Reviewer 1

In this study, the authors did lots of work to estimate AOD changes during spring 2020, based on satellite remote sensing products, and to evaluate AOD responses to emission reductions in several CovidMIP models. Unfortunately, the different satellite instruments gave such a large spread in AOD changes, even with dust (as one of the natural species) excluded in the retrievals. Strong regional dependence of the robustness of observational estimates and model performance is found, but drivers behind this are unclear. The analysis of CovidMIP models does not add much to the literature, beyond the original CovidMIP paper (Jones et al., 2021) and other published studies. These are the major concerns leading to my hesitation to recommend the current manuscript for publication. The CanESM5 sensitivity tests are more interesting and potentially revealing. The novelty and science significance of this paper may be increased by focusing more on in-depth analysis of the sensitivity experiments on the roles of meteorological factors and possibly microphysical processes driving the response of aerosols to emission reductions in spring 2020.

We thank the reviewer for their comments and suggestions on our manuscript, and hope that the revisions presented here will satisfactorily address their concerns. We note that the title of this manuscript has been updated to emphasize the focus on using COVID-19 observations to evaluate ESMs, rather than on studying the response to COVID-19 in and of itself.

To address the comments presented above:

- Contribution to the literature:
    - We believe that this work fills a gap in the existing literature on aerosol changes during COVID-19. The majority of existing studies focus on observations or simulations, but not both. Those that do compare observed and simulated responses generally use the simulations to better understand the observed anomalies, e.g. by predicting control conditions in order to determine how much of the observed anomaly can be attributed to anthropogenic emission reductions. To the best of our knowledge, no studies have yet used the observed COVID-19 response to evaluate model simulations of an equivalent reduction in emissions.
    - Although the original CovidMIP paper (Jones et al. 2021) does present a first look at simulated AOD changes during COVID-19, our work extends on theirs in several ways. Perhaps most importantly, Jones et al. (2021) looked exclusively at the simulated response, and did not include any comparison with observations. Furthermore, they present AOD anomalies without any investigation into the drivers of these changes, as their main foci were (a) the presentation of the CovidMIP experiment itself, and (b) the radiative/climatic effect of the combined aerosol and GHG emission reductions. We have added a paragraph to the introduction discussing their work and our extensions on it (see more detailed comment below).

- Drivers of regional changes:
    - We agree with the reviewer's comment that both observational estimates and model performance vary from region to region, and acknowledge that our discussion of these drivers may have been unclear. We have removed lines 451-463 of the original discussion, which discussed differences between the observed datasets in the Northern

Hemisphere, as it confused the overall message. We have also elaborated on the differences in simulated dust-subtracted AOD anomalies in the Northern Hemisphere: we have added the sentence, *"The spatial origins of these overestimations differ: in MRI-ESM2-0, the Northern Hemisphere-averaged anomaly is due almost entirely to the strong negative anomaly over Asia; in CanESM5, ensemble-median anomalies are negative throughout the entire region (Supplementary Figure S8)."* (lines 353-356), and clarified our discussion on the potential impacts of differing trends between the observed and simulated datasets (lines 357-365). When combined with the other updates to the manuscript, we hope that the existing discussion on the regional differences in both observed response and model performance will now be more clear.

- CanESM5 sensitivity tests:
    - We are glad that the reviewer finds these sensitivity tests interesting.
      We have edited the abstract to emphasize importance of these sensitivity tests to our analysis: line 10, *"we systematically assess"* to *"we conduct a series of sensitivity tests to systematically assess"*
      In addition, we have highlighted the potential of conducting similar sensitivity tests in the other CovidMIP models to determine how representative the CanAM results were (line 400/566); such analysis is outside the scope of this work, but would be illuminating. It could also be interesting to conduct a similar analysis in an air quality model such as GEM-MACH, which has the capacity to simulate gas-phase chemistry and more detailed aerosol processes; however, such a test would primarily be interesting in the context of studying COVID-19 itself, and would not aid in the present goal of assessing the ability of Earth System Models to simulate a COVID-19-like emission reduction.
      Finally, we have run an ensemble of simulations (CanAM-old-emis) to investigate the relative contributions of emission inventory and model configuration on simulated AOD. The results of this analysis are included in the discussion and explored in more detail in Supplementary Material S3.

