Response to the referee comments for the manuscript: "Anthropogenic aerosol forcing - insights from multi-estimates from aerosol-climate models with reduced complexity" by Fiedler et al.

We thank the anonymous referees for their comments that helped improving the manuscript under discussion in Atmospheric Chemistry and Physics. Our main changes of the earlier version of the manuscript are:

1. An improved presentation of our motivation with a revised introduction and introductory statements in the sections,
2. More detailed explanations of our experiment, data and analysis strategy including improvements on statements on the reasons for computing the year-to-year variability in ERF as well as on model differences and similarities for improving the clarity of the text,
3. new appendices for documenting model differences in the representation of physical processes and simulated cloud properties for improving the coherence and reading flow of the manuscript,
4. And the extension of our model ensemble with the newly available EC-Earth experiments following our protocol.

Our replies are given in blue below the referee comments in black.

Anonymous Referee #1

This manuscript examines the radiative forcing of anthropogenic aerosols in simulations with a small set of global models following the protocol for the Radiative Forcing MIP now in progress as part of CMIP6. The RFMIP aerosol specification, on which the lead authors were also a co-authors, provides a description of the anthropogenic aerosol in purely radiative terms i.e. as those parameters that enter the radiative transfer equation, and as their differential impact to cloud droplet number. Having eliminated model differences in what the aerosols are, the authors examine here how other model differences impact the radiative forcing. This could be considered a prototype for studies that might be done with the larger collection of RFMIP results when these become available. The authors report on the inter-model spread in effective radiative forcing (ERF) at present-day, show differences in the present-day distribution of background clouds and aerosols, and examine how the shift in the aerosol distribution between the 1970s and present day has impacted the RF from anthropogenic aerosols. This work is potentially interesting but not yet mature enough to publish. The work lacks an explicit motivating question, in the absence of which the variety of results presented is hard to interpret coherently. Some results, especially the off-line radiation calculations and the cursory comparison of model clouds and droplet number to observations, seem especially unconnected to the rest of the material. There are important methodological errors in how ERF is computed and in how the set of simulations is conceived of. Important opportunities for deeper understanding are also missed, especially in making connections between the background state of each model and the resulting diversity of ERF from anthropogenic aerosols. It is understandable that the lead authors wish to exploit something from the experiments they have helped design. The scientific community will nonetheless benefit more from work that exploits the simulations to answer specific questions.

> Thank you for your comments. Our work can be seen as a pilot study for RFMIP, where models use the MACv2-SP parameterisation of anthropogenic aerosol optical properties and associated change in the cloud droplet number concentration for assessing model errors in radiative transfer. It is important to underline that we only unify the treatment of anthropogenic aerosol, i.e., the natural aerosol is still model-dependent.
>
> Our aim is an assessment of the impact of the spatial change of the anthropogenic aerosol between the mid-1970s and present-day as well as the role of model-internal variability with an ensemble of modern aerosol-climate models. We improve the presentation of our motivation and coherence of the analyses in the revised manuscript. For instance, we now state our research questions already in the second paragraph rather than at the end of the introduction. Please refer to our responses below for more details on the revision.

Structure and focus:

1. What question do the authors seek to address in this work? One possibility would be “to what extent is the signal from anthropogenic aerosol detectable against the background of uncertainty and natural variability?” (I understand this to be one of the motivating questions of RFMIP although progress could be made without using formal detection and attribution machinery). Another would be “how does the background meteorological and/or aerosol state affect the radiative forcing of anthropogenic aerosols?” In the absence of a clearly-articulated motivating question it is hard to know how to interpret results. One suspects that not all the material belongs in the same manuscript. If the goal is to understand the range of values of ERF that might be expected from the same aerosol across different models then the motivation for sections 3.3, 3.4, and 3.5 is unclear. If the question is understanding how background state affects ERF then substantially more work will be required to link the quite cursory characterization of differences across models to the spread in ERF. Neither of these questions would motivate the also-cursory comparison of models and observations.

