Anonymous Referee #1

We thank the reviewer for his/her time and the thoughtful feedback. Below we list all referee remarks and suggestions (in *italics)* along with our responses.

Reviewer comment: *Keller et al. are investigating here the impact of COVID-19 restrictions on both NO_2 and O_3 surface concentrations. To estimate these changes taking into account the influence of the meteorology, they designed an interesting approach relying on a global simulation with the GEOS-CF model primarily bias-corrected using machine learning models. Compared to the recent studies covering this topic, the main strengths of this study are its spatial scale (since more than 5,000 stations in 46 countries are considered) and the fact that two important trace gases are included (NO_2 and O_3). The authors notably highlighted a strong variability of the NO_2 changes in general agreement with the level of mobility restrictions put in the different countries, while a lower response of O_3 is found. The paper is well written and relevant for our scientific community. It should thus be accepted after addressing one single major comment*
regarding the methodology and other minor suggestions.

Major comment :
- My major comment is about the machine learning methodology. Some information are missing or at least confusing, which explains why I classify it as a major comment but it might be only a minor one requiring only to provide more details in the text. My concerns are related to the way training and test datasets are obtained and how the machine learning models are tuned. Due to the substantial autocorrelation typically found in hourly air quality time series, using a random selection for splitting the datasets into training and validation data might lead to too optimistically good performances. For instance, the way I see it, if the model ingests a training data at a time t (and learn the corresponding model bias) and used for prediction at t+1, given that the model includes time features that allow to locate temporally this point, it will simply learn that around that time t, the model error is X, and then consider that the error at t+1 should also be close to X. In other words, the model might not learn properly the relationships between the model error and the features other than the temporal ones (most importantly, the meteorological parameters). In addition, it seems that no cross-validation (using for instance K-fold or time series cross-validation) is performed at any time (the word never appears in the manuscript), while this is important for tuning the models and ensuring robust estimates of their performance. Actually, it seems also that no tuning is performed during the preparation of the machine learning models. Also, there is often some confusion between the terms "training" (the phase in which you train your model), "validation" (the phase in which you tune your model and/or your select among different types of models) and "test" (the phase in which you evaluate the final performance of your final model already tuned). I guess what you mean here by "validation" is "test"? But since you are not mentioning how/if your models are tuned, it is quite confusing. Please clarify your methodology regarding these different points. On top of that, I agree with the comment of the editor that some discussion of the uncertainties of your approach should be included in the paper.

Author's response: We updated the machine learning methodology in the revised version of the manuscript and expanded its description. In the updated manuscript, all models have been trained using 8-fold cross validation, where the 8 batches represent quarterly chunks of model-observation pairs in order to minimize possible autocorrelation impacts. We also updated the notation of 'validation' and 'test' datasets.
These updates led to a slight deterioration of the model skill scores but have no discernible impact on the overall results and conclusions.

R: *Minor comments :*
- L73 : Which data availability at the hourly scale are you requiring for considering that a given day is valid? Please add this information
A: We only include days with at least 12 hours of valid data. We added this information to the manuscript.

R: *- Fig. 1 : Eventually, adding three panels zooming on North America, Europe and East Asia might be useful since the red points are completely hiding the blue and purple points in Europe and Asia (or if there are much less blue/purple points, you could plot them above the red points)*
A: We added figures with close-up maps of East Asia, Europe, and North America to the appendix.

R: *- L93 : You used OMI observations for scaling the anthropogenic emissions from 2010 to 2018. Why not scaling emissions up to 2019 included? Does the same procedure applied for 2019 highlights noticeable changes of NOx emissions between 2018 and 2019?*
A: We recognize that the initiali wording in this paragraph was misleading and we adjusted it in the updated version of the manuscript, with reference to the GEOS-CF description paper recently submitted for review (https://www.essoar.org/doi/10.1002/essoar.10505287.1).

R:*- L103 : Please indicate here that your XGBoost is making predictions at the hourly scale. Please also mention clearly that one XGBoost model is trained for each station, independently from the others.*
A: We added this information to the manuscript.

