The authors did address many of my comments, and I thank them for that. The paper now provides a much more complete description of what the model and the inversions do, and this facilitates the interpretation of the results. The paper also new comparisons with surface NO2 data (Fig. S7-S10).

Still, regarding many concerns, I remain unconvinced, as explained below. The authors keep insisting in their response that "the results in the manuscript are based on the a posteriori, not the a priori" as if the large (sometimes huge) model biases were simply irrelevant. It is obvious from Fig. S11-S14 that both the prior and the posterior model fail to match the observations at most locations except the most polluted. Over many regions (e.g. Ukraine, etc.), the model-data difference for NO2 is systematic and exceeds the TROPOMI error (~1.1E15, Verhoelst et al.). This implies serious issues in the model and/or in the data (e.g. in the bias correction). Things are even much worse regarding HCHO. If we are clueless as to the causes for such discrepancies over moderately polluted regions, why should one trust the results in very polluted areas? True, the AKs provide an indication regarding where and when the inversion results are most reliable (if we accept the hypothesis that model and data are not too biased). However following that guideline, one would have to accept as very credible the NOx results for Germany in March (Figure 2 and Figure 4) indicating a strong emission increase in Northern Germany (in 2020 with respect to 2019) and an emission decrease in Southern Germany. This discrepancy between regions in the same country is an obvious artefact, as the authors implicitly admitted by removing the discussion on that region. This is a "COVID-19 paper" and the reader should be given some clues regarding patterns of emission changes which are obviously wrong. No need for cell-phone, traffic or industrial data for that. The paper does not provide any clue, probably because of the too many issues with the data (largely due to cloudiness, Fig. S16-17) and especially with the model. For example, the very wrong distribution of NO2 columns in the model over Germany (Fig S11) should have prompted the authors to try to explore its possible causes instead of relying exclusively on the power of inverse modelling. Maybe the inversion is correct, but how does it help anyone if we don't understand why?

Thanks for your comment.

To make sure that we are on the same page, we need to explain the Bayesian inversion here. The goal of the inversion is NOT to *exclusively* match the model to the observations (TROPOMI). The cost function follows a quadratic shape consisting of two terms:

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{y} - F(\mathbf{x}))^T \mathbf{S}_o^{-1}(\mathbf{y} - F(\mathbf{x})) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_a)^T \mathbf{S}_e^{-1}(\mathbf{x} - \mathbf{x}_a)$$

The left-hand side tries to minimize the differences between the observation and the model (F(x)), and the right side uses the prior knowledge as a pseudo-observation (or more precisely, an expectation, because the Bayesian inversion tries to find the maximum likelihood of P(x|y); the mathematical part is explained in https://www.sciencedirect.com/science/article/pii/S1352231016301315 and Rodgers 2000).

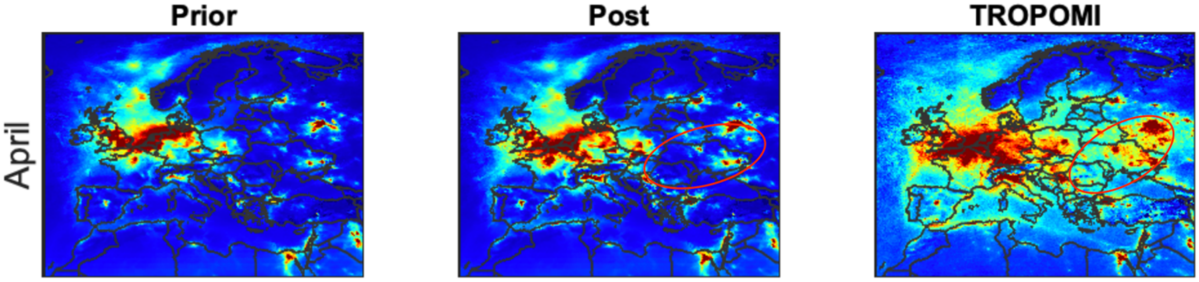
There is a competition between these two terms. If the error of prior knowledge (projected to the observational space) is very small compared to that of observations ($K_i\mathbf{S}_e\ K_i^T$ << So), the estimation will be rather independent of the first term. This is clearly described in the

Kalman gain shown in our manuscript (So>>Se will make G almost zero). For example, if we are given super accurate values of NOx emissions observed by flux measurements (kg/s) over a chimney and someone asks us to verify it using model+satellite data, it will be very unlikely for the top-down emission values to be different than the prior estimate no matter how large the differences between satellite and model are. This tendency manifests in low AKs. One may say: why doesn't your constrained model match with satellites observations in this scenario (or within a sigma value of the satellite observations)? The first question we should ask is: why should it match if the prior knowledge (flux in-situ data) is more certain? Why should we degrade our estimation with garbage? The fundamental reason behind using the Bayesian inversion in all type of data assimilation/inversion frameworks is to prevent the *"garbage in-garbage out"* problem.

