Point-by-point reply for the manuscript „Interactions between atmospheric composition and climate change - Progress in understanding and future opportunities from AerChemMIP, PDRMIP, and RFMIP“ by Fiedler et al.

We thank the reviewers for commenting on our perspective. The comments helped us to develop a substantially revised manuscript. Newly designed tables, revised figures, and added text sharpen the presentation and give more details. Below are our replies in blue to the statements of the reviewers in black along with details on our intended changes in the manuscript.

Response to reviewer 1

As much as I applaud the motivation behind this paper, to show the connections between the three MIP's and thinking about ways forward, I can't recommend publication of this article as it is missing any meaningful conclusions. I don't want to discourage the authors from writing this paper, however I would like to encourage them to further develop ideas and summarize more developed ideas about ways forward.

Overall, the paper summarizes what the three MIP's are about (which has been already done in other papers), followed by presenting some new aspects to be included in future MIP's. However, all those ideas stay at the surface by just naming them, and giving short paragraphs on what those buzz words are, without really discussing how they can be connected to future MIPs. As such the paper does not give any new information and unfortunately is not useful in its current form.

Thank you for the constructive comments. We aimed for a balanced presentation of different aspects that the three MIPs have in common while keeping their specific purposes in mind. One published manuscript, even if it was much longer than this one, could not fully consider all aspects in detail and reading further literature will therefore always be necessary. The three MIPs cover different broad areas of research in Earth system sciences and there are multiple future directions that one could pursue. Our intention here is to synthesize and emphasize different developments from the three MIPs based on available literature and the experience that we made during performing the works. We anticipate this manuscript to serve as a starting point for planning new MIPs on the interactions of atmospheric composition and climate change. We think this is useful for others, especially for those who might want to design a successful MIP of their own, e.g., given the recent call for community MIPs for CMIP7 (https://wcrp-cmip.org/model-intercomparison-projects-mips/).

Please note that we would not want to recommend single specific directions for good reason. It is our belief that diversity in scientific methods and questions are essential and integral values of science. Any restriction, whether through recommendation or maybe imposed somewhere, for certain research directions would be limiting the freedom of choice of scientists and, hence, we do not suggest, or worse even prescribe, what future research should primarily focus on. The research questions and methods should be freely picked by the future MIP leaders who may or may not use the information provided in this manuscript. We intend to revise our manuscript to even better fulfill this purpose as support of the development of ideas for future MIPs on the interactions of atmospheric composition, air quality, and climate change. The reviewer comments and questions are useful and are greatly acknowledged for this purpose.

For example:

Emulators: they are just mentioned, and it should be pointed out that some very successful work including emulators, including perturbed physics experiments have already been carried out under CMIP6. What is the future vision here? How should this be more integrated into CMIP? Challenges, possibilities etc?

In our community we try to meet requirements from such simple climate models by considering them in the experimental design, e.g., for AerChemMIP2, in addition to their use in our research. The question on how to best integrate emulators into CMIP as a whole is being addressed by the CMIP Strategic ensemble design task team (https://wcrp-cmip.org/cmip7-task-teams/ensemble-design/). A future vision for the broader utility of emulators in CMIP7 might be developed by the task team.

We refer to several studies employing emulation techniques for CMIP6 including perturbed parameter ensembles. In the revised manuscript, we add the references to Sudakow et al. (2022) and Tebaldi and Knuti (2018) to the previously existing paragraphs on emulators:

“In terms of physically-based emulators of the climate system (i.e. simple climate models), RFMIP and AerChemMIP experiments were invaluable to determine aerosol ERF, ozone ERF and the factors influencing methane chemical lifetime. Some of these relationships were developed in the lead-up to AR6 WG1 and used directly in the report (e.g., Smith et al., 2021, Thornhill et al., 2021a, 2021b). (...) A review of emulation techniques that are routed in statistical mechanics highlights the potential to further improve emulators for the use in climate sciences by using machine learning (Sudakow et al., 2022). Also the difficulty of accounting for non-parametric biases of CMIP models in emulators remains (Jackson et al., 2022). Nevertheless, emulators have already been proven useful to sample parametric differences and to study climate change (e.g., Tebaldi and Knutti, 2018).” Due to the link with machine learning, we combine the former Sections 4.1.1 and 4.1.3 into one Section 4.1.1 in the revised manuscript.

Jackson, L. S., Maycock, A. C., Andrews, T., Fredriksen, H.-B., Smith, C. J., & Forster, P. M. (2022). Errors in simple climate model emulations of past and future global temperature change. *Geophysical Research Letters*, 49, e2022GL098808. https://doi.org/10.1029/2022GL098808

Sudakow, I., Pokojovy, M., & Lyakhov, D. (2022). Statistical mechanics in climate emulation: Challenges and perspectives. *Environmental Data Science, 1*, E16. doi:10.1017/eds.2022.15

Smith, C. J., Harris, G. R., Palmer, M. D., Bellouin, N., Collins, W., Myhre, G., et al. (2021). Energy budget constraints on the time history of aerosol forcing and climate sensitivity. *Journal of Geophysical Research: Atmospheres*, 126, e2020JD033622. https://doi.org/10.1029/2020JD033622

Tebaldi, C., & Knutti, R. (2018). Evaluating the accuracy of climate change pattern emulation for low warming targets. *Environmental Research Letters*, *13*(5), 055006, doi:10.1088/1748-9326/aabef2

Thornhill, G., Collins, W., Olivié, D., Skeie, R. B., Archibald, A., Bauer, S., Checa-Garcia, R., Fiedler, S., Folberth, G., Gjermundsen, A., Horowitz, L., Lamarque, J.-F., Michou, M., Mulcahy, J., Nabat, P., Naik, V., O'Connor, F. M., Paulot, F., Schulz, M., Scott, C. E., Séférian, R., Smith, C., Takemura, T., Tilmes, S., Tsigaridis, K., and Weber, J. (2021a). Climate-driven chemistry and aerosol feedbacks in CMIP6 Earth system models, Atmos. Chem. Phys., 21, 1105–1126, https://doi.org/10.5194/acp-21-1105-2021

Thornhill, G. D., Collins, W. J., Kramer, R. J., Olivié, D., Skeie, R. B., O'Connor, F. M., Abraham, N. L., Checa-Garcia, R., Bauer, S. E., Deushi, M., Emmons, L. K., Forster, P. M., Horowitz, L. W., Johnson, B., Keeble, J., Lamarque, J.-F., Michou, M., Mills, M. J., Mulcahy, J. P., Myhre, G., Nabat, P., Naik, V., Oshima, N., Schulz, M., Smith, C. J., Takemura, T., Tilmes, S., Wu, T., Zeng, G., and Zhang, J. (2021b) Effective radiative forcing from emissions of reactive gases and aerosols – a multi-model comparison, Atmos. Chem. Phys., 21, 853–874, https://doi.org/10.5194/acp-21-853-2021

Km- scale modeling: This is a very big topic. Again, here it is just mentioned without giving any perspective? How would km-scale modeling be integrated for CMIP? What are the challenges? Aka not resolving chemistry while this paper discusses composition related MIPs. A much more critical assessment is needed. Can these experiments even be done using coupled oceans? How would that be integrated? How does that relate to TriMIP climate change experiments? What is the real use as high resolution weather models already exist since a long time. How would pushing to even higher resolution solve any climate issues? There is a lot to be discussed, like connecting climate change to impacts etc. Connecting weather modeling to CMIP modelling, etc. However this paper discusses no real issues, just mentions general topics.

Agreed, it is a big topic. The best way to integrate the new generation of climate models into CMIP in general needs to be decided through broader discussion involving more than the three MIP communities here. From our perspective, there will always be a tradeoff between model resolution, experiment length, and number of experiments, and this choice needs to be adequate for the scientific question being addressed, as stated in the

manuscript. In the revised manuscript, we intend to expand on the role of spatial resolution and possible ways forward for our community, primarily by adding the following paragraphs in Section 4.1.2:.

