Referee #3

We thank Kirsten de Nooijer for the time taken to review this paper and for the thoughtful feedback. Below we list all referee remarks and suggestions (in *italics)* along with our responses.

Reviewer comment: *Major comment 1): Statistics need to be included. The aim of the paper is to quantify, uncertainties should be quantified as well. The reported numbers are easily disregarded without the proper statistics (e.g., p-values, t-tests or z-tests) and uncertainty ranges. This problem is present in figures 4 and 6 (but applies for all given emission changes in the manuscript). Here the difference between the BCM prediction and observations is shown, but without noting whether this difference is significant (or perhaps falls within the uncertainty range of the BCM prediction).*
Author's response: We updated the uncertainty estimation based on model-observation comparisons on the test dataset, and propagate the estimated uncertainties per location site to a city and country level. The numbers in the updated version of the manuscript include the estimated uncertainties. In addition, we highlight statistically significant concentration changes in the concentration tables provided in the Appendix (Tables A3-A8), using a (stringent) p-value of 0.001.

R: *Example 1, line 145: "For Wuhan, we find a reduction in NO_2 of 60% relative to the expected BCM value for February and March 2020, and similar decreases are found over Milan (60%) and New York (45%) starting in mid-March and lasting through April (Fig. 4; Tables A1-A3)." How certain are these numbers? Is it between 62% and 58%? Or between 70% and 50%? I urge the authors to please quantify the uncertainty of these numbers by providing uncertainty ranges or mentioning of significance. This could be implemented similar to Le Quéré et al. (2020), here reductions in emissions are provided by stating the range (representing ±1σ) instead of a single number.*
A: The estimated uncertainty ranges are provided in the updated version of the manuscript. The stated uncertainties take into account the number of observation sites, so that estimates that are based on fewer sites result in higher uncertainties (all else equal).

R: *Example 2, line 228: "Compared to the BCM model, there has been an increase in the concentration of night time O3 (midnight-5.00 local time, Fig. 8a) by 1 part per billion by volume (ppbv = nmol mol-1) compared to the BCM, whereas Ox shows a decrease of 1 ppbv (Fig. 8b)." Is this reported 1 ppbv difference significant? I highly suggest you to report whether the modelled change is significantly different from the observations. The recent paper by Liu et al., (2020), also referenced by Keller et al., does report significance and thereby makes a more compelling case. Liu et al. derived uncertainty from 10000 Monte Carlo simulations from monthly statistics to estimate a 68% confidence interval. This procedure could be followed here as well. Another suggestion is to provide a paragraph on uncertainty estimation for the machine learning algorithm in the method section, similar to Petetin et al. (2020). Perhaps here the method of Hengl et al. (2017) could be useful. They describe a procedure for machine-learning uncertainty estimation with the use of the program R and the package 'xgboost'.*

A: We added a section on the uncertainty estimation to the methods (Section 2.3.4) and use these uncertainties to quantify the significance of our findings. Based on this, we conclude that the 1ppbv change is indeed statistically significant, and we now state so in the manuscript.

***R:** - Major comment (2): It's unclear how numbers in the result section are constructed from the represented data, no calculation steps are mentioned in the method section. Most importantly, how is the reduction in global NOx emissions of 2.9 TgN calculated?*
A: We revisited the description of the emission calculation, offering much more detail on the methodology to hopefully make it easier to follow.

***R:** Line 253 states the following: "This results in anthropogenic emission adjustment factors of 0.3 to 1.4 (Fig. A7)." Because of the lack of clarification on calculation steps or argumentation, it is unclear how the adjustment factors of 0.3 and 1.4 are determined. Is perhaps the approach of Mendoza & Russel (2001) used to derive adjustment factors for NO_x emissions? Please refer to the used methodology or provide the calculation steps. The in the manuscript referred figure A7 does not provide the calculations either (even though this seems to be suggested). Figure A7 only shows the monthly average perturbations applied to the 2018 anthropogenic base emissions, ranging from 0.5 to 1.5. As a consequence, the resulting quantification of reduction in emissions loses credibility.*
A: As already stated above, we updated the description of the emission calculation and also adjusted the uncertainty estimation, which now includes uncertainties for both the estimated NO_2 reductions and the assumed NO_2/NO_x ratio. The emission estimates reported in the revised version of the manuscript now include these uncertainty estimates.