Uncertain observations have a less of chance to impact the emissions because the prior knowledge is competing with them. It is intuitively clear that any instrument should have a larger uncertainty (relative to their absolute values) in low/background conditions. For instance, fitting the cross-sections to the satellite radiance will be much harder (and prone to larger errors), if the molecular absorption of our targeted gas is weaker (the reason why HCHO observations are noisier than NO_2, and why TROPOMI NO_2 columns in rural areas are less credible than over polluted ones). It is because of this reason that we see a larger discrepancy between the constrained model at rural areas with respect to NO2 columns compared to polluted areas. Is this a drawback from Bayes' theorem perspective? Absolutely not, this is in fact its power. As a matter of fact, the inversion is a solid way to extract good information (variance) from the data considering the probability distribution of x (P(x)) and P(y|x), all together.

We could have narcissistically inflated our result by increasing the emissions errors or scaling down the observational errors to artificially line up our model with the observations over rural areas. Is this ever logical in the sense that our prior knowledge is completely wrong and/or satellites are so precise? The same problem applies to the biomass burning area in April 2019. The reason why the model seemingly failed to reproduce those values was because it understood that TROPOMI columns were too uncertain over that region and especially in high latitudes (due to considering TROPOMI's uncertainty column variable that were considered as e_precision (in eq1) in addition to 4%). There are "*horror stories*" expressing the consequence of not fully accounting for biomass-induced aerosol optical properties in AMF calculations (which can result in either too low or too high AMFs depending on single scattering albedo). There might be other problems with TROPOMI HCHO. We also do not want to rule out the problems with the model that may provide biased Jacobians over that area (discussed later). Overall, we think having a model not being strongly constrained by controversial observations over that area is better (safer). Kalman gain is so low over that area such that scaling up and down the columns are not going to make a noticeable difference in the a posteriori (assume post = prior + G(model-OBS); G (Kalman gain)=*epsilon*, now scale OBS by 1.2 or 0.8, the post will be resilient to the change).

Now back to reviewer's comment on the discrepancy between the constrained model and both TROPOMI and surface measurements in non-polluted areas. For example April 2019:



First, the discrepancy between Post and bias-corrected TROPOMI (what is shown above), has nothing to do with the bias correction factors. The bias-corrected TROPOMI was used for the inversion, so if *y* was scaled by something different (or nothing at all), the a posteriori would follow that direction ((y-F(x) + bias)*G).

Second, the inversion is not a least-squares estimation meaning that we have a competition between prior knowledge and the observational constraint. Even the outdated mass balance method partly accounts for this competition. Please take a look over Martin et al., 2002; before they suggested that new_NOx = old_NOx * (satellite column/model column), they mentioned how to derive the final answer thorough performing a weighted average (combining the estimates derived from the mass balance method and the prior knowledge):

$$\ln E = \frac{(\ln E_t)(\ln \varepsilon_a)^2 + (\ln E_a)(\ln \varepsilon_t)^2}{(\ln \varepsilon_a)^2 + (\ln \varepsilon_t)^2} \qquad (1)$$

But the fact this equation had buried in the beginning of their paper has made so many scientists unintentionally forget to make some assumptions about the errors.

Third, if one wants to directly compare X (TROPOMI) and Y (the constrained model), they are required to consider the variance of X in the comparison through a Monte Carlo method, or simply a weighted chi-sq minimization. Pixels with larger uncertainty should have smaller weights in the comparison. So aiming for 100% match between the model and the TROPOMI is flawed. This is one of the primary reasons why we are using an inversion here. Granted, any inversion system will naturally get contaminated with the model parameter errors (discussed later) meaning the estimates are simply the added information on top of an ignorant model.

Here, the reason why we still see some differences in Post and TROPOMI over rural/suburban areas is because the prior emission has played an important role. We could have loosed the prior constraint (which is set to 200% for biogenic emissions, say we could have set it to 400%) to see a bigger change (and closer values) compared to TROPOMI. But that is just one hypothesis, one realization. We shall not forget: "All models are wrong but some are useful" said George Box. We just want to learn some tendencies from the model. Who knows how uncertain MEGAN soil parametrization is? We had been transparent about all numbers going into the model/inversion since the beginning of the process. Based on these numbers, the presented results are the optimal estimates of

emissions given prior/observation errors, which ultimately helped us to get a decent anomaly maps in April over the central Europe due to less cloudiness.

So many arguments are provided in one paragraph so again we need to fragment them:

If we are clueless as to the causes for such discrepancies over moderately polluted regions, why should one trust the results in very polluted areas?

The results over polluted areas are relatively more credible because the TROPOMI has a larger weight to constrain the model compared to the prior estimation (discussed above). The generalization of the fact that we do not see a good agreement (which needs to take into account the variance of TROPOMI and the prior) over less certain columns so we cannot trust the stronger (more certain) signals disregards the concept of Bayesian inversion. We had provided the independent measurements in our original manuscript for a reason. In retrospective, we wish we could have FTIR/MAX-DOAS HCHO observations (if available) to do the same thing for HCHO changes.

However following that guideline, one would have to accept as very credible the NOx results for Germany in March (Figure 2 and Figure 4) indicating a strong emission increase in Northern Germany (in 2020 with respect to 2019) and an emission decrease in Southern Germany.

As we mentioned in the review process, we had verified them with scrutiny: AKs and surface measurements.