“There is evidence that global coupled atmosphere-ocean simulations with a few kilometers resolution with first interactions with atmospheric composition, namely the carbon cycle (Hohenegger et al., 2023), can be done. Such model experiments have been carried out with a computational performance that sparks hope that kilometer-scale modeling will be possible to answer open questions concerning climate change to provide information for societal needs. Areas that kilometer-scale experiments could advance in the field of interest of the three MIPs encompass some pressing questions for the understanding of interactions of atmospheric composition and climate change. For some research questions on atmospheric composition and the associated climate response in our field, kilometer-scale experiments are already used, e.g., for a better understanding of aerosol-cloud interactions (Simpkins, 2018), which is one of the key uncertainties in ERF from ESMs (e.g., Smith et al., 2020). One question that can be better addressed with kilometer-scale experiments is the resolution dependence of radiative forcing and feedbacks, especially for those that involve clouds that are a key uncertainty in our understanding of climate change with ESMs (Stevens and Bony, 2013). Another question is to what extent more resolved meteorological processes aid in improving the representation of atmospheric composition and air quality, e.g., concerning health impacts in urban areas. “

“The community of the three MIPs will not be able to entirely rely on global kilometer-scale model experiments in CMIP7, especially in the context of a MIP since fully coupled Earth System Models (with aerosols and chemistry) at a resolution 1 km fast enough to perform multi-decadal simulations are unlikely to be ready in the timeline of CMIP7. In light of this restriction, we see two main routes forward for immediately using spatially refined information in our next MIPs. The first possible and computationally smart way is to use the output from global kilometer-scale experiments that are run for other purposes to drive offline models for aerosols and chemistry or atmospheric radiation transfer calculations. This approach is suitable to answer some but not all research questions in our community. For instance, the response of dust emission fluxes to changes in winds and moisture can be addressed with offline modeling and allows to identify underlying reasons for changes and model differences in the dust response (Fiedler et al., 2016), but the implication of such dust emission changes for air quality and climate responses can not be quantified with offline modeling. For the latter, regional one- or two-way dynamical downscaling approaches could be used. We perceive dynamical downscaling as the second main avenue for our near-future works to obtain regionally refined spatial information. Regional climate modeling is already well developed and organized via CORDEX with a focus on providing regional climate change information. Regional climate models with the capability to perform experiments with coupled aerosols and chemistry exist for instance in Europe and the US (e.g., Pietikäinen et al., 2012, Schwantes et al., 2022), but have not been used in our past MIPs. For CMIP7, UKESM2 and CESM aim also to have regional model configurations. An ensemble of regional composition-climate models therefore exists and could be used in future MIPs. The regional models will nevertheless need output from global climate model experiments with coupled aerosols and chemistry as boundary data for performing the regional experiments. As such our need for experiments with classical global ESMs is retained, at least for CMIP7, although we are not averse to the idea of moving towards global kilometer-scale modeling with a sufficient coupling of physical processes to aerosols and chemistry to address the community’s research interests.”

Bony, S., Stevens, B., Frierson, D. *et al.* (2015) Clouds, circulation and climate sensitivity.*Nature Geosci* 8, 261–268. https://doi.org/10.1038/ngeo2398

Hohenegger, C., Korn, P., Linardakis, L., Redler, R., Schnur, R., Adamidis, P., Bao, J., Bastin, S., Behravesh, M., Bergemann, M., Biercamp, J., Bockelmann, H., Brokopf, R., Brüggemann, N., Casaroli, L., Chegini, F., Datseris, G., Esch, M., George, G., Giorgetta, M., Gutjahr, O., Haak, H., Hanke, M., Ilyina, T., Jahns, T., Jungclaus, J., Kern, M., Klocke, D., Kluft, L., Kölling, T., Kornblueh, L., Kosukhin, S., Kroll, C., Lee, J., Mauritsen, T., Mehlmann, C., Mieslinger, T., Naumann, A. K., Paccini, L., Peinado, A., Praturi, D. S., Putrasahan, D., Rast, S., Riddick, T., Roeber, N., Schmidt, H., Schulzweida, U., Schütte, F., Segura, H., Shevchenko, R., Singh, V., Specht, M., Stephan, C. C., von Storch, J.-S., Vogel, R., Wengel, C., Winkler, M., Ziemen, F., Marotzke, J., and Stevens, B. (2023) ICON-Sapphire: simulating the components of the Earth system and their interactions at kilometer and subkilometer scales, Geosci. Model Dev., 16, 779–811, https://doi.org/10.5194/gmd-16-779-2023.

Fiedler, S., Knippertz, P., Woodward, S. *et al.* (2016) A process-based evaluation of dust-emitting winds in the CMIP5 simulation of HadGEM2-ES. *Clim Dyn* 46, 1107–1130. https://doi.org/10.1007/s00382-015-2635-9

Schwantes, R. H., Lacey, F. G., Tilmes, S., Emmons, L. K., Lauritzen, P. H., Walters, S., et al. (2022) Evaluating the impact of chemical complexity and horizontal resolution on tropospheric ozone over the conterminous US with a global variable resolution chemistry model. *Journal of Advances in Modeling Earth Systems*, 14, e2021MS002889. https://doi.org/10.1029/2021MS002889

Stevens, B., and Bony, S. (2013) What Are Climate Models Missing? *Science,* 340,1053-1054, doi:10.1126/science.1237554

Pietikäinen, J.-P., O'Donnell, D., Teichmann, C., Karstens, U., Pfeifer, S., Kazil, J., Podzun, R., Fiedler, S., Kokkola, H., Birmili, W., O'Dowd, C., Baltensperger, U., Weingartner, E., Gehrig, R., Spindler, G., Kulmala, M., Feichter, J., Jacob, D., and Laaksonen, A. (2012) The regional aerosol-climate model REMO-HAM, Geosci. Model Dev., 5, 1323–1339, https://doi.org/10.5194/gmd-5-1323-2012.

Machine learning: Again just mentioning a topic, without going deeper. There is so much to discuss. What about ML is useful for CMIP, what parts are not useful. Replacing models physics/chemistry with ML has very big problematic sides, creating not understandable black boxes. At the same time using ML might be unavoidable in the future, if higher resolution is necessary for others processes. How does this topic fit into the CMIP framework? What are possible ways forward? Again, no answers or ideas are discussed here, just giving a buzz word.

Yes, there are diverse opportunities arising from machine learning approaches spanning from data mining to parameterisation development. Again we do not want to make a general recommendation about what CMIP should do since it requires soliciting feedback from the much broader community involved in CMIP. We gave examples to illustrate how machine learning can be used for our research interest, specifically for “improving or speeding up process representations in ESMs, as well as designing smart tools for post-processing and evaluating ESM output”. There is explainable machine learning and algorithm code is available, such that these methods are not necessarily black boxes. Certainly care should be taken when implementing or interpreting any novel additions to modeling. In our community, we see several routes for using machine learning to advance our research. We add these routes with the following paragraph in Section 4.1.1:

“We see primarily four areas where machine learning could help in advancing the research in our community. These are (1) to include faster and more precise representations of processes in models, e.g., for replacing or modifying physical parameterizations that are thought to not work sufficiently well in all conditions in which they are needed, (2) to develop novel ways to gain a better understanding of physical and chemical interactions, e.g., through data mining employing machine learning techniques, (3) to fill observational gaps, e.g., in satellite products to allow the creation of spatially complete data to more easily validate model results against observational information, and (4) to mimic climate responses to radiative forcing, e.g., to prioritize experiments for the design of new MIP protocols. Proofs of the concept of applying machine learning in our research field exist. (...)”

Further comments:

Figures: Figure 1 contains no information that is not conveyed already in the text. Figure 2, this could be a question of preference, but again here I don’t know why a figure is needed, as the arguments between, resolution vs. complexity vs simulation length (maybe add here also ensemble size) is obvious.

We are going with the adage - a picture is worth a thousand words. We agree that the text covers what is conveyed in Figures 1 and 2 but these illustrative figures make it easier to absorb the information presented in the text. Text and figures complement each other in the manuscript. We note “ensemble size” in the revised caption of Figure 4: “ The latter axis includes the ensemble size, referring to the number of members in an experiment ensemble.”

A discussion could be useful to think if the CMIP6 TriMIPs asked for too many experiments. Was it useful to have so many tiers and experiments asked for? Would a simpler set of runs be more useful? Or was this the correct amount

to ask for? A critical assessment would fit into this paper and would allow this paper to go purely summarizing the previous experiments.