***R:** Lines 262-266: "Based on bottom-up emissions estimates for 2015 from the Emission Database for Global Atmospheric Research (EDGAR v5.0_AP, Crippa et al., 2018, 2020) and using a constant concentration/emissions ratio of 0.8 based on the best fit line obtained from the model sensitivity simulation (dashed purple line in Fig. 9a), we calculate that the total reduction in anthropogenic NO_x emissions due to COVID-19 containment measures during the first six months of 2020 amounted to 2.9 TgN (Fig. 9b and Table 2)."*
It is clear a calculation is performed, but not how. How is the quite important 2.9 TgN reduction in anthropogenic NOx emission due to COVID-19 containment constructed? The 2.9 TgN is not in the referred Table 2 nor in Figure 9b. I urge the authors to provide the taken calculation steps resulting in the (quite important) 2.9 TgN reduction in anthropogenic NO_x emission. This will improve the credibility of that given number.
A: The methodology to calculate the emissions is now described in much more detail, along with a discussion of the corresponding uncertainties.

***R:** - Major comment (3): The manuscript mentions 'lockdown' situations but does not provide a definition of 'lockdown'. The restrictions vary per country (Ravindran & Shah, 2020) and the definition will have consequences on changes in NO_2 emissions. Some countries only enforced restrictions based on time, while keeping most forms of transport, schools and business open. Others have been reported to only had restrictions for part of the country. Please provide a definition of 'lockdown'.*

A: A clear definition of lockdowns is indeed complicated by the various responses, often even within regions of a country. We provide the list of used lockdown dates in the Appendix (Table A2). In general, we emphasize that the main purpose of the lockdown dates are to guide the reader in the interpretation of the figures, rather than using them at 'face value' for statistical analysis. The interpretation of lockdown dates is further complicated by the fact that many countries issued 'soft' lockdowns before the official lockdowns, which already altered human behavior and resulted in a decrease in NO_2 concentrations in advance of the official stay-at-home orders. We discuss this problem in the newly added Section 2.4 in the manuscript.

***R:** Lines 156-162: For Taipei and Rio de Janeiro, the observations and the BCM show little difference (Fig. 4), consistent with the less stringent quarantine measures in these places. Other cities with only short term NO2 reductions of less than 25% include Atlanta (USA), Budapest (Hungary), and Melbourne (Australia), again correlating with the comparatively relaxed containment measures in these places (Fig. A1-A3). In contrast, Tokyo (Japan) and Stockholm (Sweden), which also implemented a less aggressive COVID-19 response, exhibit NO_2 reductions comparable to those of cities with official lockdowns (>20%), suggesting that economic and human activities were similarly subdued in those cities."*
This suggests that degrees of reduction in NO_2 emissions are linked to severity in measures taken by local governments (e.g. lines 156-162), however, the severity of measures per country are not characterised. I suggest providing an overview of 'lockdowns' via a table including severity of measures and start and end dates. As an example, take a look at Ravidran & Shah (2020) where countries were classified on severity by introducing colour codes.
A: We added the lockdown dates used in this study to the Appendix (Table A2) but refrain from adding a lockdown severity measure because we don't feel comfortable with such a number on a country scale. For instance, how should one evaluate the severity of the lockdown for the United States where some cities (e.g. New York) were under a complete lockdown while other places saw little (official) restrictions? Rather, we emphasize in the manuscript that the main reason for adding lockdown dates is to support the visualizations.

***R:** Line 142: The start and end dates for these are from https://en.wikipedia.org/wiki/COVID19_pandemic_lockdowns or based on local knowledge."*
Because of Wikipedia's quickly changing contents, stating the start and end dates in a table will be an improvement on the derived results and will be more concrete than the stated 'local knowledge.
A: We added the list of lockdown dates to the Appendix and also provide the date at which the lockdown dates were accessed.