> We have moved our motivation and research question to the beginning of the article. The revised introduction names the motivation and research questions in the first two paragraphs:
>
> “Despite decades of research on the radiative forcing of anthropogenic aerosol, quantifying the present-day magnitude and reconstructing the historical evolution of the forcing remains challenging. Recent work has indicated that natural variability affects estimates of the effective radiative forcing (ERF) of anthropogenic aerosol (Fiedler et al., 2017). More specifically, natural variability was identified as a cause for increases and decreases in the global mean ERF associated with the spatial change in anthropogenic AOD (τ_a) between the mid-1970s and mid-2000s. The anthropogenic aerosol pollution in the mid-1970s was herein larger in Europe and North America than in East Asia, whereas the opposite is the case in the mid-2000s. In addition to these regional changes in aerosol pollution, differences in the surface albedo, insolation, and cloud regimes between the aerosol transport regions of the Pacific and continental Europe may result in changes in the global ERF over time.
>
> In light of model uncertainties (e.g., Kinne et al.,2006, Quaas et al., 2009, Lohmann et al., 2010, Lacagnina et al., 2015, Koffi et al., 2016), a single model as used in Fiedler et al. (2017) does not necessarily represent the full spectrum of possible anthropogenic aerosol forcings. In the present study, we therefore revisit the question of Fiedler et al. (2017): "Does the substantial spatial change of the anthropogenic aerosol between the mid-1970s and mid-2000s, reflected by the change in τ_a shown in Fig. 1, affect the global magnitude of ERF?" using ensembles of simulations from five global aerosol-climate models with reduced aerosol complexity. In this context, we additionally ask: "What is the relative contribution of variability amongst and within models to the spread in ERF?", and document the model diversity for the pre-industrial aerosol and cloud characteristics that are relevant for ERF of anthropogenic aerosol. Such model differences have previously been identified for other climate models (e.g., Nam et al., 2012, Fiedler et al., 2016, Crüger et al., 2018).“

2. What is the intent of showing model-observation comparisons in section 3.3, or the offline radiation calculations in section 3.4? One might infer that the authors hope to address the ability to estimate real-world ERF from historical observations but this is not explained clearly.

> We show the observations as an orientation for realistic values for model validation. Please note that Section 3.3 has been moved to the appendix for improving the reading flow of the article.
>
> We state in the revised introduction: “We provide observational benchmarks for the inter-comparison of the complex models with satellite data and results from a stand-alone atmospheric radiation transfer model for quantifying differences in the instantaneous

radiative forcing (RF)", in Appendix B (former Section 3.3): "The model diversity in RF and ERF is larger when cloudy skies are considered. We therefore assess the model diversity in cloud properties and compare the models against observational climatologies from satellite products, (…). The observational products herein provide an orientation for realistic values, (…).", and at the beginning of Section 3.3 (former Section 3.4): "We use offline radiation transfer calculations for providing benchmarks for the instantaneous radiative forcing (RF) of the complex models. "

Methodology:
3. Effective radiative forcing relates long-term radiative perturbations and long-term response. It does not make sense to look at yearly averages. The protocol for CMIP and RFMIP, following doi: 10.1002/2016JD025320, is for 30-year simulations precisely to average out model internal variability.

> We agree, it is precisely one of our points and important for later ERF analyses from CMIP6 simulations, i.e., we need to average over sufficiently long time periods for estimating ERF of a model. Fiedler et al. (2017) discuss the precision of ERF estimates from one climate model that depends on the confidence level, the magnitude of model internal variability and the number of years for averaging. Here, we show that the year-to-year standard deviation in ERF is similar to the model in Fiedler et al. (2017), i.e., the precision estimates are applicable to the here-used models, too. Short model simulations covering a few years, like studies have done in the past, are not suitable for calculating ERF and can lead to misleading results. It is important to keep this in mind for diagnosing ERF in transient climate experiments. e.g., by following the RFMIP recommendation of using three member ensembles with ten-year averages for time-varying ERF estimates.
>
> In addition to our explanation in the last paragraph of Section 3.1, we now add in Section 2.2: "This approach is chosen for illustrating the effect of year-to-year variability on ERF estimates. (…) the RFMIP protocol recommends a thirty-year average for diagnosing the ERF of a model (Pincus et al., 2016)" and in the conclusion: "For instance, the protocol of RFMIP requests thirty-year averages for estimating the present-day ERF and three-member ensembles with ten-year averages for diagnosing decadal changes in ERF (Pincus et al., 2016)."