First, without using any model or the inversion, TROPOMI (Figure 2) is suggesting those semi-artifacts (although we should recognize that there are some similarities between surface anomalies and TROPOMI over northeastern Germany; and we still believe the smaller reduction in NO2 levels over northeastern Germany suggested by surface measurements is because it's less urban thus less impacted by the reduced mobile emissions). We also do see the semi-artifact patterns without the bias-correction (shown later). Second, it is true that AKs suggest that TROPOMI was able to provide reasonable information on the emissions relative to the prior knowledge (which is subject to errors; the true AK requires the exact TROPOMI and model parameter errors), but there are four important complications that we must consider:

i) we cannot expect that having very few number of observations (sometimes down to 4 days out of ~30 days in a month) from TROPOMI will help us to capture the full picture of emissions over the surface (surface measurements are based on all days); so we have a temporal representativity issue and ii) we do not know the exact statistics with respect to TROPOMI errors over the area for two different years. There could be an issue with the prior profile, or any sort of uncertainty resulting in relatively too high TROPOMI NO2 in 2020 over some areas iii) could this be due to column/surface decoupling issue that are usually not resolved in CTMs; the PBL is parametrized in this model (because the spatial resolution of the model is much coarser than 100-300 m^2 where you can leverage LES). iv) could it be due to not fully capturing lightning in the CMAQ model (and TM5) resulting in high values of TROPOMI NO2 in 2020, since too low AMFs can cause too high VCDs? All

of these can happen but none of these can be easily proven without real data. To our best knowledge, publicly data are not available. <u>This study is trying to portray what we can do given our current knowledge on satellite validation, surface networks, and model parametrization. It is not about enhancing our knowledge on atmospheric chemistry from the limited data during the pandemic.</u> The paper concludes with: "Unless a comprehensive air quality campaign targeting COVID-19 related lockdown is available, we recommend that the impact of lockdown on air pollution should be examined through the lens of well-established models constrained by publicly available data, especially those from space in less cloudy environments."

At this point, without having real data, all we can do is to articulate past issues related to models/observations:

In Section 3.1 we had pointed that :" However it is crucial to note that these maps are based upon sporadic clear-sky pixels that might obscure the full portrayal of emissions changes happening throughout the period (discussed later)."

we added some limitations:

The constrained model correlates reasonably well with the changes observed by the surface measurements in April, but it fails to reflect those in March and May.

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 (and TROPOMI NO_2 columns). A number of factors can contribute to these large discrepancies: i) the surface measurements were present throughout the month of March, whereas TROPOMI data were frequently absent due to cloudiness resulting in some degree of temporal representativity issues; ii) the statistics used for the TROPOMI bias-correction may not always hold true, since each individual pixel can deviate from the norm of the reported biases; iii) the shape of NO_2 profiles simulated by the WRF-CMAQ can sometimes be uncertain due to errors in the PBL parameterization or the difficulties with resolving the non-hydrostatic components (where vertical motions are comparable to horizontal ones) [e.g., Pouyaei et al., 2021]; this complication can result in unrealistic changes in the columns.

In conclusion:

"Third, the changes in NO_x emissions suggested by TROPOMI NO_2 and the constrained model over northeastern Germany in March and Eastern Europe in May were unrealistic, possibly due to observations and/or the model issues."

For example, the very wrong distribution of NO2 columns in the model over Germany (Fig S11) should have prompted the authors to try to explore its possible causes instead of relying exclusively on the power of inverse modelling.

We would need aircraft spirals and surface spectrometers to investigate that. Again, we see the wrong distribution of NO2 columns from TROPOMI, which means that there can be an issue with the observations.

That being said, the authors have accounted for many of my concerns and updated the text accordingly. The authors realized that the NO2 results are less reliable in March and May, which is why the ozone analysis is restricted to April, as it should. They state to their defense that "remote sensing data provide limited information for optimizing emissions [many references]" which I find contradictory since emission optimization is precisely the methodology adopted in this paper, and the paper provides in great detail relative and absolute differences (2020-2019) of top-down emissions (Figure 4 and Table 2) despite those limitations.

It is true that satellites can sometimes provide limited quantitative information (like HCHO in higher latitudes or NO_2 in rural areas) or depreciate the model analysis (like northeastern Germany in March, and eastern Europe in May), but the motivation of this study is not to blindly advertise their utilization. Our major motivations stated in the introduction:

"The motivations of this study are to determine the capability of a regional model constrained by satellite HCHO and NO_2 columns to capture near-surface pollution, and if local ozone production rates are the driving factors for heightening ozone pollution during the 2020 lockdown. In other words, what chemical and physical processes are associated with the elevated ozone? How representative are satellite observations at capturing surface air quality through an inversion context? Is meteorology the primary factor in shaping elevated ozone as suggested by Ordóñez et al. [2020]?"

This paper is an ozone study. The surface ozone levels are function of emissions, meteorology, transport, and etc. To perturb the emissions, one may cut mobile emissions by 50% (like what most covid-19 modeling studies have been done) or try to provide top-down estimate using real data (satellites). We chose the latter and as a result, we found a decent spatially-varying anomaly in NO2 in April. This helped us to achieve a decent anomaly in term of ozone (which is extremely hard to get from models; please take a look over Figure 2 in https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2020JD034213, April 2020 O3 (the middle right over Europe); this paper is written by great scientists, but their model over the Europe is not as responsive as our model (and OBS) to the changes in emissions/meteorology). We broke down each individual physiochemical model process to understand why surface ozone is higher in 2020.