Good idea, we add a discussion at the beginning of Section 3.1 which includes a review of number of experiments and publications from the CMIP6 MIPs AerChemMIP and RFMIP:

“MIPs in CMIP6 as a whole asked for many experiments that jointly placed a big computational demand on climate modeling centers. The requested experiments were designed to address the MIP specific scientific questions. The three MIPs discussed here contributed to that demand. However, the diversity of research interests across the modeling centers meant that some experiments received more attention than others. Setting priorities with tiers was useful to the extent that it highlighted the priority of experiments from the MIP’s perspective. In so doing, the tiers guided the participating modelers to set a focus on some experiments to have a larger model ensemble where the MIPs wanted contributions the most. However, in retrospect, some of the Tier 2 experiments may have been more useful than Tier 1. An example here is piClim-histaer (Tier 2) from RFMIP, which quantified the spread in magnitude and timing of historical aerosol forcing in CMIP6 models, was informationally rich, and a contributing factor in deriving the aerosol ERF time series for AR6 WG1. Simpler and multi-purpose experiments would be useful and less burdensome (both in terms of human and computational resources) as long as they facilitate answering the science questions laid out by the MIPs. We propose enhanced coordination across the MIPs during the experimental protocol design phase potentially aid in reducing the number of experiments. .”

The performed experiments have been used in peer-reviewed publications and the results informed the sixth assessment report of the IPCC, particularly chapters 6 and 7 of Working Group 1, and the calibration of emulators which allowed climate projections from emissions scenarios in Working Group 3. We therefore perceive the experiments of the MIPs as useful. We intend to include the new Table 1 for a new analysis of the number of experiments in the MIPs and their usage in peer-reviewed publications. Associated new text is added in Section 3.1:

“One could say more performed experiments are better for obtaining more data for the statistical analysis and for addressing more research questions, and that is certainly true but not feasible in light of restricted resources. In preparation for the next phase of AerChemMIP and RFMIP, we, therefore, revisit the question of the type and number of experiments in our requests based on refined research questions that we jointly want to address as a community. We intend to keep the computational burden for modeling centers as small as possible. In this process, we coordinate our intended activities with other initiatives close to our interests, e.g., via a series of workshops organized by us and others. It could be useful to enhance further the coordination across MIPs in designing multi-purpose experiments that could be mined for several different questions. It potentially allows to free some resources and to simplify workflows, e.g., to generate larger ensembles of identical multi-purpose experiments to account for internal variability like done for CMIP historical experiments. One such experiment type from our community would be transient coupled single-forcing experiments to quantify the contributions from different anthropogenic perturbations to climate change.

In preparation for the second phase of AerChemMIP and RFMIP, we review the current status of the number of experiments and their usage in peer-reviewed publications, summarized in Table 1. A total of 67 models performed CMIP6 *historical* experiments that were used in as many as 15100 publications. Model output to assess differences in forcing and response was, however, more restrictive, e.g., AOD output is available for 45 out of the 67 models. Most of the *historical* experiments (40) are performed with emission-driven models. The ESMs with prescribed aerosols (19) in the *historical* experiments used mostly (13) MACv2-SP (Stevens et al., 2017). MACv2-SP was developed in the framework of RFMIP and is due to the relatively broad implementation in ESMs now included in the works of the CMIP climate forcing task team, although the targeted exploitation of MACv2-SP in RFMIP was with one publication small compared to the usage of other experiments of RFMIP and AerChemMIP so far.

RFMIP and AerChemMIP received output from 103 experiments leading to 204 publications to date. We separate the RFMIP and AerChemMIP experiments here into three classes, namely experiments with full coupling between the atmosphere and ocean (*hist-X)*, with prescribed sea-surface temperatures and sea-ice at pre-industrial level (*piClim-X)*, and with prescribed transient changes in sea-surface temperatures and sea-ice from a *historical* experiment (*histSST-X)*. Inter-comparing these classes, *piClim-X* experiments were performed the most with a total of 50 contributing models followed by *hist-X* with 36 experiments. However, hist-X is used three times more often in scientific publications (146) compared to *piClim-X (*52*)*. The higher computational demand of *hist-X,* therefore, seems

justified by the much larger scientific output compared to the experiments without a coupled ocean (*histSST-X* and *piClim-X*), measured by the number of published articles."

Experiment name	**MIP**	**Number of models (ESGF as of June 2023)**	**Number of publications (google scholar as of June 2023)**
historical	CMIP6	67	15100
hist-aer	DAMIP	15	111
hist-piAer	AerChemMIP	10 8 also did hist-piNTCF	21
hist-piNTCF	AerChemMIP	11 3 did only hist-piNTCF	14
Total number coupled experiments (Total excl. *historical*)		**103 (36)**	**15246 (146)**
histSST	CMIP6	12	32
histSST-piAer	AerChemMIP	7	9
histSST-piNTCF	AerChemMIP	10	7
Total number histSST experiments (Total excl. *histSST*)		**29 (17)**	**48 (16)**
piClim-control	CMIP6	23	69
piClim-histaer	RFMIP	10 (+4 not on ESGF)	16
piClim-spAer-histaer	RFMIP	1 (+3 not on ESGF)	1
piClim-aer	AerChemMIP, RFMIP	19	27
piClim-NTCF	AerChemMIP	10	7
piClim-spAer-aer	RFMIP	3	1
Total piClim experiments (Total excl. *piClim-control*)		**73 (50)**	**121 (52)**
Total (Total RFMIP and AerChemMIP)		**205 (103)**	**15415 (214)**

Table 1: Overview on existing experiments from RFMIP and AerChemMIP and their use in scientific publications.

On page 5, L 125 challenges of MIP research does point out aerosol diversity, and only cites one paper using an extremely overly simplified methods, and ignoring all the work that has been done over decades by the Aerocom community, where a deep understanding has been collected on model processes and diversity.

Thanks for pointing this out. The line explicitly refers to examples of dedicated simplification for an improved understanding, namely: "Methods to separate out some of these model differences include experiments using, for instance, prescribed aerosols (e.g. Fiedler et al., 2019) or reactive trace gases (e.g. Checa-Garcia et al., 2018), (...)".

Members of the AeroCom community are important partners in the three MIPs, act as co-authors on this manuscript, and have been influential in designing the experiment protocols of MIPs discussed here. We add here: “Such experiments have also been used for a better understanding of reasons for model differences in aerosol forcing in the AeroCom community (Stier et al., 2013) and of circulation responses to idealized aerosol forcing (Voigt et al., 2017).”.

Stier, P., Schutgens, N. A. J., Bellouin, N., Bian, H., Boucher, O., Chin, M., Ghan, S., Huneeus, N., Kinne, S., Lin, G., Ma, X., Myhre, G., Penner, J. E., Randles, C. A., Samset, B., Schulz, M., Takemura, T., Yu, F., Yu, H., and Zhou, C.: Host model uncertainties in aerosol radiative forcing estimates: results from the AeroCom Prescribed intercomparison study, Atmos. Chem. Phys., 13, 3245–3270, https://doi.org/10.5194/acp-13-3245-2013, 2013.

Voigt, A., Pincus, R., Stevens, B., Bony, S., Boucher, O., Bellouin, N., Lewinschal, A., Medeiros, B., Wang, Z., and Zhang, H. (2017), Fast and slow shifts of the zonal-mean intertropical convergence zone in response to an idealized anthropogenic aerosol, *J. Adv. Model. Earth Syst.*, 9, 870–892, doi:10.1002/2016MS000902.

The abstract reads like an introduction, but maybe that is caused by the fact that no solid conclusions are drawn in the paper.

We revise the abstract to include the take-away messages stated in the conclusions that have also been adjusted for clarity.