Below are a few more specific comments:

- The abstract lacks quantitative results either from the observational estimates or model analyses.
  We have updated the abstract to clarify that our statements are grounded in quantitative statistical tests: line 6 now reads, *"…most regions do not exhibit* ***statistically significant*** *changes…"* These statistical tests form the basis of our analysis, and are the main results upon which we report.

- Line 4 (and several other places): strictly aerosol optical depth rather than aerosol burden is estimated in this study, which should be made clear.
  We acknowledge that the original manuscript incorrectly referred to aerosol burden in a number of places rather than AOD, and we thank the reviewer for pointing this fact out. We have clarified the text both to emphasize that we are investigating AOD, not burden, and to elaborate

on our reasons for doing so.

- Line 39-41: This statement is inaccurate. Jones et al. (2021) did specifically compare regional AOD changes among the participating Earth system models.
  While we believe that our original statement was correct, we understand that it lacked sufficient detail and could be misinterpreted. We have added a more detailed description of the Jones et al. study to the Introduction, lines 71-76:
  "The models used in this work are taken from the COVID-19 Model Intercomparison Project (CovidMIP; Jones et al., 2021), which was developed to investigate the effects of a COVID-19-like reduction in aerosol and greenhouse gas emissions. Although Jones et al. (2021) present an initial analysis of changes in aerosol optical depth, their primary foci were the radiative and climatic responses to the COVID-19 perturbation, and the drivers of the simulated aerosol changes were not investigated. Our analysis provides the first detailed investigation of aerosol changes in the CovidMIP models, as well as the first comparison between observed and CovidMIP-simulated changes."

- Line 46: Please be more specific about what kinds of observed changes being used for model evolution purposes. Global or regional climate models use many different observational data for the evaluation purpose.
  We have updated the sentence to read, *"No studies have yet leveraged the observed aerosol response to the COVID-19 lockdowns for model evaluation purposes."*

- Line 61-64: depending on the purpose of obtaining aerosol concentrations, it can be a big problem of using a column optical property (AOD) as a proxy for aerosol concentration or aerosol burden mentioned in the first science question.
  Thank you for raising this concern. In the original text, these caveats were presented in the discussion, but we agree that they should have been presented in the introduction. In fact, there are reasons that the AOD is intrinsically interesting, and not merely an imperfect proxy for aerosol burden. We have updated the text to clarify our motivation for studying AOD, and to remove references to aerosol concentration/burden except as a possible avenue for future research.
  The text now reads, *"Both the air quality and climate impacts of a reduction in emissions depend on the response of atmospheric aerosol concentrations to emission changes, which is in general a complex and nonlinear dependence (Szope et al., 2021; Kroll et al., 2020). The climate effect further depends on the resulting changes in extinction, which can be quantified in terms of aerosol optical depth. In this work we consider changes in AOD, rather than concentration, since it is readily available from both model simulations and remotely sensed observations."*

- Line 78 (and section 2.1): It is too vague and generic to name meteorological conditions as one of the determining factors of AOD. In addition to the emissions, one should at least speak to the transport and sink terms of atmospheric aerosols such as dry and wet deposition in aerosol budget equation, although the detailed aerosol chemical and microphysical processes are sometimes even more important, depending on the aerosol types.
  We have expanded Section 2.1 to more clearly describe the processes that drive AOD variability.

We now discuss separately the factors that determine aerosol burden (emissions, secondary production, transport, and the effects of meteorology on lifetime via wet and dry deposition rates) and those that determine the AOD that results from a given burden (characteristics of the aerosol such as morphology and refractive index, and microphysical processes such as hygroscopic growth).

- Table 1: There might be too few models. Are they outliers among the 12 CovidMIP models?
  We thank the reviewer for their concern on this point. We would of course have preferred to include more models in our analysis, but were limited by data availability. However, the AOD anomalies simulated by these models span the range of AOD anomalies shown in the original CovidMIP paper, as well as the range of anomalies in downwards SW radiation flux, global surface air temperature, and global precipitation response. As such we feel that it is a representative sample.
  We have updated the text to include (lines 184-185), *"These six models, summarized in Table 1, sample the range of global AOD anomalies and climatic responses simulated by the full CovidMIP suite (Jones et al., 2021)."*