The new figures S7-S11 with the comparisons with surface NO2 data are interesting but I wonder about their significance. There is an overall bias reduction (31%-45%, although the precise meaning of this range is not made clear) but, despite the TROPOMI NO2 constraint, a large systematic underestimation of modelled NO2 remains with respect to the data. Why is that? Could this be due to representativity issues? Or interferences in the NO2 measurements? Note that daily averages are strongly influenced by night-time chemistry. For a proper evaluation of TROPOMI-constrained model values, it would be preferable to sample the data as well as the model in the early afternoon (say between 12 and 15 LT), since night-time chemistry is not well constrained by TROPOMI. In addition, it would be useful to summarize the comparison over the different subregions of Fig. S7-S11 in a table (including mean bias, mean absolute deviation and correlation coefficient). The authors state in their response "The absolute amount of CEDS is too low in many places" as explanation for the model underestimation. This is not a comparison of CEDS with NO2 data. There might be issues in the model (PBL mixing, chemistry) or in the data (interferences, representativity issues) which should be acknowledged first before holding the emissions as responsible.

We actually found a very recent study done by Sun et al., 2021 (https://acp.copernicus.org/preprints/acp-2021-268/) who found the same inventory to be significantly lower than other top-down estimates; See figure 11 in their paper. Yes, chemistry, PBL, and the NOz interference partly play a role, but they can't fully be responsible for such a large underestimation of NO2 in the model. We have extensively used the WRF-CMAQ in our previous studies and we have never observed anything even close to these large negative biases [Souri et al., 2016; Souri et al., 2017; Souri et al., 2018; Souri et al., 2020] except for emission offsets happening episodically in petrochemical industries. We strongly believe the emissions are the primary culprit.

We added more explanation and added the statistics in the supplementary for the selected regions. We need to focus only on daily-averaged data to minimize the impact of discounting for the diurnal mobile emissions due to the lack of the diurnal factors in the CEDS inventory.

A major philosophical question arises here: why should point measurements absolutely match to the grid data? Point is an element of space, but model (and satellite) grids (at best) represent the average values. This is a fundamental issue in our community. We just recently tackled this taboo (https://drive.google.com/file/d/1a7qc_YHU3zP_pp1U4GfomgV0e6AZv7T7/view?usp=sharing). Vigouroux et al. may be interested in this study as their statistics could have been largely impacted by the spatial representativeness.

We also forgot to mention in the previous review that we did not see such a large deviation of NO2/(NOy) factors portrayed by Lamsal et al., 2008, from 1.0 in our previous studies over Texas and CONUS. The NO2 correction factors were topmost 10-20% in September 2013 over Houston, which is far lower than 60-80%(!). Perhaps, the chemical mechanisms have been much improved since 2008. The coefficients may need to be readjusted with newer models. Moreover, the spatial representivity factor is hugely impacting Lamsal's results. NO_2 values are spatially heterogenous making the spatial representivity error massively large.

To account for the reviewer's comment:

We added Table S3 and S4.

We added:

We observe an improvement in the statistics 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; Tables S3, S4). The large underestimation of the model in terms of surface NO_2 concentrations is most likely due to the underestimation of CEDS inventory [e.g., Figure 11 in Sun et al., 2021]. However, it is worth noting that the disagreements between the model and the surface measurements do not solely reflect the uncertainty in the emissions. A major complication arises from the fact that point measurements represent concentrations locally, whereas the model grids (15×15 km^2) are (at best) the average of infinitesimal points integrated over the grid space. Essentially, no one should expect that these quantities will completely line up, unless one transforms the point measurements to the grids (i.e., rasterization) by carefully modeling the spatial auto-correlation (or semivariograms) of the point data [Souri et al., 2021]. Additionally, there is uncertainty about the chemical mechanism used in the model. In particular, Souri et al. [2017] observed a large overestimation (~ factor 4) of daily-averaged total nitrate ($HNO_3 + NO_3^-$) in the CB05/AERO6 mechanism despite moderately reasonable nitrate (NO_3^-) simulations. This was attributed to a large overestimation of N_2O_5 hydrolysis rate [Bertram and Thornton, 2009] which is the primary loss pathway of NO_x in low photochemically active regions [Shah et al., 2020]. The interferences from the NO_z family on the surface measurements might be still present in springtime in midlatitudes (~10-30%) [Lamsal et al., 2008]. Last but not the least, the PBL parametrization controlling the level of vertical mixing rates has errors primarily due to soil moisture not being observationally constrained in the model [Huang et al., 2021].