R: *- L110 : Which "mean" are you referring to here? The overall mean over the period 2018-2019? Or the seasonal monthly mean? Did you test applying the machine learning without removing outliers? Which performance is obtained ? Strongly stagnant conditions might lead to a peak of NO_2, and you want your model to learn this type of event, so at first sight, I don't understand why this step is needed (or even wanted). Please provide here a more complete justification of your methodological choices.*
A: The main motivation for this approach was to adjust for obviously erroneous observations, such as ozone or nitrogen dioxide concentrations of several thousand ppbv. Such values can occur in the OpenAQ database, whose values are reported in real-time and are not backfilled with quality-controlled data. To support this point, we performed two sensitivity simulations using more stringent thresholds of 3 or 4 standard deviations and did not find any change in our results.

R: *- L110 and Figure 2 : Please comment a bit more your results. Notably, I am wondering why your results are different at stations around #3000. Is this a specific region? Do you have any idea of the reason for that? Also, I am curious, why not simply normalising the RMSE by the average concentration? (rather than the range between 5th an 95th percentiles)*
A: We reordered the stations to reflect the four major regions considered in this study (China, Europe, USA, rest of the world). We chose the percentile window as the denominator for the NRMSE because it offers a better reflection of the concentration variability at the given site. The results using the RMSE normalized by the annual mean would look qualitatively very similar.

R: *- L116 : You mention 49 species and 31 modelled emissions : please provide the complete list of the species taken into account here (eventually in Supplementary Material or Appendix)*
A: The full list of input features is given in Table A2 in the Appendix.

R: *- Section 2.3 : Please indicate the features importance obtained with XGBoost, for both NO_2 and O_3. This is an information especially interesting in your study since you are using a lot of features, many of them probably not very useful for making the predictions (?)*
A: We added a new paragraph to the revised version of the manuscript, discussing the SHapely Additive exPlanations (SHAP) values for both the NO_2 and O_3 bias correctors in more detail. The SHAP values are similar to the 'classic' feature importance but better take into account the role of feature interactions. The distribution of all input feature importances is shown in the Figures A4 and A5 in the appendix.

R: *- L151 : Just to know, how these cities have been selected? Were they selected following an objective approach (for instance, all largest cities or cities with strongest data availability) or arbitrarily?*
A: We chose these 5 cities rather arbitrarily. Wuhan, Milan and New York represent early outbreak 'hotspots' that received a lot of media attention, and Taipei and Rio de Janeiro offer examples of different government responses to the pandemic (as also reflected in the data). We provide more detail on our motivation for showcasing these 5 cities in the revised version of the manuscript.

R: *- Fig. 5 : It would be useful to indicate the number of stations included in each country (for instance in the title of each panel).*
A: We provide the number of sites in the inset of each figure.

R: *- L182 : I would expect that the machine learning model (trained with data from late 2018 to end of 2019) would learn the reduction of NO_2 associated to the Chinese New Year, but the results for 2019 presented in Fig. 5 suggest that it is not the case. Any idea of the possible reason for that ? Is it simply because no training data are available in the first part of 2018 ? In any case, this could reduce the trust we have in the prediction done in 2020, at least for this specific country and this specific period of the year. Maybe including a new input variable representing if a day is holiday or not could help solving this issue.*

A: This is an excellent comment and the idea to add holidays as an additional input feature is intriguing (albeit somewhat cumbersome to implement on a global dataset!). We are actually quite happy to see that the model did not learn the NO_2 reduction associated with Chinese New Year, as such a behavior in our eyes would indicate a possible overfitting. Rather, we hope to capture the 'regular' model bias with the machine learning models and accept the fact that unusual events such as holidays cannot be captured. We feel this is the more conservative approach, especially since for China, we would have only one holiday to train the model on (year 2019).

R: *- The authors evaluated their machine learning models by checking the mean biases, errors and correlations over the entire period, which is fine for the analysis conducted in Sect. 3.1. The analysis of the diurnal variations of O3 and Ox is interesting but should come with an evaluation of the performance of the machine learning models at the diurnal scale : does the bias of the machine learning models show any diurnal variability ? I think it is important to show (eventually in the Supplement) and discuss a diurnal plot of the bias (similar to Fig. 8a) for both training and test datasets, just to ensure that the very small mean biases obtained over the entire period do not hide error compensation of stronger biases during specific times of the day.*
A: We added the hourly skill scores of the test data set in the appendix, and also note it in the discussion of the results. Note that the skill scores for the training and validation data show the same indifference to the time of the day.

