We thank the 3 anonymous reviewers for their helpful comments on our manuscript. Their comments, and our corresponding edits and responses, are copied below.

Reviewer 1

I assume MAM stands for March-April-May. It is not defined in the paper.

Thank you for catching this omission. Line 254, the first time we use the acronym, now reads, *"Throughout, anomalies are calculated relative to the 2015-2019 March-April-May (MAM) mean."*

CanESM5 is noted as "CanESM50" in the plot legend of both Fig. 1 and Fig. 2. Is that a mistake?

We have clarified both the figures and text to be more explicit about model versioning. Both now refer to the model as "CanESM5.0" throughout the manuscript, except for occasional references to CanESM5.1 or CanESM5 in general as appropriate.

In the caption of Figure 4, please make a note of the model name used for the sensitivity experiments shown in this figure.

Thank you for noting this omission. The caption now reads, *"Effects of varying the magnitude of the simulated COVID-19 perturbation in CanAM5.1."*

Reviewer 2

L6: "forced with COVID-19-like reductions in aerosol and greenhouse gas emissions" The models were actually forced by greenhouse gas concentrations and aerosol and aerosol precursor emissions. Possible suggestion: "forced with COVID-19-like reductions in aerosols and greenhouse gases".

We thank the reviewer for catching this oversight. We have corrected both the abstract (line 2) and main text (lines 79, 169) as suggested. The only place where we still refer to "reductions in aerosol and greenhouse gas emissions" is in the description of the original CovidMIP methodology, which did estimate reductions in the emissions of both aerosols and greenhouse gases; as the reviewer points out, the GHG emissions were then converted to concentrations for use in model simulations.

Reviewer 3

Page 3, L60: how was COVID different than long term changes? You mention co-emission of species. Did COVID reductions change that? My guess is probably: energy emissions were the same, but transport emissions were reduced. Maybe discuss this?

Thank you for the excellent question. We have expanded on our reasons for being interested in a rapid change (lines 61-68):

There are both practical and scientific motivations for studying a rapid emission reduction. On the practical front, a short-but-strong signal is easier to disentangle from other sources of variability; we have continuous satellite observations that cover the entire study period; and the period is short enough

the instrument drift is unlikely to be a concern. Scientifically, model simulations indicate that the presence and severity of potential climate penalties, including changes in mean and extreme temperatures and precipitation, may be proportional to the rate at which emissions are reduced (Acosta Navarro et al., 2017; Hienola et al., 2018; Samset et al., 2020; Shindell and Smith, 2019; Shindell et al., 2012; Sillmann et al., 2013). Although we do not investigate the climate response to COVID-19 in this work, understanding the aerosol response itself is an important first step.

Page 4, L115: You might want to note explicitly that satellite retrieval error and model error are two other significant error sources.

We agree with the reviewer that satellite retrieval error and model error are important sources of error; these are discussed in the following sections. This section specifically discusses the factors that would contribute to differences between observed and simulated AOD signals even in the absence of those key sources of uncertainty. We have revised the text to clarify the intent of this section.

Lines 94 to 96 now read, *"We begin by highlighting the major considerations that need to be addressed in an analysis of this type: sources of AOD variability; factors that contribute to discrepancies between simulated and observed AOD fields, no matter the quality of the atmospheric model or satellite retrieval; and finally, the impacts of observational uncertainty."*

Section 2.2 has been renamed, *"Differences between observed and simulated AOD in the absence of model error or observational uncertainty"*

Lines 122 to 129 now read, *"Even given a hypothetical model that perfectly simulated atmospheric aerosol processes, and perfectly accurate satellite retrievals, differences would still arise between the observed and simulated AOD fields. These differences can be grouped into three main categories. First, a freely running model would produce a different realization of meteorological conditions than occurred in the real world, and so aerosols would be subject to different emission, transport, and depositional processes. Second, any errors in the model inputs (e.g. in the size of perturbation applied to represent COVID-19) would translate into biased simulations. Finally, the simulated and observed AODs would be recorded with different spatiotemporal sampling. Before any differences between the observed and simulated responses to COVID-19 can be attributed to model biases, then, these factors must be accounted for. (We discuss the role of observational uncertainty separately, in Section 2.3.)"*

Page 6, L170: How accurate was the 2 year blip assumption? Can you compare it to mobility data until now? Since it sounds like you are going to conclude emissions matters, maybe you can show the assumptions versus actual observations/inventories?