The authors complain that they "do not understand why this reviewer is concerned about the prior emissions. This paper is not about validating CEDS (...) We could have used EDGAR emissions and reached similar results (...) [or] even used constant emission rates throughout Europe and induce the emission changes by TROPOMI". This is wrong. The striking similarity between prior and post NO2 columns (Fig S11-S14) indicates clearly that both TROPOMI and the prior determine the solution, and therefore the choice of the prior does matter a great deal. Even in areas with high AKs where the inversion is mostly driven by the observations, the changes in the patterns of the emissions induced by the inversion must be questioned: are those real or could they be related to issues in the model or in the observations? I'm not asking you to solve these issues but only to consider and discuss them in a more balanced way.

Thanks for your comment. We both agree and disagree. If we set NOx emissions constant in the entire region, it is wrong to set the prior errors to 50%; it should be ~10000%. This will automatically simplify the problem to an iterative joint mass balance method. The final result may not be as good as considering the prior knowledge (the height of injection, rates, speciation with regards to VOCs, …), but the final estimate will be much better than a constant rate (i.e., AK~1 almost everywhere). From an optimization perspective, this is a successful case. But if we are aiming for the exact rates and to reduce the computational costs, it is obvious that we should have reasonable prior values. We already addressed these problems in the first comment.

Regarding the HCHO inversion I still believe that the results of the inversion have very little value given the high differences between the model and TROPOMI. If they think that the emission differences 2020-2019 from the inversion are significant, please provide a quantitative estimation of the uncertainties on the retrieved emissions.

The authors insist that the high TROPOMI HCHO values in April in Northern and Eastern Europe are not too high, citing the negative bias over high-HCHO level sites reported by Vigouroux et al. 2020. The hypothesis that the negative bias is due to aerosol effects is what it is, a hypothesis. In April 2019 over Saint Petersburg (a megacity right in the big HCHO plume on Figure S13), TROPOMI is overestimated by about 35% based on comparisons by Vigouroux et al. Same thing over Sodankyla and Kiruna. Those direct measurements inside the HCHO hot spot are much more relevant than speculations about the possible role of biomass burning. Your manuscript should acknowledge that TROPOMI HCHO is very probably overestimated in that area, for reasons unknown. I do not dispute the fact that the direction of the emission increment in April 2019 goes in the right direction. Of course it does. But you do not need a sophisticated inversion system to infer that emissions were higher than the model prior in April 2019.

We decided to move the VOC and HCHO part to the supplementary material as there are not independent data to verify the model/TROPOMI. We cannot remove this part entirely because the VOC emissions are constrained (especially in lower latitudes where the signal is strong). The results must be reproducible, therefore we need to inform readers about all modifications applied to the model including VOCs. We need to keep the results in the supplementary part for different reasons:

i) **Ozone chemistry is a function of NOx and VOC emissions, even though TROPOMI HCHO (not only the inversion) didn't provide significant information on emissions in midlatitude/high latitude regions, we do see a good amount of information in lower latitudes (Mediterranean basin for example).**
ii) **The atmospheric lifetime of ozone can reach to several weeks. So it is important to constrain the relevant emissions in large areas as much as we can. This is more critical for April 2020 when the high pressure system over the central Europe has extended to lower latitudes implying that there are strongly regional background contributions.**
iii) **There are some changes on VOCs in central Europe in April 2020. The results won't be reproducible if these changes are not applied (especially if we consider the larger sensitivity of ozone production rates to VOCs in NOx-statured areas). The results must be reproducible, so the analysis on VOC/HCHO should be included somewhere.**
iv) **We do not polarize the results. Each pixel should be treated as an independent instrument with different uncertainty.**

We briefly mentioned the results in abstract/conclusion:

The observational constraint on VOC emissions is found to be generally weak except for lower latitudes.

Fourth, we observed a weak observational constraint on VOC emissions from TROPOMI HCHO except for lower latitudes.