4. What motivates the use of multi-model means in 5-7, 9-10? An ensemble mean is the best estimate of the expectation value of some quantity when the samples are independent and uncorrelated, but this is unlikely to be the case in the small set of simulations here (or even in the larger collection to be collected through RFMIP).

> The multi-model mean is useful for comparing individual model results to the same reference. We add in Section 3.1: "For doing so, we first calculate the multi-model mean as a reference value."

5. Although the authors may well remove the comparisons to observations it is remiss to present inferences of drop number from satellites without mentioning the very many caveats around such estimates. See the careful review in doi:10.1029/2017RG000593.

> We agree that satellite retrievals are uncertain themselves and add in the Appendix (former Section 3.3): "The observational products herein provide an orientation for realistic values, although satellite retrievals also have caveats (e.g., Grosvenor et al. 2018)." The section on the cloud inter-comparison has been moved to the Appendix for improving the reading flow of the article.

6. Section 3.5 seems to illustrate that even a large spatial shift in aerosols has a relatively small impact on ERF. It's not clear why this bears mentioning - is there some surprise here? One might naively expect that the same aerosol burden would have roughly the same impact no matter where it was on the planet.

It is not obvious that the same change in global mean aerosol optical depth gives the same global ERF. We revise the introduction to make this clearer (refer to our reply to the first point). Additionally, we state at the beginning of Section 3.4 (former Section 3.5): “We assess the effect of a substantial spatial change of the τ_a maxima from Europe and the U.S. to East Asia between the mid-1970s and mid-2000s. One can additionally argue that the spatial differences in cloud regimes, insolation and surface albedo contribute to regionally different radiative effects resulting in a changing global ERF.”

Smaller points:
7. The word “comparably” is used incorrectly in several places in the manuscript. The authors likely mean “relatively.”

Replaced.

8. The introduction is so indirect as to be unclear. It would be better to start with motivating questions more specific to this study than “what is the anthropogenic aerosol forcing.”

We revised the introduction. Please refer to our reply to your first point.

9. Far more detail is provided about each model than is useful. The only details that are really needed are those that might have bearing on interpreting the results presented here.

We focus on model differences in the pre-industrial aerosol and clouds that are relevant to the results on radiative forcing. For the sake of brevity, we have moved the overview on the model physics packages to the appendix and refer to it in Section 2.2: “We therefore keep for instance the model diversity for the physical parameterisations of radiation and clouds (Appendix A)” and add in the same section: “All other aspects remain model-dependent, e.g., the treatment of the pre-industrial aerosol and clouds (Appendix A)” and describe the model differences for the pre-industrial aerosol optical depth in a new paragraph: ”We do not prescribe the same natural aerosol nor interfere with any other model components than prescribing the optical properties of anthropogenic aerosols and η_N. For instance, the pre-industrial aerosol optical depth (τ_p) depends on the model (Fig. 2 and 3). Regional differences occur primarily over oceans and deserts, where observations are typically sparse. It is herein noteworthy that ECHAM-HAM runs with interactive parameterisations for dust and sea-salt aerosol resulting in different spatio-temporal variability in τ_p (Fig. 3) compared to the monthly mean climatology MACv1 in ECHAM. In the interactive parameterisations, the natural aerosol emissions, transport and deposition rely on meteorological processes that are difficult to represent in coarse-resolution climate models, e.g., desert-dust emissions strongly depend on the model representation of near-surface winds (e.g., Fiedler et al., 2016) such that constraining the desert-dust burden remains challenging in bottom-up aerosol modelling (e.g., Räisänen et al., 2013, Evan et al., 2014, Huneeus et al., 2016). ”

10. The simulations run from 2000-2010 but are treated as a statistically homogeneous set. Is this fair? It certainly deserves from comment.

We add in Section 2.2.: “The first year of each 11-year run is considered as a spin-up period and is excluded from the analysis, thus all analyses are for the period 2001-2010. We have chosen the ten-year period for including variability in the boundary conditions."

11. In section 3,3 readers will appreciate a symbol for top-of-atmosphere shortwave cloud radiative effect that is not a capitalized version of the symbol for cloud fraction.

We remove the subscript in the symbol for the cloud fraction in the revised manuscript.