Revised abstract:

“The climate science community aims to improve our understanding of climate change due to anthropogenic influences on atmospheric composition and the Earth's surface. Yet not all climate interactions are fully understood and uncertainty in climate model results persists as assessed in the latest Intergovernmental Panel on Climate Change (IPCC) assessment report. We synthesize current challenges and emphasize opportunities for advancing our understanding of the interactions between atmospheric composition, air quality, and climate change, as well as for quantifying model diversity. Our perspective is based on expert views from three multi-model intercomparison projects (MIPs) - the Precipitation Driver Response MIP (PDRMIP), the Aerosol and Chemistry MIP (AerChemMIP), and the Radiative Forcing MIP (RFMIP). While there are many shared interests and specialisms across the MIPs, they have their own scientific foci and specific approaches. The partial overlap between the MIPs proved useful for advancing the understanding of the perturbation-response paradigm through multi-model ensembles of Earth System Models of varying complexity. We discuss the challenges of gaining insights from Earth System Models that face computational and process representation limits and provide guidance from our lessons learned. Promising ideas to overcome some long-standing challenges in the near future are kilometer-scale experiments to better simulate circulation-dependent processes where it is possible, and machine learning approaches where they are needed, e.g., for faster and better sub-grid scale parameterizations and pattern recognition in big model data. New model constraints can arise from augmented observational products that leverage multiple datasets with machine learning approaches. Future MIPs can develop smart experiment protocols that strive towards an optimal tradeoff between resolution, complexity, and simulation number and length, and thereby, help to advance the understanding of climate change and its impacts. ”

Revised conclusions:

“In part, the difficulty of simulating precipitation responses is related to the grand challenge of representing clouds and circulation, which can be addressed with newly evolving capabilities. Moreover, model-state dependencies affecting radiative forcing and climate responses can potentially be reduced or even resolved in the near future. Promising ideas are the use of:

- ESM experiments with prescribed land temperatures in addition to prescribed sea ice and sea-surface temperatures in more Earth System Models to quantify the effective radiative forcing free of artifacts arising from temperature adjustments over land,
- kilometer-scale model experiments with resolutions of a few kilometers to improve the understanding of interactions of atmospheric composition with circulation, clouds, and precipitation which are long-standing

challenges in climate modeling with coarse-resolution ESMs affecting for instance the representation of atmospheric composition and associated air quality assessments and aerosol-cloud interactions,
- novel machine learning approaches to speed up and improve parameterizations of sub-gridscale processes for experiments with ESMs and kilometer-scale models, to data mining for pattern recognition in big model output and to develop augmented observational products for new constraints on model output,
- emulation techniques to mimic climate responses to different forcings within the solution space of existing experiments to reduce the computation burden on modeling centers, and
- sufficiently long experiments or many ensemble members for an experiment to better distinguish climate and air quality responses to atmospheric composition changes from internal variability and therefore substantially reduce the risk of ambiguity in attributing responses to anthropogenic perturbations."

Page 6 Line 162 etc, This is referring to the comment: Process complexity is not reducing uncertainty. It should not be the goal to have all models to agree with each other. Model diversity should be the goal. Increased 'uncertainty' comes with the territory of increased complexity. Understanding 'uncertainty' model diversity should be the goal. The community will never get to a place where all models agree, neither should they.

The comment refers to the sentence: "High process complexity, although desirable and needed, also poses a challenge to reducing uncertainty in the assessment of the climate response to various forcings." Maybe our words can be interpreted differently than was intended. We did not mean to imply that models should all agree. We clarify the meaning in the revised manuscript by adding: "Model diversity in terms of for instance the combination of parameterizations, number and fidelity of represented processes, choice of coupling of model components, choice of the dynamical core, and the resolution is wanted. Differences in model results cannot be fully resolved to reach a perfect model agreement, but the models ideally converge to similar solutions for a given question, e.g., how the Earth's temperature responds to anthropogenic perturbations. The diversity in model results should therefore reduce over time to gain confidence in our conclusions drawn from simulated responses to imposed perturbations. AerChemMIP emphasized two challenges for reducing uncertainty. (...)"

Page 8, the reason behind differences in CMIP modelled dust trends and observational evidence, lies in the fact that most/all CMIP models are not coupled to dynamic terrestrial/dynamic vegetation models, and as such missing many feedbacks that lead to changes in dust emissions. As surface process modeling on CMIP time scales is extremely difficult, hybrid approaches could be investigated to get more interactions between emissions and ESM processes included in future CMIP experiments. This goes beyond dust, and is true for many other emission sources.

We use the simulation of dust across models as an example to describe the process understanding abyss. In the revised manuscript, we include a new overview table that visually highlights the gaps (see Table 4 in the reply to reviewer 2). We further add in the revised text: "Modeling surface conditions is a challenge and a potential source of the diversity in simulated dust trends. Not all models participating in CMIP6 have the capability to simulate interactive vegetation dynamics but some do, e.g., UKESM and GFDL-ESM4. A lack of coupled vegetation dynamics is not the only potential reason for differences in dust and other aerosols. Winds emit and transport dust aerosols and the soil erodibility is influenced by moisture from rain events, but both regional changes in circulation and precipitation differ across models such that their changes with warming are not fully understood." It would be great to look into these dependencies in depth in future model studies. Potential ways forward are presented in the revised Section 4.1.2.

The Radiative forcing paragraph, what is new here that has not already been summarized in the papers cited in this section? Conclusion Nr 1: using prescribed SST's, this is AMIP, which is part of CMIP.

We give an overview of available methods. In the conclusion, we stated: "prescribed land and ocean surface temperatures". This is different from prescribing sea-surface temperature and sea ice only, which is done in AMIP included in the CMIP6 experiment protocol. In the revised conclusion, we make this point clearer by writing: "prescribed sea ice, land- and sea-surface temperatures". Only a few models have carried out experiments in which both the sea AND land surface temperatures have been prescribed. This is described in the forcing section of the manuscript ("If the capability of fixed land-surface temperatures (Andrews et al., 2021) was facilitated in more ESMs, biases in ERF arising from surface temperature adjustments would be virtually eliminated in the future. By adopting the fixed sea- and land-surface temperature method (Figure 6), the change in the radiation budget would then be

equal to the change in the energy budget of the system, which overcomes the limitations of other methods for estimating ERF.") and we will elaborate more on it for clarity as follows:

"Prescribing sea- and land-surface temperature is different from the experiments carried out for CMIP6 and RFMIP. The requested experiments used prescribed sea-surface temperatures and sea ice following the experimental design of the Atmosphere Model Intercomparison Project (AMIP, Gates, 1992), but the land-surface temperatures were freely evolving. Prescribing the sea ice, sea- and land-surface temperatures has not been done in a MIP to date."

We further rephrase the conclusion to make the distinction clearer at that point too: "(...) performing experiments with prescribed land temperatures in addition to prescribed sea ice and sea-surface temperatures in more Earth System Models to quantify the effective radiative forcing free of artifacts arising from temperature adjustments over land (...) "

Gates, W. L., 1992: AN AMS CONTINUING SERIES: GLOBAL CHANGE--AMIP: The Atmospheric Model Intercomparison Project. *Bull. Amer. Meteor. Soc.*, **73**, 1962–1970, https://doi.org/10.1175/1520-0477(1992)073<1962:ATAMIP>2.0.CO;2.

In summary, I feel this paper can't be fixed by adding some more aspects to the chapters, I feel it needs much more holistic ideas how to push forward and eventually needs a complete rewrite.

Based on the comments from both the reviewers, we substantially revise the manuscript including highlighting new ideas and more details for clarifying content. An example is in conclusion number 1 which requires new experimental setups that not all models are yet able to do and was not done in any existing MIP.

Reply to Reviewer 2

[General]

I think this work serves several purposes. It reports on the different MIPs, on their interaction, on their challenges and future opportunities, and is a vision paper for the CACTI activity, which might be a future link between the different MIPs. Learning from the past and stressing what are the large/important remaining knowledge gaps, is possibly important to guide future research directions.
I think the work of these MIPs in the past 5 to 10 years has been very valuable, and their importance can hardly be overstated. These MIPs contributed largely to trying to understand conceptually the chain from initial perturbations to climate responses.
In this manuscript, the authors have put the different MIPs in a similar framework, to allow them to characterize what they differ in and what they have in common. It is nice that this manuscript tries to connect the various MIPs - such an approach can never be expected to be fully comprehensive or satisfying, but I very much appreciate the work of the authors.
It is also important to (critically) look backward to earlier existing MIPs, and synthesize ideas on ways to go forward. One can expect that MIPs will be popular for some time, but might loose interest over time. One can also expect that a new generation of scientists will come up with new ideas and areas to focus on. Within this environment, it is nice that these three MIPs have worked (closely) together (e.g., TriMIP), try to reflect on and synthesize their achievements (this manuscript), and make efforts to find synergies for the future (CACTI).
The manuscript comes up with suggestions for future research directions, and I am fully aware that it is not always easy to be precise in general prospective studies. Coming up with a list of possible future directions is brave, and one cannot expect such lists to be complete.
I think this work is valuable and is appropriate for the journal. E.g., the original description papers for RFMIP and AerChemMIP were both published in GMD. Having this prospective paper in the suggested journal is therefore appropriate.
I think however that the manuscript needs modifications before being fit for publication. Below you can first find a list of my main concerns. It is followed by a more detailed list, referring to specific lines in the manuscript. Both the main concerns and the detailed comments should be addressed by the authors.