Assessing the accuracy of the 2-year blip assumptions is beyond the scope of this analysis, and up-to-date aerosol emission inventories, such as those from CEDS or ECLIPSE, are not yet available for 2020. Although regional assessments have been attempted by other researchers, a global evaluation of the scenario is not yet available (Forster et al. 2023). However, we assess the degree to which uncertainties or errors in the two-year blip assumptions would impact our results in Section 5.2.3.

Page 7, L186: Is Sigmond et al the reference for the CanESM5 AOD spread or do you need to show a figure?

Yes, Sigmond et al is the reference for the CanESM5.0 AOD spread; see in particular Section 5.1. This paper has now been published in GMD and the citations updated accordingly.

Page 10, L260: I don't think using the models to define observational anomalies is really appropriate. Why don't you use the pre-covid period and it's variability to define the mean and variability for the t-test?

We appreciate the reviewer's concern over this approach. To clarify, the anomalies themselves are calculated with respect to the mean over the 2015-2019 reference period. Unfortunately, using previous years' observations to determine the AOD expected in 2020 in the absence of the COVID-19 perturbation (an observational "control") is not practical. We have clarified the motivation for our approach in the manuscript; lines 261-267 now read,

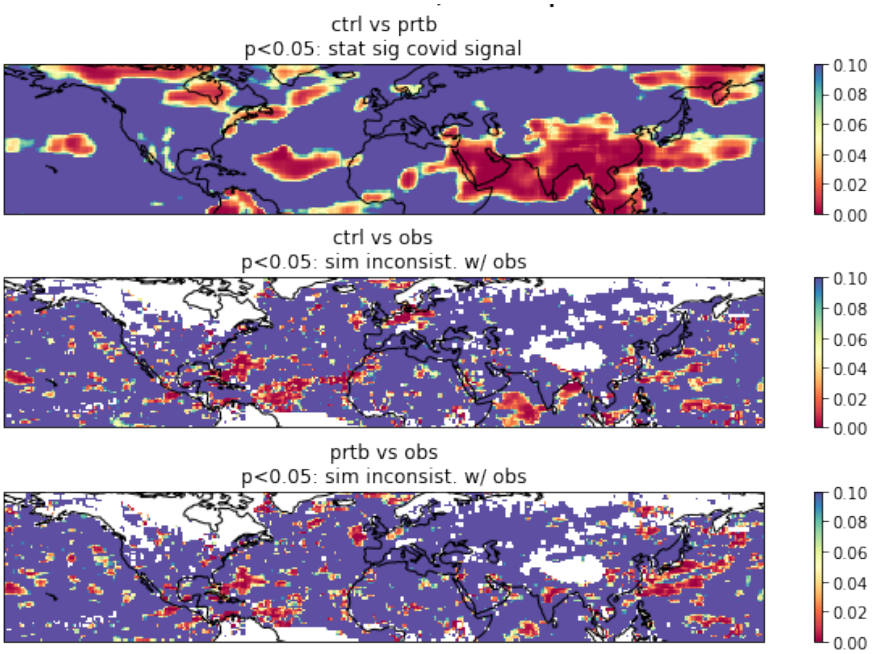
Defining detectability for the observations is more challenging, because there is no ``control observation'' from which to estimate the AOD that would have been measured in 2020 had COVID-19 not occurred. It would not be sufficient to use the mean and variance calculated from the reference period as a control: using the mean would neglect the impacts of underlying trends in the emissions, and a five-year reference period is too short to provide a robust estimate of the variance. Instead we borrow an approach from the field of detection and attribution (Eyring et al., 2021) and compare the ensemble of observed anomalies to a multimodel control ensemble (MMEc) constructed by randomly drawing an equal number of control simulation anomalies from each model.

As described in lines 270-278, this approach is supported by our comparison between the observed and simulated data over the reference period, as there is not a statistically significant difference between the interannual variability of the MMEc and the observations.

Page 10, L280: are you doing this at each point or some global mean? I hope it is at each point: there is information in the pattern.

We are doing this analysis using region-mean values, for each of our 4 analysis regions. We agree that a global-mean result would not be meaningful! Spatially-resolved t-tests do not change our results, and we have added a sentence clarifying this fact to the manuscript: *"We present results for t-tests performed on region-mean values; using spatially-resolved comparisons adds little information and does not change our results."* (line 289-291).

For the reviewer's interest, we have copied a figure showing the results of our t-tests for the model CanAM-new-emis over the Northern Hemisphere domain. In each panel, regions in orange/red have $p<0.05$, and blue/purple have $p>0.05$. The top panel compares control and perturbed ensembles, so regions in red exhibit statistically significant differences between the ensembles. As in the region-mean results, significant anomalies are found over India and China, but not Europe. The middle and bottom panels compare the control and perturbed ensembles respectively against the observations. Here, purple indicates areas of good agreement between the observations and simulations (where there is not a statistically significant difference between the two). White indicates regions of missing data. The spatial patterns shown in these panels add little information to our analysis, especially considering the fact that by definition, 5% of the region will have $p<0.05$.