12. Do the conclusions in the last paragraph differ from the RFMIP protocol, or from community practice?

Past community practices partly differed from what is recommended in the RFMIP protocol and tested in the framework of our article. We have added: “The protocol of RFMIP requests thirty year averages for estimating the present-day ERF and three-member ensembles with ten-year averages for diagnosing decadal changes in ERF (Pincus et al., 2017)."

Anonymous Referee #2

The manuscript presents a 4-model ensemble assessment of simulation variability for anthropogenic aerosol radiative forcing simulations. The four models represent a reasonable (if small) cross-section of the global models available. My main comments are focused on improving the clarity of analysis and presentation.

Thank you for your comments. We now additionally include EC-Earth experiments for a larger ensemble of five complex aerosol-climate models. We have worked on the language and added details throughout the manuscript for improving the clarity. Please refer to our more detailed responses below.

13. The estimate of variability in ERF seems to be overestimated: it is based on differentiating the time-series of pre-industrial simulations from those with anthropogenic aerosols. Should not an average of the pre-industrial simulations be used for the differencing baseline to avoid this? This is relevant to the discussion of inter-model variability relative to natural variability as well.

We define variability in ERF internal to the models as year-to-year variability, i.e., we compute annual means of the radiation budget for determining ERF. We herein subtract years with identical boundary conditions in the simulation without anthropogenic aerosol from the simulation with anthropogenic aerosol for each model. Using a mean of just the pre-industrial simulation would compute a yearly anomaly that would be different from what we define here as year-to-year variability.

In addition to our explanation in the last paragraph of Section 3.1, we now add in Section 2.2: "This approach is chosen for illustrating the effect of year-to-year variability on ERF estimates. (…) the RFMIP protocol recommends a thirty-year average for diagnosing the ERF of a model (Pincus et al., 2016)" and in the conclusion: "For instance, the protocol of RFMIP requests thirty-year averages for estimating the present-day ERF and three-member ensembles with ten-year averages for diagnosing decadal changes in ERF (Pincus et al., 2016)."

14. Further, since the differences are done for each of the three anthropogenically-influenced simulations, does it make sense to discuss correlations due to common variations driven by this approach? I found it difficult to nail down exactly what was fixed between the different models in the simulations. Line 20 of page 2: ".. prescribing identical anth. aerosol optical properties across models allows us: : :. if we : : : know the aerosol distribution" - suggests that optical properties and concentrations are prescribed. Line 9 of page 3 indicates that they "prescribe identical optical properties of anthropogenic aerosols and an associate effect on the cloud reflectivity : : :. ", which I assume to mean only the intrinsic optical properties. However on page 5 , line 24, it appears, again, that the optical depth is prescribed (".. with pre-industrial aerosol optical depth: : : as of the year 1850, three experiments with with tau-p and anthropogenic aerosol from MACv2-SP for the year: : :."), an extensive prescription that appears to fix also the emissions/atmospheric loads of the aerosol. This is fundamental to the paper and should be made crystal clear to the reader, especially in light of the findings about intra-model variability. For example, at line 19 of page 2, the point is made that "uncertainties in process modeling of anthropogenic aerosol" can be separated, but if optical depth is prescribed, I don't see how this is correct.