- Line 228-231: How was the first assumption tested?
  Section 4 (lines 245-289) has been revised to include an explanation of how this assumption was tested, and slightly reordered for overall clarity and readability. The text directly addressing this comment reads,
  "We assess the first of these assumptions by comparing observed and simulated variability over the reference period in our regions of interest. We calculate the variance of the region-mean AOD field over the reference period for each simulated ensemble member, the MMEc, and the observational datasets, and compare the spread in these estimates of variance. Although individual models may over- or underestimate the interannual variability in some regions, we cannot reject the null hypothesis that the MMEc and observations have the same interannual variability, based on a two-sided Welch's t-test. The only exception is for India, where the MMEc overestimates the variability in total and dust-subtracted aerosol optical depth; as a result, estimates of observational detectability will be conservative (i.e., anomalies are less likely to be found to be statistically significant). In a similar analysis of the variability over a longer baseline (2007-2019), using a subset of models for which these data were available, the total-AOD variability of the MMEc is consistent with that of the observations in all four regions."

- Line 284: This statement about contribution of dust to total AOD is inaccurate and can be misleading. It highly depends on season and region. Globally, dust contributes to less than 25% of annual total AOD.
  We thank the reviewer for this correction. We have conducted further analysis and, while the dust indeed does not dominate the total AOD, it does dominate the variability in our regions of interest (see also Gkikas et al. (2022), Fig. 10;). The text has been updated accordingly. Lines 324-326 now read, *"We next investigate the AOD signal when the contribution from mineral dust has been removed. In our regions of interest, the variability in total aerosol optical depth is dominated by the variability in mineral dust, which was not directly impacted by COVID-19 lockdowns."*

References:

Gkikas, A., Proestakis, E., Amiridis, V., Kazadzis, S., Di Tomaso, E., Marinou, E., Hatzianastassiou, N., Kok, J., and Garcia-Pando, C. P.: Quantification of the dust optical depth across spatiotemporal scales with the MIDAS global dataset (2003–2017), ACP, 22, https://doi.org/10.5194/acp-22-3553-2022.

Reviewer 2

The manuscript "How well do Earth System Models reproduce observed aerosol changes during the Spring 2020 COVID-19 lockdowns?" use the COVID lockdown and the following emission reduction for model evaluation. Modelled changes in aerosol optical depth (AOD) due to COVID restrictions and satellite retrieved AOD in March, April, May (MAM) 2020 are compared. The Earth System Models and observations show consistent results in Europe and India, where India is the only region considered with a significant reduction in AOD in MAM 2020 in the observations. In China and Northern Hemisphere as a whole, the modelled reduction in AOD is overestimated. Using one model, a systematic assessment of the influence of meteorology, baseline emissions, size of COVID emission reductions are done. The spread in the observations of AOD is a limiting factor of further constraining the models.

We thank the reviewer for their very helpful comments on this manuscript. Our responses to their specific suggestions are provided below. In addition, we wish to note that the title of this manuscript has been updated to emphasize the focus on using COVID-19 observations to evaluate ESMs, rather than on studying COVID-19 in and of itself.

The manuscript is well structured and presented, and I have only a few comments.

- The uncertainties in the satellite AOD products precludes a further constraint on the models responses to emission reduction. As this method outlined here, also can be used to evaluate response to future emission reductions, a bit more on future direction of satellite AOD evaluation would have been good.
  This is an excellent point. We have added to the conclusion (line 582-585), *"The substantial uncertainty in remotely-sensed observations of AOD precludes a detailed assessment of the relative biases in different models. As such, this analysis motivates future research into the drivers of the systematic biases in satellite retrievals of aerosol fields, particularly in the context of monitoring future emission reductions which are expected to take place over the coming decades."*

- L4: "observed regional aerosol burdens during" It is not aerosol burden that is assessed, but AOD (as a proxy for aerosol burden).
  We acknowledge that the original manuscript incorrectly referred to aerosol burden in a number of places rather than AOD, and we thank the reviewer for pointing this fact out. As detailed in our responses to Reviewer 1's comments on Lines 4 and 61-64, we have clarified the text both to emphasize that we are investigating AOD, not burden, and to elaborate on our reasons for doing so. References to aerosol concentration and burden have been removed from the text except as a possible avenue for future study.