Thank you for the positive appraisal of our manuscript and the well-reflected comments on the content. Below are our replies to the suggestions including how we intend to revise the manuscript in response to the comments.

[Main comments]

(1)
In the backward looking view of the manuscript, one reports mainly on the collaboration, interaction and links between the different MIPs. However, based on the title of the manuscript, one would expect that it would also report on the general "scientific progress" (or lack of it) made within the field due to these MIPs. In that sense, I think that the title and content of Section 2 (Advancement through MIP's cross-linkages) is rather limited, and should maybe be widened up to also go more deeply in the scientific progress made. I think Figures 4 and 5 cover part of the scientific progress, but I think more achievements can/should be mentioned to motivate the existence of these MIPs. This should not be very long, but possibly a table with main scientific lessons learned from the MIP experiments might be an idea.

Agreed, there can be more detail on the scientific progress through the three MIPs. To that end, we split Section 1 with the revised title “Scientific Advancement” into a new Section 2.1 on “MIP's Key Results” and Section 2.2 “MIP's Cross-linkages”. We include a newly designed table like suggested and add accompanying text in the new Section 2.1 like follows: „Table 2 summarizes key results along with the used experiments, organized by topics that were addressed by the three MIPs.

The primary objective of PDRMIP is to understand global and regional responses of precipitation statistics to the different forcing agents CO2, CH4, O3, irradiance, sulfate and black carbon aerosols (Myhre et al., 2017). Based on eleven aerosol-climate models contributing experiments to PDRMIP, energy budgets and the hydrological cycles were inter-compared for fast (days to months) and slow (years to decades) response times (e.g., Myhre et al., 2017,Samset et al., 2016, Sillmann et al., 2019). Rapid adjustments are a key in understanding precipitation responses (e.g., Hodnebrog et al., 2020, Myhre et al., 2018, Smith et al. 2018). Taking advantage of multiple forcing agents in PDRMIP, model spreads in radiative forcing and efficacy for the forcing agents are quantified (Forster et al., 2016, Richardson et al., 2019), and responses to greenhouse gases and aerosols inter-compared across the PDRMIP ensemble (Sillmann et al., 2019, Stjern et al., 2020). Others examined the climate response to forcing for selected regions, e.g., the monsoon regions, the Arctic, and the Mediterranean (Stjern et al., 2019, Tang et al., 2018, Xie et al., 2020).

The main goals of AerChemMIP are to quantify the climate and air quality responses of aerosols and chemically reactive gases, specifically near-term climate forcers (NTCFs) including methane, tropospheric ozone, aerosols, and their precursors (Collins et al., 2017). Amongst TriMIP, AerChemMIP emphasizes transient coupled atmosphere-ocean simulations to estimate the real-world evolution and timing of emission changes and associated climate responses. AerChemMIP experiments are novel in CMIP6 in that they follow the “all-but-one” design, whereby the forcing of interest is held fixed. For example, *hist-piNTCF* simulations are parallel to *historical* simulations, except anthropogenic emissions of NTCFs are held fixed at pre-industrial level (1850) and all other forcing agents evolve as in a historical simulation. The climate impacts are diagnosed by subtracting the perturbed runs from the historical climate experiment. Such an experimental design seeks to minimize the contribution of non-linear climate responses that may occur under the more traditional experimental design where the emissions or concentrations of the species of interest are perturbed (Deng et al., 2020). The model output from AerChemMIP is used to investigate 21st-century climate and air quality responses to future NTCF changes (Allen et al., 2020, Allen et al., 2021).

Another focus of AerChemMIP is quantifying natural emission feedbacks (Thornhill et al., 2021a) with an AerChemMIP-specific experimental design that is unique in CMIP6. That is a set of idealized simulations with fixed boundary conditions, except for the preindustrial natural emissions or concentrations that are systematically doubled across the ensemble of simulations, e.g., for dust aerosols *piClim-2xdust*. Paring the effective radiative forcing (ERF) from these experiments gives ERF per Tg yr-1 change in emissions or concentrations of the climate forcer. The result allows to obtain the feedback parameter (W m-2 per K) for the climate forcer through scaling the simulated changes in emission fluxes per K temperature change from either the 4xCO2 or 1\% yr-1 CO2 experiments of CMIP6. The protocol of AerChemMIP also includes transient historical simulations with prescribed sea-surface temperatures (SSTs) to diagnose transient ERFs. Similar to the coupled experiments, these simulations followed the “all-but-one”

experimental strategy. Including such analogous prescribed-SST experiments allows for a better understanding of the drivers of the climate response in the fully coupled experiments. Furthermore, time-slice experiments are performed with emissions of one species set to the present-day value but all other boundary data held fixed at pre-industrial values (Thornhill et al., 2021b).

RFMIP strives for accurately quantifying and identifying errors in the radiative forcing of composition changes in CMIP6 models (Pincus et al., 2016). The largest of the three parts of RFMIP (RFMIP-ERF) is the quantification of ERF across CMIP6 models using a time-slice approach similar to AerChemMIP. It allowed the first quantification of the CMIP inter-model spread in ERF for all major climate forcers as bulk estimates, i.e., for all anthropogenic aerosols taken together, and of the contribution from rapid adjustments to ERF (Smith et al., 2018, 2020). The second part of RFMIP (RFMIP-IRF) focuses on the instantaneous radiative forcing (IRF) excluding contributions from rapid adjustments. Errors in IRF of greenhouse gases are identifiable using benchmark calculations from line-by-line models \citep{pincus2020}. The third RFMIP part (RFMIP-SpAer) assesses model differences in ERF for identical anthropogenic aerosol optical properties and effects on clouds. Participating in RFMIP-SpAer requires implementing the simple-plumes parameterization (Stevens et al., 2017), which is a new approach in CMIP6. The pilot study for RFMIP-SpAer documents the retention of a model spread in ERF when moving to identical anthropogenic aerosols due to differences in the atmospheric host models (Fiedler et al., 2019). Through the combined analysis of output from RFMIP-ERF and RFMIP-SpAer, reasons for model differences in anthropogenic aerosol forcing were inferred (Fiedler et al., 2023).”

Topic	**Experiments**	**Key results**	**References**
Atmospheric Composition	hist, histSST-X	Model consistency in historical OH trends driven by NTCF emissions; Evolution of tropospheric and stratospheric ozone, its attribution to different drivers, and evaluation against observations; Evaluation of aerosol lifecycle, optical properties and trends in CMIP6 generation models	Stevenson et al. (2020) Griffiths et al. (2021) Keeble et al. (2021) Zeng et al. (2022) Gliß et al (2021) Mortier et al (2020)
Air Quality (AQ) & Human Health	hist, sspX.Y, ssp370-lowNTCF, ssp370-lowNTCFCH4, ssp370SST ssp370pdSST ssp370SST-X	Historical and future evolution of air pollution; Climate penalty and benefit for surface ozone; Impact of climate mitigation on AQ and human health	Turnock et al. (2020) Zanis et al. (2022) Brown et al. (2022) Allen et al. (2020, 2021) Turnock et al. (2022) Turnock et al. (2023)
Radiative Forcing	hist, histSST-X, piClim-X, ssp370SST-X	Recommendations for diagnosing forcing from CMIP models; Estimates of rapid adjustments for different forcing agents; First estimates of present-day effective radiative forcing in CMIP; Historical evolution of ozone forcing; Observationally-constrained estimate of present-day halocarbon forcing; Observationally constrained time series of historical aerosol forcing; Role of chemistry-aerosol-cloud coupling in estimates of forcing; Impact of climate mitigation measures on climate forcing; new anthropogenic aerosol parameterization for use in CMIP6; little change in aerosol forcing between 1970s and 2000s	Forster et al. (2016) Smith et al. (2018) Smith et al. (2020) Skeie et al. (2020) Morgenstern et al. (2020) Smith et al. (2021) Thornhill et al. (2021) O’Connor et al. (2021) O’Connor et al. (2022) Turnock et al. (2022) Stevens et al. (2017) Fiedler et al. (2017, 2019, 2023)
Climate Response	piControl, hist, hist-piAer, ssp370,	Climate and AQ impacts of mitigating SLCFs and non-methane SLCFs; Fast	Allen et al. (2020, 2021) Zanis et al. (2020)