Page 11, L300: note that MAM = March - May?

We thank the reviewer for noting this omission. Line 254, the first time we use the acronym, now reads, *"Throughout, anomalies are calculated relative to the 2015-2019 March-April-May (MAM) mean."*

Page 11, L304: define "detectable"

We have replaced the word "detectable" with "statistically significant."

Page 14, L370: Please clarify this is the same model as the CanAM5 in the previous plots (you corrected it for this paper).

We have revised the text to be more explicit about model versioning. The preceding analysis used the coupled earth system model CanESM5.0, and the sensitivity tests described here use the atmospheric model CanAM5.1. CanAM5 is the atmospheric component of CanESM5; the upgrades from version 5.0 to 5.1, described in Sigmond et al. (2023), resolve the dust issues described earlier in the manuscript.

Page 15, L389: I don't agree, it is in line with one of the data sets, and that is unchanged from the regular emission version given the spread.

We acknowledge that by eye, the updated-inventory results appear similar to those with the original CMIP6 emissions. Our statement refers to the results of our t-test: with the original emissions, the separation between the observations and the perturbed ensemble was statistically significant, but with the updated emissions, the separation is not statistically significant. We have clarified the text to emphasized that our comparisons with the observations are grounded in statistics (and to clarify where our discussion is qualitative). Lines 399-403 now read:

The effects of this reduction vary: in East China, the update is sufficient to bring the simulated COVID-19 signal into agreement with the observed anomaly ***(i.e., the separation between observed and simulated***

ensembles is no longer statistically significant), *whereas in the Northern Hemisphere the simulated response is brought somewhat closer to the observations but t****he difference remains significant****. In Europe the main effect of the update is to remove the simulated trend over the 2015-2019 reference period, bringing the simulations into* ***qualitatively*** *better agreement with the observations over this period…*

We have also clarified these comparisons earlier in the results section. The figure legends have been updated so that the black dot used to denote agreement between the simulated and observed ensembles is now labeled *"sim-obs dif. not stat. sig.,"* and the caption to Figure 1 now reads,

"*Black dots indicate simulated anomalies that are not significantly different from the ensemble of observed anomalies."*

Lines 322-324 have been updated to read,

In East China, four of the six models exhibit a statistically significant COVID-19-perturbed anomaly, which in general appears to be over-estimated: in these models, there is a statistically significant between the observations and the perturbed ensemble, but not between the observations and the control.

Page 15, L397: you just said Europe is improved with the new emissions. Be consistent or more precise.

We thank the reviewer for pointing out this discrepancy, and agree with their assessment. Lines 407 to 411 now read, *"These results suggest that in India, inferences drawn from the other CovidMIP models will likely not be affected by biases in the underlying baseline inventory as long as the applied perturbation is realistic. In the Northern Hemisphere, East China, and Europe, the apparent overestimation of the COVID-19 response identified in CanESM5.0, MIROC-ES2L, and MRI-ESM2-0 may have been partially caused by overestimates of the control emissions and thus of the absolute magnitude of the COVID-19 disruption."*

Page 16, L417: for the 2020 covid decrease. It does NOT improve earlier agreement.

We have clarified the text such that lines 44-52-453 now read, *"… the 2020 perturbed anomaly is in excellent agreement with the observations, …"* See also our more comprehensive revisions, discussed under your comment on Page 17 L43.

Page 17, L420: Again, I don't agree with this interpretation. What about anomalies in 2017 and 2018? Please be specific here.

This statement was intended to refer to specifically to the anomalies in 2020. We recognize that it was unclear, and have revised the text to be more explicit about our interpretation. In the process, we have substantially revised Section 5.2.2, and the paragraph referred to in this comment has largely disappeared. Please refer to the following comment for a description of the changes made.

Page 17, L431: but the inter-annual variability is not reproduced. the observations tend to have the same sign of inter-annual anomalies and the nudged model does not. Please explain how this is consistent?

We have substantially revised Section 5.2.2 to clarify the interpretation of the nudged results, in the context of differences in interannual variability between the different datasets. In particular, we explicitly address the higher variability of the nudged simulations (lines 425-429):

In general the nudged simulations exhibit higher interannual variability than the free-running simulations do, although this variability generally falls within the envelope of the larger coupled ensemble. In Europe, the nudged ensemble exhibits substantially higher interannual variability than do either of the free-running ensembles or the observations, indicating that the model may underpredict the variability of and AOD sensitivity to temperature, winds, and/or humidity in this region.