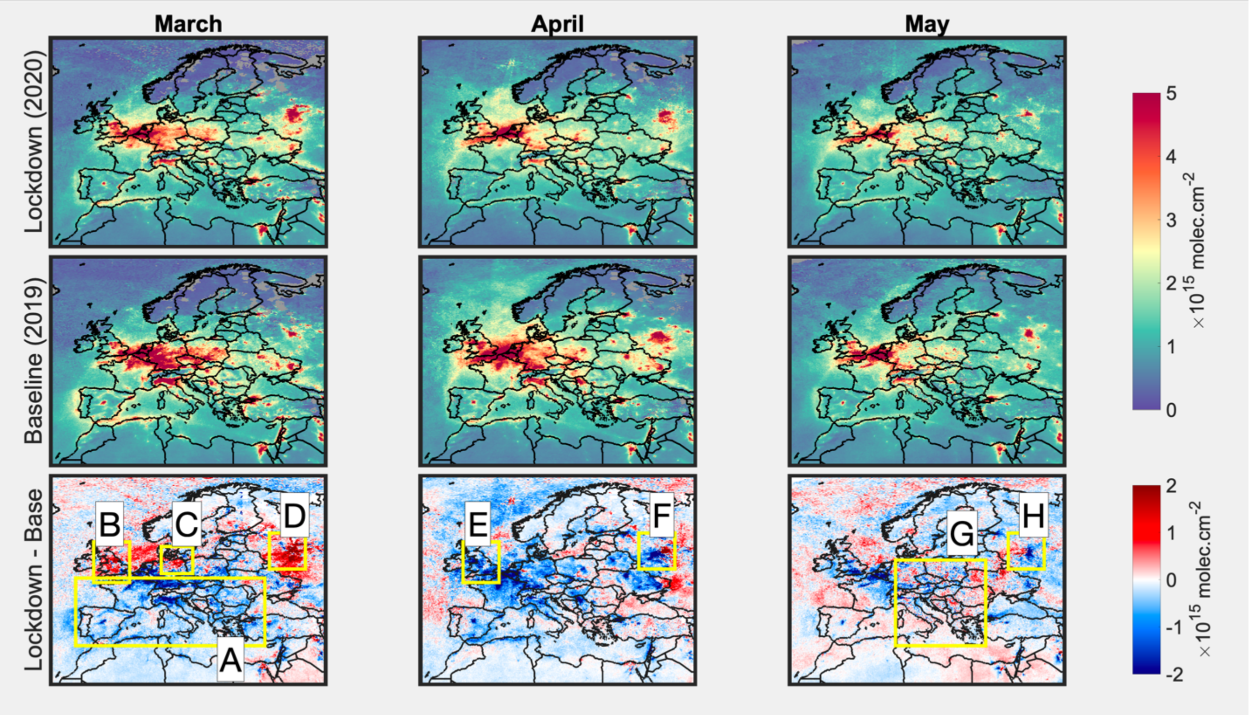
We added in the supplement:
It is worth noting that the TROPOMI bias-correction factors used here based on Vigouroux et al. [2020] are not necessarily correct over this area possibly due to snow cover, the profile shapes, or non-linear aerosol impacts on AMFs (see Figure5 in Vigouroux et al. [2020]).

The title has changed to weigh down the VOC part: Unraveling Pathways of Elevated Ozone Induced by the 2020 Lockdown in Europe by an Observationally Constrained Regional Model using TROPOMI

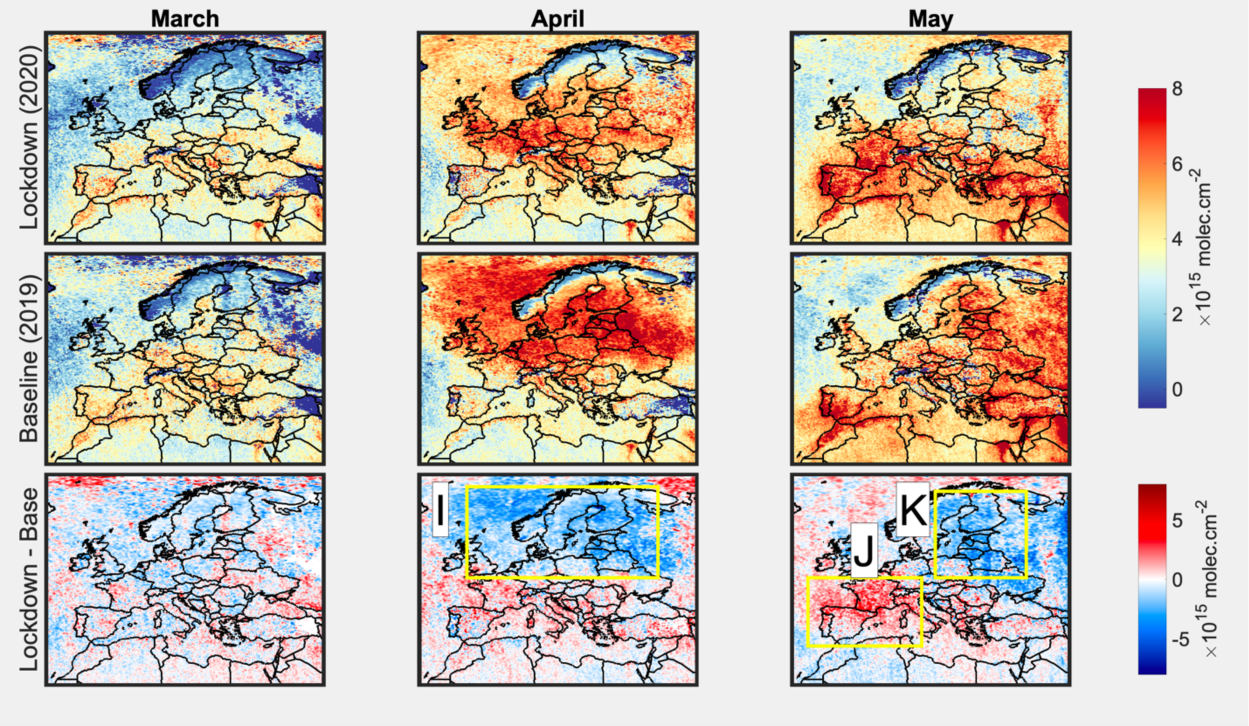
The authors did not address my comment that the bias correction might lead to different corrections being applied in 2019 and 2020, thereby creating artificial patterns in the differences between the two years.

We added that (above): ii) the statistics used for the TROPOMII bias-correction may not always hold true, since each individual pixel can deviate from the norm of the reported biases

FYI: NO2 results without the bias correction; please note that those positive values over C and B are still present. With these results, the underrepresentation of the model in terms of the NO2 reduction in April would have worsened.