	ssp370-lowNTCF, ssp370-lowNTCFCH4, piClim-X	responses from aerosols in PI climate; Role of aerosols in historical climate; Impact of SLCFs on Atlantic Meridional Overturning Circulation; Regional climate extremes; Fast and Slow Precipitation responses	Zhang et al. (2021) Hassan et al. (2022) Li et al. (2023) Samset et al. (2016)
Non-CO2 Biogeochemical Feedbacks	piControl, Abrupt-4xCO2, piClim-2xEms	First multi-model estimates for biogeochemical feedbacks	Thornhill et al. (2021)

Table 2: Key results from the three MIPs for their research topics.

Added references:

Brown, F., Folberth, G. A., Sitch, S., Bauer, S., Bauters, M., Boeckx, P., Cheesman, A. W., Deushi, M., Dos Santos Vieira, I., Galy-Lacaux, C., Haywood, J., Keeble, J., Mercado, L. M., O'Connor, F. M., Oshima, N., Tsigaridis, K., and Verbeeck, H. (2022). The ozone–climate penalty over South America and Africa by 2100, Atmos. Chem. Phys., 22, 12331–12352, https://doi.org/10.5194/acp-22-12331-2022.

Fiedler, S., van Noije, T., Smith, C. J., Boucher, O., Dufresne, J.-L., Kirkevåg, A., et al. (2023). Historical changes and reasons for model differences in anthropogenic aerosol forcing in CMIP6. *Geophysical Research Letters*, 50, e2023GL104848. https://doi.org/10.1029/2023GL104848

Hassan, T., Allen, R.J., Liu, W. *et al.* (2022). Air quality improvements are projected to weaken the Atlantic meridional overturning circulation through radiative forcing effects. *Commun Earth Environ* 3, 149, https://doi.org/10.1038/s43247-022-00476-9.

Li, Y., Wang, Z., Lei, Y., Che, H., and Zhang, X. (2023). Impacts of reductions in non-methane short-lived climate forcers on future climate extremes and the resulting population exposure risks in eastern and southern Asia, Atmos. Chem. Phys., 23, 2499–2523, https://doi.org/10.5194/acp-23-2499-2023.

O'Connor, F. M., Johnson, B. T., Jamil, O., Andrews, T., Mulcahy, J. P., & Manners, J. (2022). Apportionment of the pre-industrial to present-day climate forcing by methane using UKESM1: The role of the cloud radiative effect. *Journal of Advances in Modeling Earth Systems*, 14, e2022MS002991. https://doi.org/10.1029/2022MS002991.

Smith, C. J., Kramer, R. J., Myhre, G., Forster, P. M., Soden, B. J., Andrews, T., et al. (2018). Understanding rapid adjustments to diverse forcing agents. *Geophysical Research Letters*, 45, 12,023–12,031. https://doi.org/10.1029/2018GL079826.

Skeie, R.B., Myhre, G., Hodnebrog, Ø. *et al.* (2020). Historical total ozone radiative forcing derived from CMIP6 simulations. *npj Clim Atmos Sci* 3, 32, https://doi.org/10.1038/s41612-020-00131-0.

Turnock, S. T., Allen, R., Archibald, A. T., Dalvi, M., Folberth, G., Griffiths, P. T., et al. (2022). The future climate and air quality response from different near-term climate forcer, climate, and land-use scenarios using UKESM1. *Earth's Future*, 10, e2022EF002687. https://doi.org/10.1029/2022EF002687.

Turnock, S. T., Reddington, C. L., West, J. J., & O'Connor, F. M. (2023). The air pollution human health burden in different future scenarios that involve the mitigation of near-term climate forcers, climate and land-use. *GeoHealth*, 7, e2023GH000812. https://doi.org/10.1029/2023GH000812.

Zeng, G., Morgenstern, O., Williams, J. H. T., O'Connor, F. M., Griffiths, P. T., Keeble, J., et al. (2022). Attribution of stratospheric and tropospheric ozone changes between 1850 and 2014 in CMIP6 models. *Journal of Geophysical Research: Atmospheres*, 127, e2022JD036452. https://doi.org/10.1029/2022JD036452.

Zhang, J., Furtado, K., Turnock, S. T., Mulcahy, J. P., Wilcox, L. J., Booth, B. B., Sexton, D., Wu, T., Zhang, F., and Liu, Q. (2021). The role of anthropogenic aerosols in the anomalous cooling from 1960 to 1990 in the CMIP6 Earth system models, Atmos. Chem. Phys., 21, 18609–18627, https://doi.org/10.5194/acp-21-18609-2021.

(2)
The study is generally well written and structured, although specific sections should be improved for better understanding. I also think that there should be an effort in making the style more homogeneous over the whole manuscript. Also, one sometimes gets the impression that some paragraphs (or sentences) do not really belong in a section : they should be brought more in harmony with the text around them. This is indicated in the detailed list below.

Thank you, we have revised the text for coherence and clarity. There will be changes in the structure for clarity, namely adjusted titles of (sub)sections, combining two subsections into one, and adding a new subsection. Please refer to the manuscript with color-highlighted changes in the text..

(3)
Some synergies between the MIPs are mentioned, but one should maybe make an effort to elaborate more on this. Some examples have been given (use the same ERF definition, estimate length of simulations, diagnostics, complementary simulations, ...), but maybe the authors could try to analyze and refine it more.

Thanks, we will elaborate more in detail on synergies. We add the following in Section 2.2 :

„RFMIP asked for experiments to diagnose radiative forcing for greenhouse gases and aerosols as bulk quantities with setups parallel to DECK experiments. As such, RFMIP was able to characterise forcing in CMIP for the first time. Due to the parallel setup of the RFMIP experiments to those requested in DECK and additional overlap of experiment requests with other MIPs (DAMIP), RFMIP experiments also allowed model analyses of climate responses and climate feedbacks for well-estimated radiative forcing. AerChemMIP further separated contributions to radiative forcing into individual gases and short-lived climate forcers including different aerosol species. As such, the AerChemMIP experiment request was tailored to gain insights into why model differences in the forcing-response paradigm arise based on individual perturbations in atmospheric composition. The RFMIP tier 1 experiments were carried out by many modeling centers. Some of these contributions, e.g., from UKESM1 and CNRM, arose because the experimental setup was identical to the request in AerChemMIP. It meant that the technical workflow for performing and postprocessing the experiments was already in place such that contributing another variant of such experiments required only little effort.

Experiment requests that were differently designed in RFMIP and AerChemMIP for a similar purpose were the transient historical experiments to identify the response to individual perturbations. Specifically, RFMIP varied the quantity to be assessed over the historical period while keeping all other boundary conditions at the pre-industrial level (piClim-histX, where X is the forcing of interest), whereas AerChemMIP held the quantity to be assessed at the pre-industrial level and varied the boundary conditions for all other climate forcers over the historical period (histSST_piX). These differences in the setup hold the potential to understand where interactions and potential feedbacks arising from chemical composition changes play a role for the climate response, which has not yet been fully explored with the existing model output from the MIPs.“

We comment on the time period for forcing calculations in Section 4.2.1: “RFMIP requested 30-year long experiments for ERF calculations (Pincus et al., 2016) following earlier recommendations based on CMIP5 output (Forster et al., 2016). That experiment length proved to be sufficient for CMIP6 models contributing experiments to RFMIP, e.g., for ERF of anthropogenic aerosols although more simulated decades further improve the precision of the ERF calculation (Fiedler et al., 2017, 2019). Differently from RFMIP, AerChemMIP found that a spin-up time associated with the long-lived trace gases, e.g. halocarbons, is necessary before calculating the ERF. This meant that the approach of 30 year long time slice experiments was not entirely appropriate for the AerChemMIP experiments for all individual climate forcers. The longer spin-up period should be accounted for in future requests for new experiments for ERF calculations of such climate forcers. ”

Forster, P. M., Richardson, T., Maycock, A. C., Smith, C. J., Samset, B. H., Myhre, G., Andrews, T., Pincus, R., and Schulz, M. (2016), Recommendations for diagnosing effective radiative forcing from climate models for CMIP6, *J. Geophys. Res. Atmos.*, 121, 12,460– 12,475, doi:10.1002/2016JD025320.

