In our response, reviewer comments are marked in bold, our responses and original text in plain text, and altered text in the paper in bold italic.

Response to reviewer 2 (Anonymous reviewer)

We thank the reviewer for their interesting and useful comments on our manuscript. Our responses to these comments are given below.

Minor Comment 1: p1 l29 "improvements in the physical realism"... I don't think Mann et al 2014 is the right citation at exactly this point.

We have changed this to "Although extensive improvements in the physical realism of aerosol-climate models have been made in recent years (Ghan and Schwartz, 2007***), resulting in a set of quite sophisticated models (***Mann et al., 2014)." Mann et al, 2014 is really the only reference for where a large set of microphysics models was assembled.

Minor Comment 2: p2 l2: "although the set of models is different to those used to assess aerosol microphysical properties in Mann et al. (2014)," not really an argument for the stubbornness of the ERF uncertainty, can be omitted here.

We have deleted that sentence. It wasn't really an argument, but really just a reminder that we should not compare these two inter-comparisons (Mann et al, 2014 microphysics models and Boucher et al, 2013 climate models) – they are completely different models.

Minor Comment 3: P2 l17 I think this paragraph and equation is misleading in pretending "that the forcing depends on the interlinked sensitivities of aerosols, clouds and their radiative properties to changes in aerosol emissions". Direct radiative effects, fast adjustments are not readily folded in into this equation. Please rephrase.

We are referring only to the aerosol-cloud forcing here. This equation is not pretending anything; it is the community's main approach to understanding how aerosol emission changes affect cloud properties. We have re-written the start of this paragraph (3rd paragraph in the introduction) on page 2 line 14 to clarify this applies only to aerosol-cloud forcing:

"There are three ways in which observations help to constrain the uncertainty in aerosol ERF. The first, ***which applies to the aerosol-cloud-related forcing,*** is based on..."

Minor Comment 4: P3 l23: "there is no equivalent to Equation 1 defining how a bias in simulated aerosol properties affects the forcing " => I think this is overly critical to bias inspections. An underestimate in fine mode AOD or bias in absorption can be translated in forcing bias. Measurements of fine mode AOD estimates can constrain anthropogenic AOD to some extent. And there might be other clever interpretations of bias. Please rephrase.

We disagree. In fact this is one of the main results of our paper: we show that aerosol-radiation interaction forcing (direct effect) is not strongly constrained by state variable measurements (AOD, etc.). There are many ways in which a model can be configured to get a particular AOD, but these

model variants (as we call them) predict very different forcings. To make this clear, and to signpost the result, we add at this point:

"(i.e., there is no equivalent to Equation 1 defining how a bias in simulated aerosol properties affects the forcing). One aim of our study is to make that link***, and we show in section 3.5.1 that observational constraint of many state variables only weakly constrains the direct and indirect radiative forcings.*****"**

Minor Comment 5: P3 l31 "Model variants that produce implausible results are rejected and, likewise, the forcings that they calculate are also rejected. " => would be nice to explain this at this point a bit more. Do you look at all observations at the same time? What is the criterion for rejecting?

The sentence referred to here is in the introduction section where we very briefly summarise / introduce the approach we take in this study. We therefore don't want to go into too much detail, as full details of the methodology and constraint approach are given in the following methodology section. Hence, we have only added very brief extra explanations of these points in this introductory paragraph/section of the manuscript.

In the paragraph on page 3, line 31 of the original manuscript (page 3, line 32 of the revised manuscript), we have added:

"... Model variants that produce implausible results ***(i.e., output outside of an observation's estimated uncertainty range)*** are rejected and, likewise, the forcings that they calculate are also rejected. A similar constraint methodology has been applied to..."