And for HCHO without the bias correction:



We don’t believe the TROPOMI bias-correction has impacted the conclusions drawn from our analysis. It worked in the favor of capturing NO2/NOx in April. Again, over the biomass burning area, the Kalman gain is so low that it doesn’t really matter if the correction factors or +30% or -30%. Also, we did not conjure up some numbers out of nowhere. The correction factors are based the recent validation studies.

I strongly recommend to drop that part on HCHO and the VOCs which, despite relying on a sophisticated inversion scheme, cannot do better than a simple visual inspection of model results and observations. For the NOx part, I recommend strongly to drop Table 2 and reword many parts (including discussion, abstract and conclusions) in order to convey the known limitations and uncertainties of inverse modeling (as acknowledged by the authors, see above).

Mentioned above. The abstract is already too long to include them.

Minor comments:

- Thanks for the clarification on the AMF and the profile shapes. How does that relate to the averaging kernels (in the definition of e.g. Eskes and Boersma 2003, www.atmos-chem-phys.org/acp/3/1285/) used by other groups to derive total columns from model profiles?

To be able to use “averaging kernel” variables in TROPOMI data, we need to multiply those values by AMF. This quantity will be identical to box AMF (or sometimes wrongly called scattering weights in some literature; box AMF = AMFg*SW; final AMF = box AMF*SF). See eq 7.2 in https://sentinels.copernicus.eu/documents/247904/2476257/Sentinel-5P-ATBD-HCHO-TROPOMI.pdf/db71e36a-8507-46b5-a7cc-9d67e7c53f70?t=1625507823781

- Thanks for the clarification on the observation error. Note that the TROPOMI precision estimation might actually contain non-random parts. Your inversion system does not account for model errors. Can some crude estimate be provided for those? How could their omission impact the results?

Propagating the model error parameters (such as winds, PBL, clouds and etc.) to the final estimation requires a fully explicit calculation of Jacobians (here linking columns to that specific parameter, and finally columns to emissions) which is computationally burdensome and sometimes not possible (do we have the 4D error of wind vectors or cloud microphysics for example?); from Rodgers, 2000:

$$\mathbf{S}_f = \mathbf{G}_y \mathbf{K}_b \mathbf{S}_b \mathbf{K}_b^T \mathbf{G}_y^T \qquad (3.18)$$

G is the column-emission relationship, Kb describes the column-model parameter relationship. Sb is the covariance matrix of the model error parameter.

That’s an oversight which we had touched upon in Souri et al., [2020]. So, we essentially tend to under-predict the errors in the top-down estimation because of treating the model parameters as perfect. An alternative way to quantify these errors is to run ensemble of models with different parametrizations/initializations/reanalysis data. That would beautifully provide a set of solutions for the estimates and the resultant anomaly maps (like ozone), which in turn, it would help us with detecting outliers. We have the scripts ready to do that (see https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD031941), and our first attempt for this study was to follow that approach with different emissions/parametrization (https://twitter.com/AmirHSouri1/status/1271485748915519491?s=20). But we just could not afford the computational cost at some point, so we had to switch our inversion to what we did in Souri et al., 2020.

To account for the reviewer comments:

“An important caveat with this inversion system is that we do not take the model parameter error (such as errors in chemistry, cloud microphysics, and PBL) into account. To properly estimate the forward model parameter errors, one needs to calculate the sensitivity matrix of the columns to the model parameters combined with the sensitivity matrix of the columns to the emissions (*K*) [Rodgers, 2000]. The former calculation is computationally expensive. Moreover, the spatiotemporal varying model parameter errors may not be known in detail. The consequence of disregarding the model parameter errors is the overconfidence in the top-down estimates (i.e., overestimations of AKs).”

- The HCHO TROPOMI product provides random and systematic error estimates, why not using those? The 4% seems too low considering the large variability among the different sites in Vigouroux et al.

We addressed in the supplement: We assume the constant term of errors (e_{const}) to be equal to 4% of HCHO total columns based on Vigouroux et al. [2020]. The precision error ($e_{precision}$) is populated with the column uncertainty variable provided with the data.

- What are the implications of neglecting diurnal variations of anthropogenic emissions?

