Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2020-685-RC1, 2020

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Global Impact of COVID-19 Restrictions on the Surface Concentrations of Nitrogen Dioxide and Ozone" *by* Christoph A. Keller et al.

Anonymous Referee #1

Received and published: 14 October 2020

Keller et al. are investigating here the impact of COVID-19 restrictions on both NO2 and O3 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 (NO2 and O3). The authors notably highlighted a strong variability of the NO2 changes in general agreement with the level of mobility restrictions put in the different countries, while a lower response of O3 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.

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 - 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) - 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? - 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. - 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 NO2, 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. - 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) - 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) - Section 2.3 : Please indicate the features importance obtained with XGBoost, for both NO2 and O3. 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 (?) - 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? - Fig. 5 : It would be useful to indicate the number of stations included

in each country (for instance in the title of each panel). - L182 : I would expect that the machine learning model (trained with data from late 2018 to end of 2019) would learn the reduction of NO2 associated to the Chine 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. - 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. - 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) - 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. - L262 : "natural background NO2" - L260 : Is this value of 80% obtained at global scale? How variable it is spatially (and more specifically, from one country to the other)? - L263 : Why using

EDGAR[2015] rather than HTAP[2018] ? - 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, NO2 emissions (estimated using the OMI NO2 tropospheric column taken here as a proxy of the NOx 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 NO2 tropospheric columns. 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. - 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? - Figures in Appendix : Please increase the resolution of these plots.

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2020-685, 2020.