In the following, the text with italicization indicates the Reviewers' comments, and the normal text is our response.

**Replies to Reviewer's comments:**

Thank you very much for the insightful comments, which lead to a significant improvement. Here is the item-by-item reply to your comments.

*Reviewer(s)' Comments to Author(s):*

*Reviewer 2:*

*Zhang et al. investigate the role of aerosol forcing in CMIP6 models for a period in which temperatures were too low compared to observations. The study is important and diligently performed. It is of large interest to the readers of Atmos Chem Phys and mostly well written.*

*I have only two major comments, but a number of specific ones.*

*Major remarks*

1. *It would be necessary to list the aerosol effective radiative forcing for each model, as defined from the fixed-SST simulations. This could be in Table 1 or 3; it would be useful to discuss this in comparison to the transient diagnostics.*

*Response:* **Done.** We list the ERFaer for each ESM in Table 3 and discuss their relationship with the transient diagnostics (the aerosol-forcing-sensitivity) in our study in L188-L192: "The aerosol-forcing-sensitivity is different from the commonly used aerosol effective radiative forcing (ERFaer), which is the change in net TOA downward radiative flux after allowing adjustments in the atmosphere, but with sea surface temperatures and sea ice cover are fixed at climatological values." and in L454-456: "The aerosol-forcing-sensitivity is not correlated with the aerosol effective radiative forcing (ERFaer, Table 3), largely due to the strong influence of UKESM1-0-LL on the result."

2. *The term A in Eq. 1 is wrong (A = 1 in general). This has consequences for the analysis in Section 5.*

*Response:* Parameter A estimates the sensitivity of the shortwave flux reflected by clouds to changes in cloud amount. Therefore, the parameter A depends on the baseline cloud albedo (radiative flux per cloud amount unit) and is calculated as $\Delta$(OSR-OSRclr_p)/$\Delta$CLT in our study: $\Delta$ means differences between historical and hist-piAer simulations; OSR-OSRclr_p is the cloudy part OSR; CLT is the total cloud amount.

Therefore, as estimated in the bottom plots in Fig. 7, parameter A varies among ESMs and generally ranges from 1.5 to 2.4. We include a scaled down version of the decomposition in the main text to take the reader through the derivation (L174-238) and try to explain parameter A more clearly in L221 to L223: "That is, the parameter A generally represent the baseline cloud albedo which is sensitive to the cloud parameterizations relative to Cloud Droplet Number Concentration (CDNC), cloud-droplet effective radius, and the other factors."

*Specific remarks*

**l22 The term "aerosol-forcing-sensitivity" is not a standard term and I don't understand it at this point. Either the abstract needs to define it or manage without this new term.**
*Response:* **Done.** We added the definition of "aerosol-forcing-sensitivity" in the abstract (L19-21): "The aerosol-forcing sensitivity, estimated as the outgoing shortwave radiation (OSR) response to aerosol concentration changes, cannot well explain the diversity of PHC biases in the ESMs." We also try to better clarify it in section 2.3 (L173-186): "The sensitivity of the OSR-response to aerosol changes, i.e., the aerosol-forcing-sensitivity, can be measured by the linear fit slope between the annual mean globally averaged OSR differences and loadSO4 differences between the historical and hist-piAer simulations"

*l45 It might be wise to make clear that greenhouse gases accumulate, aerosols do not at time scales longer than a week.*
*Response:* **Done**. We add the statement in L43-44: "Aerosols are generally not evenly distributed around the planet as greenhouse gases, and they have relatively short lifetimes of the order of a week."

*l48 The authors need to clarify where the end point of this sentence is. In the most recent years, anthropogenic aerosol emissions clearly decreased.*
*Response:* **Done.** We add the statements about emission in recent years from L46 to L54: "The rate of change of global aerosol emissions slowed down in the late 20th century (Hoesly et al., 2018), and the trend of global emission has been negative since the mid-2000s (Klimont et al., 2013). There has also been a shift in emission source regions. European and US emissions have declined following the introduction of clean air legislation since the 1980s, while Asian emissions have risen due to economic development. East Asian emissions clearly increased from 2000 to 2005, followed by a decrease with large uncertainties (Aas et al., 2020; Wang et al., 2018). The decade long emission reduction since 2006 over East China is not well represented by the CMIP6 emission (Wang et al., 2021)."

*l50 But Chinese ones reached a peak already in 2011 and declined sharply since then.*
*Response:* You are right. Since the historical run in CMIP6 end in 2014, we cite Wang

et al. (2021) about the decade long emission reduction in China in L51-L54: "East Asian emissions clearly increased from 2000 to 2005, followed by a decrease with large uncertainties (Aas et al., 2020; Wang et al., 2018). The decade long emission reduction since 2006 over East China is not well represented by the CMIP6 emission (Wang et al., 2021)."