R: *- L205 : I am not sure we can consider that the stronger seasonality of O3 compared to NO2 would bring more challenges since the seasonality is expected to be taken into account by the "month" feature. However, prediction business-as-usual O3 is possibly more challenging than for NO2 due to the more complex processes driving its concentration (e.g. secondary pollutant produced by more complex chemical reactions, involving more numerous precursors, potential strong influence of dry deposition, long-range transport)*
A: We agree with the reviewer and changed the wording in the updated version of the manuscript to reflect the fact that compared to NO_2, O_3 concentrations are much more influenced by large-scale processes and the local O_3 signal is thus expected to be much smaller.

R: *- L225 : The results shown for Ox are based on a subset of stations where both NO2 and O3 collocated measurements are available? If yes, is the same subset used for showing the results of O3 alone? Please clarify this point. In any case, it would be nice to have both O3 and Ox on a similar subset of stations to allow fair comparisons.*
A: The analysis is indeed based on the subset of stations where both NO_2 and O_3 observations are available. We clarified this in the manuscript.

R: - L262 : "natural background NO2"
A: We changed the wording as suggested.

R: *- L260 : Is this value of 80% obtained at global scale? How variable it is spatially (and more specifically, from one country to the other)?*

A: The 80% average sensitivity is the global mean value over the simulated sensitivity period (Dec-Jun). For the emission calculation, we updated the methodology and now use a variable NO_x/NO_2 ratio, depending on the inferred (percentage) NO_2 decrease. We acknowledge that this is still a simplification as the NO_x/NO_2 ratio is variable in both space and time. To account for this, we assign a rather large (absolute) uncertainty of 15% to the NO_x/NO_2 sensitivity ratio. We updated the manuscript, figures and tables accordingly.

R: *- L263 : Why using EDGAR[2015] rather than HTAP[2018] ?*
A: We chose EDGAR over HTAP because it's baseline inventory is more up-to-date (2015 vs. 2010). We added this information to the manuscript.

R: *- Fig. 9a : Results shown in Fig. A7 figure are not exactly what I would expect and thus deserve more discussion, highlighting more clearly the potential uncertainties. For instance, if I understand correctly, NO_2 emissions (estimated using the OMI NO_2 tropospheric column taken here as a proxy of the NO_x emissions) would have decreased more strongly during February 2020 (before the lockdown) than in March-April (during the lockdown), which is likely not true. A potential issue I see here is that the authors are not taking into the influence of the meteorology on the NO_2 tropospheric columns.*
A: The main goal of the sensitivity simulation was to obtain NO_x/NO_2 sensitivity ratios for a wide variety of (realistic) emission changes. Rather than using a fixed NOx emission ratio (as e.g., done in Lamsal et al., 2011), we chose to use the OMI NO_2 tropospheric columns as a proxy for emission changes. This is an obvious oversimplification but serves the stated goal of the sensitivity simulation. We clarified this aspect in the updated version of the manuscript.

R: *Also, the final number of 5% of reduction of global NOx emissions should also be discussed. Is it consistent with what we could expect during that COVID-19 period, namely a strong reduction of traffic emissions? (what is the contribution of traffic to global NOx emissions?). Therefore, please discuss in more detail this section.*
A: Traffic emissions are approximately 27% of total anthropogenic NO_x emissions. Using this information, we estimate that our derived NOx emission reductions correspond to 17-24% of global traffic emissions. We added a discussion on this to the manuscript.

R: *- L305-317 : This paragraph corresponds to a new analysis and should thus be included in a dedicated section rather than in the conclusion. Also, a more detailed information should be provided regarding this work. The authors say "we assume a sustained reduction in global anthropogenic emissions of NOx, CO and VOCs". Which reductions were used for CO and VOC emissions ? Also, given that the estimated reduction of NOx emissions is highly variable in time (Fig. 9b), to what "sustained" corresponds here?*
A: We moved this analysis to its own paragraph (Section 3.4.) and added more detail on the methodology of this sensitivity experiment. The emission reduction used for the forecast simulation was fixed at -20%, i.e., assuming no variability in time. This is an obvious simplification but serves the stated purpose of the sensitivity experiment.

R: *- Figures in Appendix : Please increase the resolution of these plots.*

A: We changed the layout to 4 panels per column to increase the resolution.