And we have added the following sentence to the end of this paragraph on page 4 line 3 of the original manuscript (page 4, line 4 of the revised manuscript):

"***We constrain using each aerosol/cloud observation individually and combinations of all observations.***"

We have then added more detail on our criteria for retaining/rejecting model variants in the constraint process in Section 2.7 (Identification of plausible model variants) of the methodology section, to make this process clearer. The start of Section 2.7 now reads as:

"Observationally plausible model variants are defined to be those that simulate aerosol and radiation properties within the uncertainty ranges of the observations, defined in Table 2. ***As we use statistical emulators to generate the simulated output values for each model variant, rather than using the climate model directly, an emulator prediction error φ (valued at one standard deviation on the emulator prediction from the Gaussian process uncertainty) is also taken into account. Hence, for a given observed variable, a model variant is rejected as implausible if the range defined by its emulator prediction +/- φ lies outside the corresponding observation's uncertainty range in Table 2. Furthermore, for a joint observational constraint we retain only the model variants that are classed as plausible for all individual observation types that make up the joint constraint.***"

Minor Comment 6: P5 l9 "The analysis is restricted to Europe for the month of July. We do this primarily because regional observations can provide a better constraint on model uncertainty than global mean observations... but with the disadvantage of being less straightforward to understand. . . . We choose Europe because there are many long-term measurements" => I don't buy these arguments. With synthetic observations this should not be a big problem to do globally. There are no long term measurements used. I assume this is done to save computer time. I think its ok to use just Europe and just July. But the discussion should be more honest and open here. Paragraph please rewrite.

There was no intension on our part to not be honest and open in terms of our arguments for only using observations over the Europe region in July for this study.

Our full reasoning to base the presented study on only Europe in July is as follows:

- Previous work (Regayre et al, 2018, for example) has shown that using global mean quantities for constraint can mask many compensating regional parameter effects, leading to a very weak 'watered down' constraint on both the parameter space and model outputs like forcing that can be difficult to interpret.
- To constrain forcing globally we need to constrain the parameters that affect the forcing across the globe. Regayre et al, 2018 show that different parameter sources control the uncertainty in aerosols and forcing in different regions. Therefore, the global problem essentially breaks down to be the sum of constraining the forcing in a set of key regions, of which Europe is one. The Europe region in July provides a single region/month for which a distinct set of parameter uncertainties affect ERF. If we cannot constrain the forcing regionally in Europe, then we are unlikely to obtain a constraint on a global/multi-region scale. Hence, Europe in July provides us with the full insight we aim for here on the potential of our approach.
- It is true that our analysis is highly computational and generates a significant amount of data. We have investigated other regions in this work, including China and the North Pacific (not shown), but including more regions in the presented study would only significantly expand the results in terms of quantity and complexity, with no real gain as to our actual aim of establishing the overall potential for constraint with our methodology.
- Real observations of multiple aerosol observable quantities are sparse in many regions around the globe, and temporally (Reddington *et al.*, 2017), but Europe is a region for which a diverse set of aerosol observations are available. These observations provide realistic estimates of observational uncertainty for the synthetic study.
- The presented study was a specific stage in our model evaluation work, at which we aimed to test our methodology for constraint using synthetic observations before moving forward to our now current work where we are looking at using real observations. Using Europe in July for our synthetic study supplies a good test for evaluating the potential constraint we may achieve from using the real observations in the future – a test that we are now working towards verifying.

We have edited the penultimate paragraph in the introduction section at page 5 line 9 of the original manuscript (page 5 line 11 of the revised manuscript) to better reflect these reasons (The start of this paragraph also contains revisions with respect to our response to Reviewer 3's minor comment 2):

"The analysis is restricted to the region of Europe (defined in this study by the longitude range: 12°W to 41°E, and latitude range: 37.5°N to 71.5°N) for the month of July. We take a regional approach primarily because regional observations provide a better constraint on model uncertainty than global mean observations (Regayre et al., 2018). The sources of uncertainty in aerosols and forcing vary regionally (Lee et al., 2016; Reddington et al., 2017; Regayre et al., 2015)***. Therefore,*** a global analysis would essentially be a scaled-up version of what we present here – i.e., a set of ***regional evaluations.*** We choose Europe ***in July as this is a region and month for which a distinct set of parameter uncertainties affect the aerosol properties and the ERF, providing a good test case for our methodology. Europe is also a region for which a diverse set of*** long-term measurements of different aerosol and radiative properties ***are*** available, ***that*** we can use to inform our assessments of the observational uncertainty."