*l89 should read "van Noije"*
*Response:* **Done.** Modified in L93 and Table 1.

*l90 Mauritsen et al. (2019) is not the appropriate reference for MPI-ESM-1-2-HAM*
*Response:* **Done.** The reference for MPI-ESM-1-2-HAM is changed to Neubauer et al. (2019).

*l153 This term "aerosol-forcing-sensitivity" is introduced here the first time and needs to be defined precisely.*
*Response:* **Done.** We modify "aerosol-forcing-sensitivity" to "aerosol forcing" in this sentence which is more appropriate (L157 and L160). The definition of "aerosol-forcing-sensitivity" is given in the next paragraph (L173-186): "The sensitivity of the OSR-response to aerosol changes, i.e., the aerosol-forcing-sensitivity, can be measured by the linear fit slope between the annual mean globally averaged OSR differences and loadSO4 differences between the historical and hist-piAer simulations."

*l156 The authors need to provide a reference for this statement, or demonstrate by other means that it applies. It is not obvious, since the organic aerosol emissions often occur in very different places and are only partly linked to the fossil fuel burning that generates most anthropogenic SO2.*
*Response:* **Done.** You are right. The estimates of organic aerosol emissions are less available. In CMIP6 simulations, the emissions in Community Emissions Data System (CEDS) inventory are adopted by CMIP6 models. The CEDS emissions show similar evolutions for SO2, BC, and OC (Hoesly et al, 2018), especially for the steadily increasing in the mid-20th century (the pole-hole cooling period). We cite the reference about CEDS as suggested in L157-L163: "Whilst the overall impact of aerosol forcing will also depend on other aerosol species, we adopt this approach because the sulphates dominate estimates of aerosol forcing during this period and other aerosols species can be assumed (as a 1st order approximation) to have covaried with the $SO_2$ emissions during this period as presented by the Community Emissions Data System (CEDS) inventory adopted by CMIP6 models (Hoesly et al, 2018)."

*l164 drop "is"*
*Response:* **Done.**

*l167 And of course natural variability*
*Response:* **Done.** Modified in Line170-173: "$\Delta OSR$ of course involves any differences in natural variability and planetary albedo between the two simulations, including clear-

sky albedo changes and any adjustments in the microphysical or macroscopic properties of clouds."

*l169 It would be useful to clarify at which scale the data are aggregated. Are these monthly or annual means? global or regional means? If little aggregation, can the change in sulfate load not become very small?*
*Response:* We use annual mean globally averaged OSR and loadSO4 differences to estimate the aerosol-forcing-sensitivity. We clarified this in L173-L176: "The sensitivity of the OSR-response to aerosol changes, i.e., the aerosol-forcing-sensitivity, can be measured by the linear fit slope between the annual mean globally averaged OSR differences and loadSO4 differences between the historical and hist-piAer simulations."

*l170 Forcing or effective forcing?*
*Response:* Sentence is deleted.

*l177 It would be very useful to disentangle the two. The aerosol effective forcing is readily defined by the fixed-SST simulations designed for this purpose. The authors could investigate this in comparison to the same period in the runs they investigate here. Another option would be to make use of the DAMIP simulation with varying aerosol.*
*Response:* **Done (as the response to Major Remark 1).** We compare the "aerosol-forcing-sensitivity" in our study and the "aerosol effective forcing (ERFaer)" calculated from the fixed-SST data provided in RFMIP and shown in Table 3 of the manuscript.

It would be very interesting to compare with the DAMIP simulation with varying aerosol (hist-aer). However, there is only a small overlap between models participating in AerChemMIP (for the hist-piAer simulation) and those participating in DAMIP. Of the ESMs used in this study for the analysis of hist-piAer, only GFDL-ESM4 has done the hist-aer simulation.

*l183 To which extent is it possible to include the feedbacks in this distinction? Are the feedbacks mainly in the ACI term in this definition?*
*Response:* The aerosol-cloud feedbacks are mainly in the ACI term which includes cloud spatial extent (amount), cloud albedo on radiative fluxes, and cloud particle swelling by humidification. There is also a (smaller) effect of feedback on the ARI term it is also affected by cloud amount changes. We add the statement in L250-255: "The aerosol-cloud feedbacks are mainly in the ACI term which includes cloud spatial extent (amount), cloud albedo on radiative fluxes, and cloud particle swelling by humidification (Christensen et al., 2017; Neubauer et al., 2017). There is also a (smaller) effect of feedback on the ARI term that is also affected by cloud amount changes insofar as increased/decreased cloud cover can obscure/reveal clear-sky radiative fluxes."

The small and large effects of feedback in the ARI term and the ACI term are evident in the amplitude of ARI differences and the ACI (Cloud-Albedo term PLUS Cloud-Amount term) differences between ESMs (Fig. 8). This is consistent with Chen et al. (2014). We cite Chen et al. (2014) in L530-L532: "Chen et al. (2014) also

suggested that ACI is the main contribution to the Aerosol radiative forcing uncertainty and the response of marine clouds to aerosol changes is paramount.”