The revised introduction states: "Here, we prescribe observationally constrained optical properties of anthropogenic aerosol and an associated effect on the cloud droplet number concentration (…), but keep the full model diversity in other aspects. It allows us to eliminate the uncertainties in process modelling of anthropogenic aerosol and focus on the uncertainties in other processes influencing the radiative forcing. In other words, prescribing identical anthropogenic aerosol optical properties and an associated effect on the cloud droplet number concentration across models allows us to study those sources of uncertainty that remain if we pretend to know the spatial distribution of anthropogenic aerosol. We can thereby quantify the sole impact of other model differences, such as the natural aerosol, meteorology, radiative transfer, and surface albedo, on the radiative forcing of observationally constrained anthropogenic aerosol in a state-of-the-art multi-model context.", we further add in Section 2.1: "All other aspects remain model-dependent, e.g., the treatment of the pre-industrial aerosol and clouds (Appendix A)" and document the model differences for the pre-industrial aerosol optical depth in a new paragraph: "We do not prescribe the same natural aerosol nor interfere with any other model components than prescribing the optical properties of anthropogenic aerosols and η_N. For instance, the pre-industrial aerosol optical depth (τ_p) depends on the model (Fig. 2 and 3). Regional differences occur primarily over oceans and deserts, where observations are typically sparse. It is herein noteworthy that ECHAM-HAM runs with interactive parameterisations for dust and sea-salt aerosol resulting in different spatio-temporal variability in τ_p (Fig. 3) compared to the monthly mean climatology MACv1 in ECHAM. In the interactive parameterisations, the natural aerosol emissions, transport and deposition rely on meteorological processes that are difficult to represent in coarse-resolution climate models, e.g., desert-dust emissions strongly depend on the model representation of near-surface winds (e.g., Fiedler et al., 2016) such that constraining the desert-dust burden remains challenging in bottom-up aerosol modelling (e.g., Raisanen et al., 2013, Evan et al., 2014, Huneeus et al., 2016). ", and in Section 2.2: "Moreover, each participating model was free to individually set up all other aspects than the anthropogenic aerosol treatment. We therefore keep for instance the model diversity for the physical parameterisations of radiation and clouds (Appendix A)." The model diversity for clouds is documented in the appendix in the revised manuscript.

15. On numerous occasions, I was confused by wording and lack of specificity. I recommend that the authors perform a through line-by-line reading to make everything as clear as possible.

We have worked on the text and made the following changes in response to your examples:

16. Here are a few examples:

0) The term "multi-estimates" in the title does not appear to be widely used. Perhaps "multiple model estimates" might be more intuitive and familiar to the reader.
Changed to: "multiple estimates"

1) Abstract, line 4: "In those models we reduce: : :" - this makes it sound like a reference to only the models in the CMIP6. Better: "Here we reduce: : :"
Changed to: "We calculate the instantaneous radiative forcing (RF), effective radiative forcing (ERF), and rapid adjustments by comparing 10-year long ensemble simulations with aerosol distributions for 1850, the mid-1970s and the mid-2000s. The complexity of the anthropogenic aerosol is herein reduced"

2) Abstract, line 11 : "model diversity in clouds and use: : :" here "model diversity in clouds" is too vague - what is it referring to?
We removed the statement in the abstract and document the model differences in cloud droplet number, cloud cover, cloud radiative effects and cloud liquid water in the new appendix that we created in response to reviewer #1

3) final sentence: what does "more stringent test" mean?
Changed to: "better test"

17. In Sec. 2.1, it is stated that anthropogenic aerosols are included in the pre-industrial burden, but don't form the majority contributor of AOD in the NorESM. However, the reader needs more information about this to evaluate not the difference between anthropogenic and natural aerosols, but between pre-industrial and more contemporary simulations. One way to do this would be, for example, by providing the absolute anthropogenic contribution to global AOD in the two cases, to

show if the pre-industrial case the anthropogenic contributions are small enough not to invalidate the results from this model relative to the others in the difference.

> We have calculated the contributions of the anthropogenic AOD in 1850 in NorESM and add in the description of NorESM: “The 1850’s global-mean τ_p in NorESM is 0.096, to which anthropogenic fossil-fuel emissions make a contribution of 0.002. For comparison, the year 2005 global-mean τ_a for MACv2-SP aerosols is 0.029.”. This Section has moved to a new Appendix A in response to reviewer #1.

18. Last sentence of page 9: please provide some quantitative estimate of possible differences in natural emissions between pre-industrial and current day (for example due to land use changes etc.)

> We add: “Quantitative changes in natural aerosol burden between the pre-industrial and present-day remain unconstrained, e.g., model estimates of the anthropogenic fraction of desert dust are 10-60% associated with changes in land use and climate (Mahowald and Luo, 2003; Tegen et al., 2004; Stanelle et al., 2014).”

19. Line 17 of page 10: Clarity: it is not clear how consideration of variability does not affect an actual change in ERF. Perhaps the authors mean that they perceive the change as small relative to additional changes reflecting variability? This point is made more clearly in the conclusion.

> Replaced with: “The ensemble-averaged change in ERF is small relative to natural year-to-year variability in modelled ERFs (…).”