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

Replies to Reviewer's comments:

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

Reviewer: 1

The manuscript firstly points out that most CMIP6 earth system models (with interactive atmospheric sulfate cycle) present cold biases in the period 1960-1990 (called PHC in the manuscript, pot-hole cooling). The authors performed then a series of investigations searching the relation of this cold bias with sulfate sources and sinks across the available CMIP6 models. The authors finally proposed a single parameter "ESRT", effective sulfate retention time. This is an interesting diagnostic, relatively stable for a given model and quite useful to characterize its sulfate cycle. It was shown that ESRT has a good capacity to explain the cold bias across models. It is also interesting to see that the authors use the temperature anomalies of the PHC period to "constrain" the optimal value of ESRT. This optimal value is then used to approximate the "right" sulfate deposition rate which is furthermore used in the BCC model with improved performance.

All that said, I have a small concern for what shown in Fig. 1a displaying temperature time series. From those curves, I can deduce that the cold bias of models in the PHC period is not exceptional, not as the authors pointed out, since there is a good trend compared to observation. But the cold bias (at least in the multi-model ensemble mean) occurred before the PHC period, roughly at the point of 1935 where models drift significantly from observation and the cold bias remains for the rest of the time, including the PHC period (1960-1990).

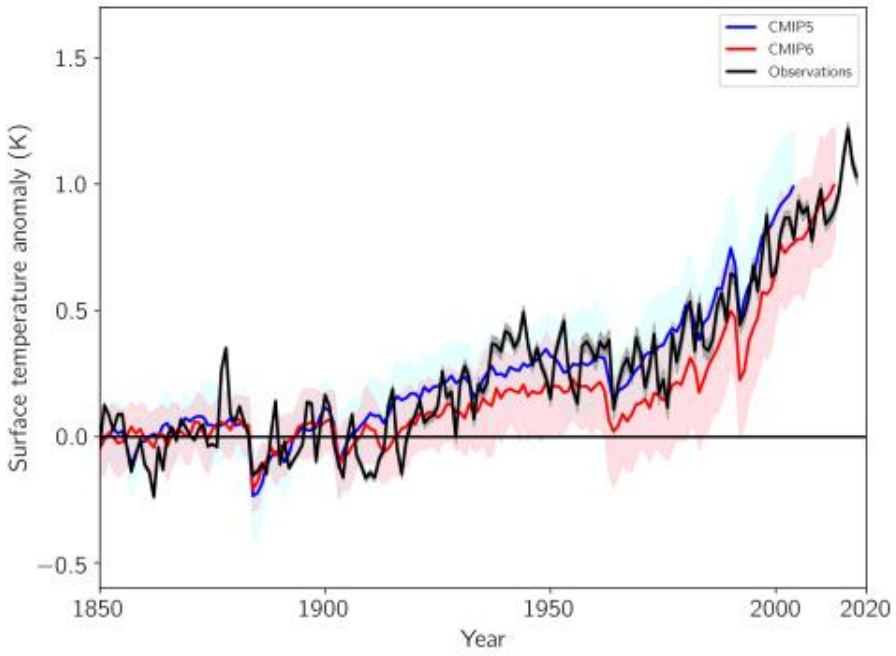

Figure 12. Ensemble mean historical surface warming in CMIP5 and CMIP6 compared with observations. Shading on the models is the ensemble SD. The baseline is 1850–1900.

https://doi.org/10.5194/acp-20-7829-2020

Response: Thank you for your comments. We reference Figure 12 from Flynn and Mauritsen (2020), which evaluates historical surface temperature anomalies in CMIP5 and CMIP6 models. Their analysis indicates that the CMIP5 multi-model ensemble mean effectively captured the instrumental record with observation falling well within model spread – a consistency also noted in CMIP3 models assessed in the IPCC Third Assessment Report (IPCC AR3). In contrast, a majority of CMIP6 models exhibit a cold bias in surface temperature, marking a notable departure from earlier model generations. We clarify this in L146-149: "The anomalous cooling in CMIP6 model marked a notable departure from earlier model generations, which can effectively capture the instrumental SAT record with observation falling well within model spread (e.g., Flynn and Mauritsen, 2020; Hegerl, et al., 2007)."

You are right. The cold bias occurred before the PHC period, roughly at the point of 1935. We think it can also be attributed to elevated sulfate aerosol burdens. As shown in Fig.1b, the sulfate burden increased steadily since the Industrial Revolution. The selection of the 1960–1990 as PHC period in our analysis stems from its alignment with accelerated anthropogenic emissions, particularly of sulfate precursors (e.g., SO_2). The large anthropogenic emissions in PHC amplify the model-observation divergence during this era. By focusing on this interval, we aim to quantify the climate impacts of anthropogenic aerosols during a period of rapidly increasing industrial activity. We clarify this in section 3.1 (L145-L146): "All the models tend to underestimate SATa since the 1930s.", L150: "The cooling bias is most pronounced from 1960 to 1990, i.e., the PHC period.", and in L169-172: "The PHC coincides with increased anthropogenic emissions, particularly of sulfate precursors such as SO_2 (Zhang et al., 2021a). Global emissions of SO_2 grew steadily after the 1950s and peaked in the 1970s at 180Tg yr^{-1}, which is about 3.6 times the 1950s' emissions (Hoesly et al., 2018)."

Replies to Dr. Stephen E. Schwartz's review on: Unveiling Sulfate Aerosol Persistence as the Dominant Control of the Systematic Cooling Bias in CMIP6 Models: Quantification and Corrective Strategies by Jie Zhang et al, for ACP.

In the following, the text with italicization indicates Dr. Stephen E. Schwartz's comments, and the normal text is our response.