***l184 This is an approximation that makes several mistakes. It neglects the (regionally very important) absorption above clouds, and it also neglects the adjustments to aerosol-radiation interactions.***

*Response:* Agreed. These effects might not be proportional to cloud fraction or clear-sky OSR, so they would appear as residuals in this framework. However, there is a suggestion in the literature that aerosol forcing and emissions may respond more linearly at a global mean scale than they are known to do so at regional scales (e.g. Booth et al, 2018, https://doi.org/10.1175/JCLI-D-17-0369.1, Kretzschmar et al, 2017). Our formulation explicitly assumes that there is a broadly linear relationship between loadSO4 and emissions, OSR and loadSO4, and non-linearity due cloud albedo or amount or any interaction is small. We now acknowledge this is in the text on L255-265: “We acknowledge that the linear approximation in our method doesn't explicitly account for the absorption above clouds, or the adjustments to aerosol-radiation interactions (e.g., Carslaw et al., 2013) that are known to be locally important. Our formulation explicitly assumes that there is a broadly linear relationship between loadSO4 and emissions, and aerosol radiation with loadSO4 (and non-linearity due to cloud albedo or amount or any interaction is small at global scale as suggested in Booth et al. (2018)). Should these interaction terms be non-negligible in this analysis, we still expect the broader attribution of the reasons for the model diversity in temperature response over the PHC period, either how they simulate aerosol concentrations or how they simulate the response to this, to generally hold.”

***l186 Lifetime is only measurable in terms of horizontal and/or cloud albedo.***

*Response:* Good point! In this framework, cloud lifetime effects will show up via the albedo and areal extent changes. For simplicity, we've removed mention of 'cloud lifetime' from this list (L236).

***l200 This seems to be wrong. Instead of A, one should use OSRcld_hist in Eq. 1.***

**Response:** We think this is right (please see our derivation in Appendix A for details). The decomposition is a bi-linear regression, where we first regress out the linear dependence on clear-sky OSR difference (which gives the first term, and Parameter M, in Eq. 4) and then regress out linear dependence on cloud-amount differences (which gives the Parameter A). Parameter A is therefore the estimated radiative-flux change per cloud amount unit. We use the cloudy part OSR and cloud amount differences between historical and hist-piAer simulations( $\Delta$ (OSR-OSRclr_p)/$\Delta$CLT) to calculate parameter A. As estimated by the slope of linear fit in Fig.7g-7i, parameter A varies among ESMs and generally ranges from 1.5 to 2.4.

***l208 This is not true, firstly because a lifetime effect may also involve changes to cloud albedo, and second because this lifetime effect is (by far) not the only influence***

*on cloud extent (as the authors immediately acknowledge).*

**Response:** Agreed – we've removed the claim that the ACI-interaction term is 'analogous' to the Albrecht effect.

*l214 Presumably, "increases in cloud droplet number concentration"? In fact, this is the key impact of aerosols on clouds (Twomey effect, radiative forcing due to aerosol-cloud interactions). But adjustments of cloud water path are also included.*

*Response:* You are right. We add the impact of cloud droplet number concentration in L236 to L239: "Such adjustments would include increases in cloud droplet number concentration and increases in simulated cloud-droplet effective radius without accompanying changes in cloud cover."

*l223 Maybe it is noteworthy that, e.g., Chen et al. (Nature Geosci 2014) or Christensen et al. (ACP 2017)*

*Response:* These two papers are very interesting and support our study. They have been cited in L250-L253: "The aerosol-cloud feedbacks are mainly in the ACI term which includes cloud spatial extent (amount), cloud albedo on radiative fluxes, and cloud particle swelling by humidification (Christensen et al., 2017; Neubauer et al., 2017).", and L530-532: "Chen et al. (2014) also suggested that ACI is the main contribution to the Aerosol radiative forcing uncertainty and the response of marine clouds to aerosol changes is paramount."

*l226 In fact, the decomposition is so far off "first" and "second indirect effects" (and, by the way, the terms are obsolete since AR5) that it is better to drop this paragraph.*

*Response:* Agree. The paragraph is deleted.

*l241 Anomalies with respect to which time period average?*

*Response:* The anomalies are relative to 1850-1900 mean. We have rewritten the figure capture for Fig.1 to make it clear.

*l242 The terms "MMM" and "MME" seem strange to me. Often, MMM is multi-model mean, but here it seems it is, in contrary, the single-model ensemble mean. What is a "multi-member ensemble"? is that not tautological? Is the term "MMM" necessary at all? Why not simply "for each model, the ensemble mean is shown"? What does the acronym "MME" stand for?*

*Response:* Thank you for your explanation. I modified the statements about single-model ensemble mean and now "MMM" stand for the multi-model mean.

