°N, 23°E	Urban	IUP-UB	09/2012 - 10/2016
Mainz (DE)	50°N, 8°E	Urban	MPIC	06/2013 - 12/2015
Thessaloniki (GR)	41°N, 23°E	Urban	AUTH	01/2011 - 05/2017

Table 1. List of QA$_4$ECV MAXDOAS stations

440. "The surface measurements reinforce the less pronounced reduction in NO2 in northern Germany and UK, although the magnitudes are not as large as those suggested by the model." This is not clear from the figure. For example the observations suggest that decreases over the UK in April and May are quite large compared to the rest of Europe.

We changed it to be more specific: "The surface measurements in March reinforce increases (or negligible changes) in NO_2 in northeastern Germany and UK, although the magnitudes are not as large as those suggested by the model."

492-496. "This apparent discrepancy is caused by the differences in boundary and initial conditions which are not quantifiable by the process analysis and would require additional sensitivity test." Is it just the ICs and BCs, or is it that these processes being examined are not strictly independent and additive?

From the modeling perspective, the processes are strictly independent. As a matter of fact, if one knows the exact IC and BC contributions and add each chemical/physical components incrementally, they will be able to reproduce the concentration of ozone at a given time and location, very similar to what simulations output. Mathematically speaking, ozone is given by:

O3 = adv + diff + chem + emiss + drydep + cloud_chem + aerosol_chem + IC

Ruling out emiss, cloud_chem and aerosol_chem components, we should be able to know the difference in O3 between two years as long as IC are known. IC for this case is the beginning of April.

Equation 6 is incorrect (the wrong rate constant is indicated).

Thanks, corrected.

544. "This analysis strongly coincides with Lee et al. [2020] and Wyche et al. [2021] who observed roughly constant O3+NO2 concentrations over the UK before and during the lockdown 2020." With this in mind, why not actually just show the modeled Ox = O3 + NO2 change (and measured change too, if available)? This seems like the most direct way to make this point.

Ox is highly influenced by transport. It's difficult to pinpoint which areas have contributed to less/more Ox for a given region. We did not only want to see the general conclusion about the "partitioning between NO-NO2-O3", but also to demonstrate the rates of P(O3). The current analysis adds more information: i) where the NO-NO2-O3 partitioning is mostly occurring which is tied to titration and ii) how fast those rates are changing.

Additionally, those studies should have also included "NO" because Ox is not strictly defined as NO2+O3. But the measurements of NO in Europe are lacking.

572. "The reduced anthropogenic VOC emissions were a result of two key assumptions: the reduced NOx emissions in NOx-rich areas increased HCHO made from VOCs (evident in larger Jacobians derived from the regional model), and TROPOMI HCHO suggested a negligible difference in HCHO concentration between the two years." Again the wording here is really confusing. It appears to be arguing that changes in NOx emissions and in the ensuing chemistry changed the actual VOC emission rates. I think I know what is meant (i.e., that these factors change the emission rates that one infers for a given HCHO level) but it really needs clearer description.

This part has been dropped.

Minor edits.

73. “atmospheric composition” not “compositions”

Corrected.

78. “particulate matter”

Corrected

157 and 185. “clean regions” and “clean areas” rather than “clear”

Corrected

177. “mainly located” or “predominantly located”

Corrected.

437. “by only considering grid cells”

Corrected.