Comments to Author(s):

I have major concern over the definition of the quantity that the authors call the ESRT, effective sulfur retention time scale. This is a non-conventional definition of a residence time that may account for the short (ca 1 day) values reported, and certainly precluding comparison with other measures of lifetime in the literature.

The authors seem unaware of the large prior literature pertinent to this study.

I elaborate on these concerns in the pdf review.

Response:

Thank Dr. Schwartz for your constructive critique, particularly for emphasizing the importance of distinguishing the "effective sulfur residence time" (ESRT) from other lifetime measures reported in the literature. We acknowledge that the term "effective sulfur residence time (ESRT)" is potentially misleading.

ESRT was envisioned primarily as a diagnostic tool (not a physical timescale) for model tuning. Because it accounts for both sulfate and SO_2 deposition, its value is typically lower than the sulfate atmospheric lifetime. As noted, it is fundamentally a metric for model evaluation rather than a conventional definition of atmospheric residence time. Therefore, we renaming it to the "Sulfur Assessment Metric for ESMs" (SAME) in the revised manuscript.

We acknowledge that sulfate lifetime remains critical for validating the model's physical realism. And there are three major changes in the revised manuscript according to Dr. Schwartz's comments. The manuscript is amended accordingly based on these three major modifications, including the introduction.

1. Discussion about sulfate lifetime in Section 4.

We calculated sulfate lifetime as the ratio of sulfate burden to total sulfate deposition (wet plus dry) in the CMIP6 models, BCC-ESM1-1, and UKESM1-1-LL (Table 2) to ensure model credibility. The detailed analysis and discussion are presented in the newly added Section 4: "**Discussion: Sulfate lifetime in CMIP6 models and the two post-CMIP6 models**".

The analysis and discussion are shown in L348-L363: "As shown in Table 2, sulfate lifetime in CMIP6 models ranges from 1.65 days in MIROC-ES2L to 6.57 days in EC-Earth3-AerChem. The mean sulfate lifetime is 3.93 days, consistent with previous literatures, particularly the mean value of 4.12 days in AeroCom models with standard deviation of 18% (Textor et al., 2006). The wide sulfate lifetime range in CMIP6 models is attributed to variations in both sulfate burden (0.33 to 0.75 Tg S) and deposition rates (0.75 to 7.58 Tg S yr^{-1} for dry deposition, and 31.68 to 69.41 Tg S yr^{-1} for wet deposition).

Sulfate lifetimes in the two post-CMIP6 models, 8.53 days in BCC-ESM1-1 and 5.77 days in UKESM1-1-LL, are generally longer than those of their CMIP6 versions. The longer sulfate lifetimes in the two post-CMIP6 models may be due to lower SO2 in these revised models but also could be due to physical climate changes (e.g., temperatures, clouds, rainfall). Compared to prior lifetime measures reported in the literature and considering the range of lifetimes found in recent models, the sulfate lifetimes in BCC-ESM1-1 and UKESM1-1-LL also appear reasonable (e.g., Charlson et al, 1992; Kristiansen et al. 2012; Textor et al., 2006)."

2. Definition of SAME index.

To eliminate the effect of differing climatological states across models, in the revised manuscript the SAME metric is defined as the ratio of sulfate anomaly during the PHC period to the sum of sulfate and SO_2 deposition anomalies:

SAME = loadSO4a / (DSO4a + DSO2a)

where:

- loadSO4a is the total sulfate loading anomaly in the atmosphere,
- DSO4a denotes the total (wet plus dry) sulfate deposition anomaly, and
- DSO2a denotes the total (wet plus dry) SO_2 deposition anomaly during the PHC period.

Since the definition of SAME is central to our analysis, we add a new section (**Section 2: Model, data, and method**) to introduce the data and methods as suggested.

3. Refine the constraint of SAME.

As shown in Fig.4a, the SAME ranges from 1.1 days in MIROC models to 2.86 days in EC-Earth3-AerChem. The correlation coefficient between SATa and SAME is -0.90 (Fig. 4b). In addition to the linear regression between SATa and SAME (black line in Fig. 4b), **in the revised manuscript, we also calculate the 95% confidence interval (CI, blue curves) and the 95% prediction interval (PI, red curves):**" (L292-294) We calculate the linear fitting between SATa and SAME (black line in Fig. 4b), the 95% confidence interval (CI, blue curves), and the 95% prediction interval (PI, red curves), respectively.", "(L303-310)As shown by the red asterisk in Fig.4b, the SAME reduced from 2.51 days to 1.43 days in updated BCC-ESM1 (BCC-ESM1-1), falling right within the PI constraint. The new SAME index is 57% of its previous values. Accordingly, the SATa in PHC is 0.34ºC, falling within the observational range from 0.165ºC to 0.515ºC. We also examine the SAME in UKESM1-1-LL with modified SO_2 dry deposition parameterization. The SAME is shortened from 2.19 days to 1.71 days, falling within the CI constraint. Accordingly, the SATa in PHC period increases by about 0.25ºC."

We also discuss the uncertainty in SAME estimate in L311-322: "Given that most models underestimate SATa relative to observations, extrapolating SAME values for

SATa exceeding the observation (0.34^{o}C) becomes highly uncertain. Result from BCC-ESM1-1 suggests that the rate of decrease in SAME predicted by the regression line may not hold for SATa values above the observed lower bound (0.165 °C). Therefore, we recommend a central SAME estimate of 1.35 days. Critically, this value carries inherent uncertainties that must be quantified:

- The 95% confidence interval (CI) of ±0.25 days (i.e., 1.10–1.60 days).

- The wider 95% prediction interval (PI) of ±0.6 days (i.e., 0.75–1.95 days).

The substantial difference between the CI and PI ranges underscores the challenge in precisely constraining SAME. We advise using the PI for applications requiring robustness against individual model deviations."