We address differences in the sign of the anomalies in Europe (lines 453-457):

The pattern of positive and negative anomalies is consistent with, although substantially amplified relative to, observations from MODIS-Aqua; however, this pattern of variability is not reproduced in the two CALIOP-derived datasets. The simulated COVID-19 perturbation is small relative to this variability and, as in all previous ensembles, the control and perturbed 2020 anomalies are both consistent with the observed ensemble.

We soften our description of the nudged results in India (lines 449-452):

Because the observed interannual variability is reproduced less well in the nudged than the free-running simulations, detailed comparisons between the observed and simulated 2020 anomalies would not be robust; however, it is reassuring to note that the 2020 perturbed anomaly is in excellent agreement with the observations, and the results of our statistical comparisons remain unchanged.

We discuss the variability of the different datasets when summarizing results for East China and the Northern Hemisphere (lines 439-441 and 458-460 respectively):

In East China, both CanAM-new-emis-free and CanAM-new-emis-ndgd simulate lower interannual variability than do the observations; the pattern of variability is perhaps somewhat improved in the nudged ensemble.

When averaged over the entire Northern Hemisphere (0N-70N), neither the free-running nor nudged ensembles reproduce well the observed pattern of interannual variability. Overall, our results here are unaffected by nudging: the control ensemble is consistent with the observed anomalies, and the perturbed ensemble values are too negative.

Finally, in lines 461-464 (the original focus of this comment) we have provided more support for our argument that improvements in the observations are necessary:

In all regions, the separation between the control and perturbed CanAM-new-emis-ndgd ensembles is on the order of the spread in observational estimates. Furthermore, in two regions (India and Europe), the three observational datasets disagree on the pattern of positive and negative anomalies throughout the reference period. These findings taken together suggest that further observationally-based model evaluation may not be feasible given current observational uncertainties.

Page 18, L465: in the previous section you implied that nudging did improve agreement (please explain or make consistent).

We hope that the revised Section 5.2.2 will address this concern, as we are now more explicit about the areas in which nudging did and did not improve the agreement between observations and simulations.

References

Eyring, V., N.P. Gillett, K.M. Achuta Rao, R. Barimalala, M. Barreiro Parrillo, N. Bellouin, C. Cassou, P.J. Durack, Y. Kosaka, S. McGregor, S. Min, O. Morgenstern, and Y. Sun, 2021: Human Influence on the Climate System. In Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 423–552, doi: 10.1017/9781009157896.005.

Forster, P.M., C. J. Smith, T. Walsh, W.F. Lamb, R. Lamboll, M. Hauser, A. Ribes, D. Rosen, N.P. Gillett, M.D. Palmer, J. Rogelj, K. von Schuckmann, S.I. Seneviratne, B. Trewin, X. Zhang, M. Allen, R. Andrew, A. Birt, A. Borger, T. Boyer, J.A. Broersma, L. Cheng, F. Dentener, P. Friedlingstein, J.M. Gutiérrez, J. Gütschow, B. Hall, M. Ishii, S. Jenkins, X. Lan, J.-Y. Lee, C. Morice, C. Kadow, J. Kennedy, R. Killick, J.C. Minx, V. Naik, G.P. Peters, A. Pirani, J. Pongratz, C.-F. Schleussner, S. Szopa, P. Thorne, R. Rohde, M.R. Corradi, D. Schumacher, R. Vose, K. Zickfeld, V. Masson-Delmotte, and P. Zhai.: Indicators of Global Climate Change 2022: annual update of large-scale indicators of the state of the climate system and human influence, Eart System Science Data, 15, https://doi.org/10.5194/essd-15-2295-2023, 2023.

Sigmond, M., Anstey, J., Arora, V., Digby, R., Gillett, N., Kharin, V., Merryfield, W., Reader, C., Scinocca, J., Swart, N., Virgin, J., Abraham, C., Cole, J., Lambert, N., Lee, W.-S., Liang, Y., Malinina, E., Rieger, L., von Salzen, K., Seiler, C., Seinen, C., Shao, A., Sospedra- Alfonso, R., Wang, L., and Yang, D.: Improvements in the Canadian Earth System Model (CanESM) through Systematic Model Analysis: CanESM5.0 and CanESM5.1, Geosci. Model Dev., 16, https://doi.org/10.5194/gmd-16-6553-2023, 2023.