*l255 What distinguishes the "first" member from any member?*

*Response:* The member used here is actually the first realization for each model. We rewrite the sentence in L287-291: "Figure 1a shows the near-global averaged time series of annual mean TAS anomaly relative to 1850 to 1900 in HadCRUT5 during the historical period from 1850 to 2014, and the ensemble means for each model except for EC-Earth3-AerChem and GFDL-ESM4 (where only a single realization is available for

the hist-piAer experiment).”

***l292 "Sensitive to initial condition" seems strange wording for an influence of internal variability.***
*Response:* **Done.** We modified the wording as "The PHC bias are large (-0.37ºC) in EC-Earth3, which has prescribed chemistry and aerosol. The large bias may be a reflection of the large internal variability on TAS in EC-Earth3 (Döscher et al., 2021), for which we have only one member." in L329-L332.

***l320 A colour scale that evenly is distributed in positive and negative directions should be chosen. Or else this discrepancy should be pointed to in the caption.***
*Response:* **Done.** Since the warming amplitude in early 2000s is about twice of the amplitude of "pot-hole" cooling, we choose to use unevenly distributed color scale. We declare differences in the color scale in the caption: "Note that the color scale intervals in the positive and negative directions are 0.2 ºC and -0.1 ºC, respectively."

***l370 This is only true at a very superficial glance. There is not intimate link evident.***
*Response:* **Done.** We delete this sentence.

***l374 This correspondence is hardly evident.***
*Response:* **Done.** The northern-hemisphere anthropogenic surface $SO_2$ emissions are shown by line contours in Fig.2. The contours start from 20 to 40 ng $m^{-2}s^{-1}$ with an interval of 10 ng $m^{-2}s^{-1}$. The surface $SO_2$ emissions are small before 1950s and rapidly increase during the PHC period. We modified the sentence and mainly focus on the large surface $SO_2$ emissions and sulphate loading during the PHC period (L403-405): "The growing sulphate loading during the PHC period corresponds with the increase in northern-hemisphere anthropogenic surface $SO_2$ emissions (line contours in Fig.2)".

***l389 why "relative"?***
*Response:* **Done.** We delete the word "relatively".

***l389 The more immediate reason is the higher incoming solar flux at lower latitudes.***
*Response:* **Done.** We add the impact of solar flux in L417-419: "Due to the more rapid oxidation and higher incoming solar flux at lower latitudes, an equatorward shift in SO2 emissions around 1990s result in a more efficient production of sulphate and stronger aerosol forcing.".

***l399 Perhaps this should be more cautiously "approximately linearly". See debate of Stevens (2015, doi: 10.1175/JCLI-D-14-00656.1), Kretzschmar et al. (2017, doi:10.1175/JCLI-D-16-0668.1), Booth et al. (2018, doi:10.1175/JCLI-D-17-0369.1).***
**Response: Done.** We change the wording to "approximately linearly" as suggested in L431.

***l402 What is "this nonlinearity"? Before, linear relationships were described.***

*Response:* **Done.** We rewrite the sentence as follows (L429-435): "In both cases, the approximation of linearity holds less well for UKESM1-0-LL, especially at small sulphate loadings. This reflects the behaviour of HadGEM2, a predecessor of UKESM1 (Wilcox et al., 2015), and is likely to be due to the strong aerosol-cloud albedo effect in these models. The global mean annual mean reff decreases by about 0.7 μm since pre-industrial era, more than twice the magnitude of change seen in the other models (Fig.1b in Wilcox et al., 2015 and Fig.9b in this study)."

*l479 Why "generally"? It is of course a coarse approximation only.*
*Response:* **Done.** We replace "generally parametrized as" to "approximated to" in L508.

*l519 The cloud albedo term may also carry a substantial contribution by adjustments, namely via the adjustments of liquid water path.*
*Response:* The ACI term will be influenced by factors contributing to cloud albedo. Liquid water path changes can often be one of the strong drivers of cloud albedo change and can be driven by changes unrelated to aerosol changes (such as global warming). In this study we consider only the radiative changes between the historical simulations and the parallel simulations where aerosol emissions are fixed (hist-piAer). We cannot hold cloud liquid water path fixed in these experiments. So, whilst liquid water changes will be one of a number of factors influencing our ACI term, these should just be capturing those liquid water path changes that are related to the presence or absence of aerosol emission changes.

*l529 This is not quite true. Delta reff is also influenced by changes in liquid water path. The forcing can be identified when investigating droplet number concentrations.*
*Response:* Agreed. we've tried to make this sentence more explicit (L571-573): "$\Delta r_{eff}$ is generally related to the cloud-optical depth and cloud water path, and $\Delta CLT$ is related to adjustments in cloud cover due to ACI."