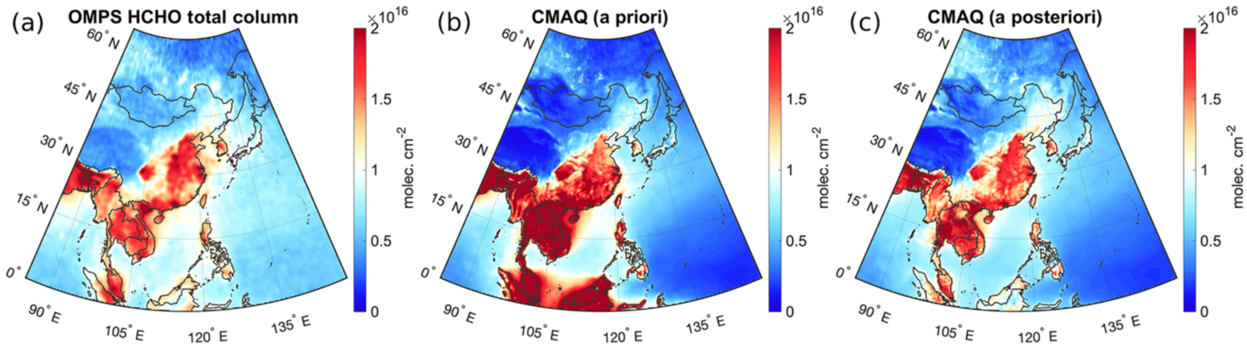
For ozone? It depends on the underlying chemical conditions. Our former studies over industrial areas in Texas in summertime revealed that it was sometimes important (by sometimes we mean on very stagnant and photochemically active day) to have the right timing of mobile emissions if we want to replicate the pick of ozone in NOx-sensitive areas. We should recognized the fact that not all anthropogenic sectors have diurnal cycles. But we are unsure if this matters when it comes to the springtime ozone over Europe. The chemical conditions are majorly NOx-saturated, and O3 titration through NOx is prevalent. So having a fixed emission rate over a diurnal variability would decrease/increase the titration during daytime/nighttime making ozone slightly higher in daytime, and lower in nighttime. We are unsure if this is critical for MDA8 on the monthly basis in low photochemically active areas. It could have been problematic if we had studied the changes in diurnal shape of ozone due to covid-19.

- Table 2: please provide uncertainties if you provide numbers which you think could be used by regulatory agencies. There is ample evidence that these numbers should be taken with great caution. Or delete this table if you cannot estimate the uncertainty.

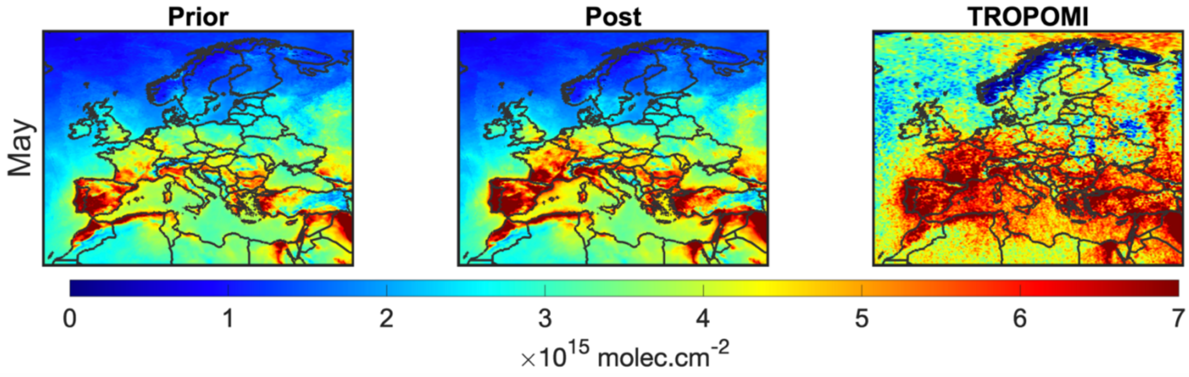
Using eq.4, we can provide errors for these estimates, but those are theoretical. To provide an exact error, we will need eddy covariance flux measurements. Therefore we removed the table.

- Seasonality of top-down VOC emissions in 2019: I reiterate my comment. The retrieved patterns indicate primarily anthropogenic emissions, with hot spots in Ruhr and Rhine Valleys, Southern Holland, London, etc. Those are not biogenic emission hotspots. The inversion system is simply unable to bring useful information on the emissions, except that the total VOC emissions were much higher than the prior (and than their 2020 counterparts) in Northern Europe (which is of course very obvious from TROPOMI).

The inversion system used for this study is based on a work presented in Souri et al., 2020. This system significantly improved HCHO over east Asia using the same model (WRF-CMAQ):



The inversion does not work in *mysterious ways*. The reason that we do not see such an agreement in this study is primarily because of large errors of TROPOMI in less photochemically active areas over Europe (please compare these columns to those in Europe) or the inability of model to provide sensitive Jacobians over certain areas (due to clouds, chemistry, and etc.). Another important caveat with comparing two datasets (here post vs TROPOMI) is that we must consider their variance. Pixels with higher uncertainty (larger variance) located over higher latitudes have a smaller weights in the comparisons.



As for the enhancements of TROPOMI HCHO (and VOCs) in some urban areas in 2020 with respect to 2019, we would need speciated VOC measurements to understand why. HCHO is too crude to determine the main reason. One hypothesis is that the number of VOC compounds in CEDS is too limited such that the model had to increase the anthropogenic emissions to compensate for. Another reason could be due to uncertainty in the yield of HCHO in CB06 mechanism. We added in the supplementary:

“However, the reason behind of the enhancement of VOCs over several urban areas such Paris and Po Valley is not fully understood. This can be caused by the errors in the chemical mechanism or the limited VOC compounds provided by the CEDS emission inventory.”

The reviewer provided very constructive and remarkable comments, which we have taken to heart; as a result, we believe our study has become stronger and are hoping this reviewer will find the manuscript merit publication in ACP.