Minor Comment 7: Chapter 2.1 and 2.2 and 2.3: I think they can be reversed. Some simple questions are not clear to me: Are the simulations global? Is it a one year simulation with a 4 month spinup (eg Sep-Dec of the preceding year) and is then just July analysed? Is the emulator producing global fields, from which data are sampled at European stations?

We have considered the reviewers suggestion to change the order of the sub-sections in our methodology section (Section2) of the manuscript. However, we think that the current order is most suitable, as we prefer to keep the overall summary of our approach at the start (section 2.1), with the different aspects further explained in the order they are mentioned in that summary.

We have clarified the model set up in the final paragraph of section 2.3 (page 9 lines 11-12 of the original manuscript; page 9 lines 15-17 in the revised manuscript):

"... In total 217 perturbed parameter simulations ***of the global model*** were run ***for a full year*** for each anthropogenic emission period (1850, 1978 and 2008 emissions). ***Each simulation had a spin-up period of seven months from a consistent starting simulation, where the parameters were set to their median values for the first four months and the perturbations then applied in the final three months***."

The emulators do not produce global fields. We have clarified this at the beginning of Section 2.4 (page 10, line 6 of the original manuscript; page 10 line 5 of the revised manuscript):

"For each model output (such as ***the regional mean ToA flux, CCN conc., etc. for Europe in July***) we construct a statistical emulator model over the 27-dimensional parameter uncertainty..."

Minor Comment 8: Page 5 counts 191 simulations, while page 9 counts "in total 217 perturbed parameter simulations". Better to harmonize numbers.

The reasoning for this difference is explained in the next sentence on page 9 (lines 12-13). We only use ensemble members that completed the full year of simulation in all periods, which reduces the number of runs used for analysis to 191 from the total of 217 that were originally run. We feel it is important that we are transparent about this and so we continue to state both numbers. We have amended the sentence at page 9 line 12 of the original manuscript (page 9 line 18 of the revised manuscript) to improve the clarity on this point:

"Twenty-five simulations did not complete ***the full*** annual cycle ***so were not used in our analysis. Consequently,*** the ensemble of simulations ***used for analysis*** for each period was made up of the remaining 191 simulations, all of which were used to build the final emulators."

Conclusions: I wonder how general the findings are if the ERF is in essence tested only over Europe and July with synthetic observations, but that might be shown in future publications.

The aim of this paper is to demonstrate the potential for model constraint using multiple observations. As we argue in the paper (and in our replies to the reviewer comments) a global analysis would essentially be a scaled up version of what we are doing here – i.e., constraint of global ERF will be dependent on the extent to which we can constrain the model in all the key regions. Global forcing is the sum of regional forcings, and each region has its own unique combination of uncertainties.

References:

Boucher, O., Randall, D., Artaxo, P., Bretherton, C., Feingold, G., Forster, P., Kerminen, V.-M., Kondo, Y., Liao, H., Lohmann, U., Rasch, P., Satheesh, S. K., Sherwood, S., Stevens, B. and Zhang, X. Y.: Clouds and Aerosols, in Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by V. B. and P. M. M. Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. 571., 2013.

Mann, G. W., et al.: Intercomparison and evaluation of global aerosol microphysical properties among AeroCom models of a range of complexity, Atmos. Chem. Phys., 14(9), 4679–4713, doi:10.5194/acp-14-4679-2014, 2014.

Regayre, L. A., Johnson, J. S., Yoshioka, M., Pringle, K. J., Sexton, D. M. H., Booth, B. B. B., Lee, L. A., Bellouin, N., and Carslaw, K. S.: Aerosol and physical atmosphere model parameters are both important sources of uncertainty in aerosol ERF, Atmos. Chem. Phys., 18, 9975-10006, doi:10.5194/acp-18-9975-2018, 2018.

Reddington, C. L., et al.: The Global Aerosol Synthesis and Science Project (GASSP): Measurements and Modeling to Reduce Uncertainty, Bull. Am. Meteorol. Soc., 98(9), 1857–1877, doi:10.1175/BAMS-D-15-00317.1, 2017.