(4)
I am wondering whether the topics which are treated under Section 4 (Methodological opportunities) are representative for the future research environment and directions. I also do not know why it should be limited to "methodological" opportunities only - that is certainly not what the title of the manuscript suggests. I have the impression that the mentioned opportunities are all relevant, but some of them are maybe rather specific, and their total does not seem to cover the broad opportunity space. E.g., although kilometer-scale experiments might be important, they seem rather far in the future for most of the ESMs which have resolutions in the order of 50-200 km. As MIPs are designed for (many) models to participate, one should, e.g., take into consideration that part of the models are maybe not at the forefront of the scientific progress. Although the 5 topics mentioned in Section 4 are relevant, I could imagine a longer, more comprehensive and equilibrated list of relevant research questions and directions. My general impression is that these 5 opportunities do not cover well enough the directions in which these MIPs probably will go. Possibly an extended list of these opportunities could be put in a table.

We kept the exact direction of the future MIPs open as the process of defining the research goals and experimental requests for new future MIPs in CMIP7 out of the CACTI community are only now beginning. This process should be bottom-up driven by the scientists as much as possible, hence we wanted to provide an overview in this perspective piece to support a broader awareness of opportunities for others that we think are useful for the development of the next MIP protocols from CACTI. For the point on kilometer-scale experiments, we agree that a MIP is not possible with the most comprehensive ESMs on the CMIP7 timeline.

There are ways to accelerate such a development in the future, if the scientific community would strive towards it. One such way is to use machine learning to make models faster and better at representing sub-grid scale processes which includes aerosols and chemistry. This aspect is now more clearly discussed in Section 4.1.1: “(...) contribute to improving or speeding up process representations in ESMs, as well as designing smart tools for post-processing and evaluating ESM output. We see primarily four areas where machine learning could help in advancing the research in our community. These are (1) to include faster and more precise representations of processes in models, e.g., for replacing or modifying physical parameterizations that are thought to not work sufficiently well in all conditions in which they are needed, (...)”

Another way is to revisit the use of regional downscaling that in the past has been successfully applied to obtain spatial resolutions on limited-area domains of a few tens of kilometres that today are now also feasible with global models. We would elaborate more on these points with links to CORDEX in the revised manuscript.

A third way could be to include simpler models in MIPs that complement the complex model experiments to diagnose responses of interest. One example is the use of simulated atmospheric and soil conditions from CMIP models as input for aerosol emission models to understand model differences in dust responses, e.g., to pinpoint reasons for model differences in dust-aerosol emissions. This point is also newly added to the manuscript.

We add the text in Section 4.1.2:
“The community of the three MIPs will not be able to entirely rely on global kilometer-scale model experiments in CMIP7, especially in the context of a MIP since only a few models with the necessary capability for coupled aerosols and chemistry might exist in time. In light of this restriction, we see two main routes forward for immediately using spatially refined information in our next MIPs. The first possible and computationally smart way is to use the output from global kilometer-scale experiments that are run for other purposes to drive offline models for aerosols and chemistry or atmospheric radiation transfer calculations. This approach is suitable to answer some but not all research questions in our community. For instance, the response of dust emission fluxes to changes in winds and moisture can be addressed with offline modeling and allows to identify underlying reasons for changes and model differences in the dust response (Fiedler et al., 2016), but the implication of such dust emission changes for air quality and climate responses can not be quantified with offline modeling. For the latter, regional one- or two-way dynamical

downscaling approaches could be used. We perceive dynamical downscaling as the second main avenue for our near-future works to obtain regionally refined spatial information. Regional climate modeling is already well developed and organized via CORDEX with a focus on providing regional climate change information. Regional climate models with the capability to perform experiments with coupled aerosols and chemistry exist for instance in Europe and the US (e.g., Pietikainen et al., 2012, Schwantes et al., 2022), but have not been used in our past MIPs. For CMIP7, UKESM2 and CESM aim also to have regional model configurations. An ensemble of regional composition-climate models therefore exists and could be used in future MIPs. The regional models will nevertheless need output from global climate model experiments with coupled aerosols and chemistry as boundary data for performing the regional experiments. As such our need for experiments with classical global ESMs is retained, at least for CMIP7, although we are not averse to the idea of moving towards global kilometer-scale modeling with a sufficient coupling of physical processes to aerosols and chemistry to address the community's research interests."

We change the title of section 4 to "Opportunities" to better reflect the content and list opportunities in a new Table 3 for a better overview:

Theme	**Opportunities**
Machine learning	Development of new parameterization schemes that are faster and better than existing schemes or learning from machines
	Data mining to better characterize processes in big data
	New observational products to constrain model simulations
	Improvements of emulators to better inform decisions for future experiments with ESMs
Kilometer-scale modeling	More resolved physical processes that potentially better link changes in atmospheric composition to clouds and circulation
	Possibly better regional information on climate change and air quality impacts
	Global quantification of scale-dependence of forcing and response from synoptic to submesoscale
	Better understanding of scale-dependent processes relevant for atmospheric composition, such as natural emissions including mineral dust, marine organics, and others

Table 3: List of opportunities for our community that can arise from novel capabilities.

(5)
In Section 3.2 open research questions are mentioned. However, that section is rather short, and I have the impression that it is possibly under-valued. To maybe stress the ideas in this section, it might be an idea to elaborate them a bit more, and possibly list them in a table in the manuscript. Some of these topics are : non-DMS marine volatile organic compounds, natural primary biological aerosol particles, fire, dust, natural aerosol, aerosol optical properties, and aerosol optical depth.

Good idea, we design an overview table and add in the the text: "Known gaps are listed in Table 4.". We elaborate more on some examples like dust aerosols.

Topic	**Gap in knowledge**
Non-DMS marine volatile organic compounds	Forcers are not well represented in ESMs, uncertrainties in process understanding
Natural primary biological aerosol particles	Not included in ESMs

Fire	Interactive fires are not simulated by most ESMs and their role for climate changes can therefore currently not be assessed
Mineral dust	Unclear trends of dust aerosol concentrations in a warming world, role of anthropogenic versus natural dust emissions on forcings and feedbacks
Natural aerosol	Pre-industrial state of aerosol burden and properties
Aerosol optical properties	Aerosol absorption substantially differs across ESMs
Aerosol optical depth	Unknown reasons for large model spread in aerosol optical depth
Aerosol-cloud interactions	Unclear resolution dependency of the magnitude on global scales
Emissions inventories of near-term climate forcers	Emissions not well characterized
Biogenic VOCs	Not included in all ESMs and wide variety of emissions and responses in those that do

Table 4: List of gaps in our knowledge for different topics of the three MIPs.

(6)
I miss in the manuscript some perspective on what we have learned from approaches which did not work (if so) in these three MIPs. Although this should not be the focus of the manuscript, these questions might come up for a reader. In addition, such reflections might contain important lessons for future activities, and improve the effectiveness of future research and MIPs. I list here some thoughts which might be covered.

Thank you for the useful questions. We address them in the following and will add details in the manuscript to reflect these points.

- Were there MIP experiments with a too small signal-to-noise ratio to be useful? Did the suggestions from Forster et al. [2016] appear to be correct? Is it valid for all variables, on all scales : regional and seasonal, or global and annual? Were the simulation lengths and number of simulations appropriate in the MIPs? Were the ensemble sizes (number of identical setups by individual model) large enough?