***R:** Lines 21 22: Reductions in NO_2 correlate with timing and intensity of COVID-19 restrictions, ranging from 60% in severely affected cities (e. Wuhan, Milan) to little change (e. Rio de Janeiro, Taipei)."*
Also, the manuscript mentions correlations in timing and intensity of COVID-19 restrictions and reductions in NO_2 (e.g. lines 21-22). A quantification of this correlation is however missing. Are these findings only based on eying the figures? Was a correlation test performed? I recommend adding quantification of the correlations.

A: We didn't mean to use the word correlation in the literal sense here, and recognize that its use was misleading. We changed the wording accordingly as we don't think that a correlation analysis of the derived concentration changes to the lockdown dates is scientifically warranted.

-Minor comments:
R: *Table 1: The links for AEROS (Japan) and EPA Victoria (Melbourne, Australia) do not work.*
A: We couldn't find any issues with the links but updated them again in the updated version of the manuscript.

R: *119: Provide an argumentation on why all observations below or above 2 standard deviations from the mean are removed, contrary to Ma et al. (2020) where observations below or above 3 standard deviations were removed.*
A: We updated the discussion about the removal of outliers (and its motivation), and also conducted two sensitivity studies using a threshold of 3 and 4 standard deviations, respectively. These sensitivity runs did not show any change in our results.

R: *Figures 2 and 3: The presentation of the machine learning statistics could be simplified in form of a table. I fail to see how the representation of the machine learning statistics in a graph are useful to the reader (including the location#, since no information is supplied to deduct which location# is which location). I suggest replacing figures 2 and 3 by a table providing statistical performance, similar to Table 4 of Ivatt & Evans (2019).*
A: We updated the figure so that statistics are grouped by region (as suggested by another reviewer), and discuss the statistics in more detail in the newly added Section 2.3.3.

R: *Figures 4 and 6: Reductions in % are difficult to read in the figures, one must go back to the text for the actual numbers. Consider including the numbers in the figures, so they stand stronger by themselves. Both figures could be shortened on the x-as as well, starting at 2019. The (incomplete) data from 2018 does not contribute to the results. I would even consider replacing both figures 4 and 6 entirely by new figures that better meet the objective of quantifying the difference in reductions of NO_2 and O_3 concentrations (including notification of significance or uncertainty ranges, see major comment 1).*
A: The percentage reductions are provided in the figures in the appendix as well as the tables, and the uncertainties are stated in the tables. The main objective of Figures 4 and 6 is to introduce the overall concept of our methodology and to show comparisons of observations and model values before and after the bias-correction. The time range 2018-mid-2020 is shown to highlight the full extent of the analysis data and to highlight how the model-observation comparisons look like for the entire previous time period (where available). Most other figures in the manuscript focus on relative changes derived from the bias-corrected model (e.g., the figures in the Appendix or Figures 5 and 7), and we find it important to show the full time series of the baseline model (as well as the effect of the bias-correction) for both O_3 and NO_2 in at least one figure.

R: *191: Consider replacing the vague terms 'some countries' and 'most countries'. These results are stronger when presented in numbers, for example: '42 out of 46 countries...'*

A: We updated this to ‘29 out of 36 countries…’.

R: *208: Reconsider the phrasing of this result. Belgium, Italy, Luxembourg and Switzerland do not all four show pronounced peaks in early April, based on Figure 7.*
A: We changed the wording in the updated version of the manuscript.

R: *221-225: Consider including chemical equations of the mentioned processes to improve readability of this paragraph.*
A: Detailed explanation of ozone chemistry, including the chemical equations, are provided in the references. We added an additional reference to Seinfeld and Pandis (2016) and also provide another reference for the NO_x/NO_2 ratio (Shah et al., 2020) in the updated discussion of the emissions calculation.

R: *252-258: Move this text to methods, it seems out of place here in the result section.*
A: We expanded the description of the sensitivity simulation, so that this paragraph now hopefully seems less out of context. We prefer keeping it in this paragraph so that the entire section stands on its own.

R: *305-309: Move this text to methods as well, it seems out of place here in the conclusion section.*
A: This is now discussed in newly added section 3.4.

R: *Figure 10: Consider moving this figure to the result section instead of below the conclusion.*
A: This figure is now discussed in the newly added section 3.4.