- Consider swapping section 2.3 and 2.2?
  We have decided to keep the sections in their current order, but have expanded the text to clarify our logic for doing so.
  Lines 87-89 now read, *"We begin by highlighting the major considerations that need to be addressed in an analysis of this type: sources of AOD variability; differences that are expected to arise between simulated and observed AOD fields, no matter the quality of the atmospheric model or satellite retrieval; and finally, the impacts of observational uncertainty."*

We have also added the following sentence to the end of Section 2.1 (lines 111-113): *"In the following sections, we describe first the differences that would be expected even if both models and observations were perfectly accurate, and then the impacts of observational uncertainty."*

- Table 1: The mineral dust column can be misread as if models include mineral dust or not in the simulations. Maybe replace "Mineral dust?" by "od550dust" or "Mineral dust output".
  Thank you for pointing out that this is unclear. We have updated the column heading to read, *"Published od550dust?"* and in the caption specify that this indicates whether od550dust was available on ESGF.

- Section 3.2: Could be useful with a table of the satellite AOD products.
  Thank you for the suggestion. We have added a table summarizing the major features of the satellite AOD products.

- ACROS-C is used in Figure 1, but not mentioned in section 3.2.1.
  Thank you for catching this omission. ACROS-C has been added to Section 3.2.1.

- L253: "the Northern Hemisphere as a whole" From figure 1 and text elsewhere, the "as a whole" is not entirely correct as it is 0N-70N. Replace "as a whole" with (0N-70N).
  We thank the reviewer for pointing out this inconsistency. The text has been updated accordingly. In some cases, the phrase "as a whole" has been retained, but the latitudinal range has been added, e.g. *"When the Northern Hemisphere (0-70N) is considered as a whole, …"* to avoid making it sound as though we consider the Northern Hemisphere to be somehow separate from the other regions which are contained within it.

- Figure 1 (and 2 and 3) contain a lot of information. It could be useful to add more information to the legend, maybe first present what is included in the timeseries (the six models and the observations with symbol and black line). Then, as a separate box or just below, the 2020 values (Square: control, diamond: covid pert, MMEc). For MMEc maybe only show the square and not the line, as I was looking at the time series when I first looked at the plot. See also if filled, black outline, opaque/semi-transparent can be indicated inside the figure as well. I am not able to see if the results are plotted opaque or semi-transparent. Possible to use filled or not filled symbols instead?
  We thank the reviewer for their recommendations, and in particular for highlighting the challenge in differentiating between opaque and semi-transparent markers. We have made the following changes:
  - The statistical significance of the 2020 anomalies is now indicated by filled vs open markers, as opposed to opaque vs semi-transparent.
  - Simulated ensembles that are consistent with the observed ensemble are indicated by a black centre dot, rather than a black outline on the marker.
  - The legend has been split into two, with the second legend summarizing the formatting conventions of the 2020 points.

- Figure caption: Delete “horizontal offset for visual clarity” Already mentioned that the right side of the panel was for 2020 and “2020 values” are the titles of the subpanels.
  Updated.

- L313: It is hard by eye to see the difference in the trend between observations and models for the reference period (2015-2019).
  The language in this section has been softened to avoid claiming the existence of a trend; we agree with the reviewer that a difference is not clearly visible, especially given the variability in the observations. We do still include some discussion around the potential impacts of a difference in trends, since the raw data do suggest that such a difference *might* exist, and as discussed in Section 4 our use of the MMEc assumes that the observations and simulations have similar trends. As such it seems important to address the possibility. The original text from lines 312-319 has been replaced with, *“Given the short reference period and substantial interannual variability of the observations, it is challenging to identify whether the simulated trends are representative of those observed. In East China there is some indication that the models may simulate marginally more negative trends than the observations (more visible in the raw data shown in Supplementary Figure S6, and when the 2020 anomaly is not included in the timeseries), which could imply that the MMEc may inadequately represent the range of plausible “control observations.”' However, this discrepancy -- if it exists -- does not appear sufficient to explain the absence of a statistically significant anomaly in the observations. Visual inspection suggests that the observed 2020 anomalies are consistent with AOD excursions measured over the preceding 5 years, even taking any potential trends into account, whereas the simulations show an obvious decrease in 2020. This behaviour is particularly clear when considering the raw data shown in Supplementary Figure S6.”*

- L359-362: This was a bit unclear. Just note that the MMEc is as in Figure 2.
  Thank you for the suggestion. This paragraph has been removed, and the necessary information added to the caption of Figure 3. In fact, the MMEc is different in Figures 2 and 3 (it is drawn from only the models included in the figure) but as it turns out, the choice of MMEc does not change whether the observed anomaly is statistically significant; we agree that this was unclear in the original text.