We add the following in Section 3.1: “The necessary data amount for separating the climate response from internal variability depends on the scientific interest. The response to variability ratio is for instance sufficiently good for the effective radiative forcing in the global multi-annual mean for most climate forcers. The suggestion from Forster et al. (2016) for performing 30 years of model experiments with the same boundary data, therefore, proved useful to diagnose global effective radiative forcing in most time-slice experiments, except for land-use changes (piClim-lu, Smith et al., 2020). We learned that the exact precision of the simulated effective radiative forcing depends on the model due to model differences in the internal variability that induced radiative perturbations from year to year (Fiedler et al., 2019, 2023). Longer simulations of 45 years are needed to diagnose the forcing of some longer-lived trace gases due to the time scale for gas transport through the stratosphere via the Brewer-Dobson circulation (O’Connor et al., 2021). For regional radiative effects, the 30 and 45 year long simulations are not sufficiently long in all regions to obtain a statistically significant for all anthropogenic perturbations. In UKESM, the aerosol radiative effects are for instance statistically significant at the 95% level over about 50% of the globe, but the effects are only statistically significant for 10% of the globe for land use and non-methane ozone precursors (O’Connor et al., 2021). Similarly, regional aerosol forcing is not statistically significant over all world regions in models contributing to RFMIP (Fiedler et al., 2019, 2023).

For model responses, the ensemble sizes and lengths were not sufficient for addressing all research questions of interest in the three MIPs. This is particularly true for regional responses that require a larger number of simulations or longer averaging for sufficiently reducing the impact of model-internal variability on the climate response. The

ensemble sizes were, however, typically large enough for quantifying annual and global responses, e.g., for the global multi-annual mean of precipitation (Myhre et al., 2018, Allen et al., 2020). Quantifying the regional response of climate to forcing requires larger ensembles of simulations, which the Regional Aerosol MIP (RAMIP, Wilcox et al., 2023) is currently addressing through requesting larger ensembles of experiments with regional perturbations of aerosols than available from AerChemMIP. Typically larger data amounts and/or magnitudes of a perturbation with decreasing spatial scale are necessary for separating a response from the internal variability."

O'Connor, F. M., Abraham, N. L., Dalvi, M., Folberth, G. A., Griffiths, P. T., Hardacre, C., Johnson, B. T., Kahana, R., Keeble, J., Kim, B., Morgenstern, O., Mulcahy, J. P., Richardson, M., Robertson, E., Seo, J., Shim, S., Teixeira, J. C., Turnock, S. T., Williams, J., Wiltshire, A. J., Woodward, S., and Zeng, G.: Assessment of pre-industrial to present-day anthropogenic climate forcing in UKESM1, Atmos. Chem. Phys., 21, 1211–1243, https://doi.org/10.5194/acp-21-1211-2021, 2021.

Wilcox, L. J., Allen, R. J., Samset, B. H., Bollasina, M. A., Griffiths, P. T., Keeble, J., Lund, M. T., Makkonen, R., Merikanto, J., O'Donnell, D., Paynter, D. J., Persad, G. G., Rumbold, S. T., Takemura, T., Tsigaridis, K., Undorf, S., and Westervelt, D. M.: The Regional Aerosol Model Intercomparison Project (RAMIP), Geosci. Model Dev., 16, 4451–4479, https://doi.org/10.5194/gmd-16-4451-2023, 2023.

- RFMIP and AerChemMIP had quite some distinct and complementary experiments. Possibly, there were also some experiments which better would have been combined into one experiment. Is that something that should be better organized in the future?

This is a very good point, which we are currently addressing for RFMIP2 and AerChemMIP2. There is potential to even better organize the coordination of different experiments which we are currently discussing for the next phase of the MIPs. We reflect on the differences in experimental designs in the three MIPs in the new Section 2.1 (see reply to reviewer one).

And in Section 2.2: "Experiment requests that were differently designed in RFMIP and AerChemMIP for a similar purpose were the historical experiments to identify the response to individual perturbations. Specifically, RFMIP varied the quantity to be assessed over the historical period while keeping all other boundary conditions at the pre-industrial level, whereas AerChemMIP held the quantity to be assessed at pre-industrial level and varied the boundary conditions for all other climate forcers over the historical period. These differences in the setup hold the potential to understand where interactions and potential feedbacks arising from chemical composition changes play a role for the climate response, which has not yet been fully explored with the existing model output from the MIPs."

- Was the degree of participation in these MIPs sufficient? I assume RFMIP and AerChemMIP (and maybe also PDRMIP) were aimed to explain the results/behaviour of the CMIP models (explain better the final model response and its diversity). However, not all CMIP models contributed to (all) RFMIP and AerChemMIP experiments. Should having more CMIP models participating in AerChemMIP/RFMIP/PDRMIP be a priority?

It would be fantastic if we can win more modelling centres to do the RFMIP2 and AerChemMIP2 experiments. From our perspective these experiments are crucial to quantify the effective radiative forcing and understand associated climate responses in CMIP7. In RFMIP, the participation for the quantification of effective radiative forcing from Tier 1 experiments was very good, with up to 20 models for some experiments. It would have been very useful to have diagnostic output for aerosol optical properties from more models to explain the differences in forcing. In AerChemMIP less models participated, due to less models being able to do such experiments for individual climate forcers compared to RFMIP and the computational demand on modeling centers. An additional reason was the requirement to perform the DECK experiments as a prerequisite to participate in CMIP6 and hence AerChemMIP which chemical transport models cannot do, although their simulations might have been informative for addressing the research question. This has led to some AerChemMIP articles with only few participating models (e.g., Griffiths et al., Stevenson et al.,). In AerChemMIP2, there is the potential to include more types and numbers of models to increase the ensemble size for specific questions, even if they cannot contribute to all scientific aims of the protocol.

We add in Section 2.2: “RFMIP asked for experiments to diagnose radiative forcing for greenhouse gases and aerosols as bulk quantities with setups parallel to DECK experiments. As such, RFMIP was able to characterise forcing in CMIP for the first time. Due to the parallel setup of the RFMIP experiments to those requested in DECK and additional overlap of experiment requests with other MIPs (DAMIP), RFMIP experiments also allowed model analyses of climate responses and climate feedbacks for well-estimated radiative forcing. AerChemMIP further separated contributions to radiative forcing into individual gases and short-lived climate forcers including different aerosol species. As such, the AerChemMIP experiment request was tailored to gain insights into why model differences in the forcing-response paradigm arise based on individual perturbations in atmospheric composition. The RFMIP tier 1 experiments were carried out by many modeling centers. Some of these contributions, e.g., from UKESM1 and CNRM, arose because the experimental setup was identical to the request in AerChemMIP. It meant that the technical workflow for performing and postprocessing the experiments was already in place such that contributing another variant of such experiments required only little effort.”

We further add a new Table 1 (see reply to reviewer 1) and new text in Section 3.1: “One could say more performed experiments are better for obtaining more data for the statistical analysis and for addressing more research questions, and that is certainly true but not feasible in light of restricted resources. In preparation for the next phase of AerChemMIP and RFMIP, we, therefore, revisit the question of the type and number of experiments in our request based on refined research questions that we jointly want to address as a community. In this process, we coordinate our intended activities with other initiatives that are close to our interests, e.g., via a series of workshops organized by us and others. It could be useful to further enhance the coordination across MIPs in designing multi-purpose experiments that could be mined for several different questions. It potentially allows to free some resources and to simplify workflows, e.g., to generate larger ensembles of identical multi-purpose experiments to account for internal variability like done for CMIP historical experiments. One such experiment type from our community would be transient single-forcing experiments to quantify the contributions from different anthropogenic changes.

In preparation for the second phase of AerChemMIP and RFMIP, we review the current status of the number of experiments and their usage in peer-reviewed publications, summarized in Table 1. A total of 67 models performed CMIP6 *historical* experiments that were used in as many as 15100 publications. Model output to assess differences in forcing and response was, however, more restrictive, e.g., AOD output is available for 45 out of the 67 models. Most of the *historical* experiments (40) are performed with emission-driven models. The ESMs with prescribed aerosols (19) in the *historical* experiments used mostly (13) MACv2-SP (Stevens et al., 2017). MACv2-SP was developed in the framework of RFMIP and is due to the relatively broad implementation in ESMs now included in the works of the CMIP climate forcing task team, although the targeted exploitation of MACv2-SP in RFMIP was with one publication small compared to the usage of other experiments of RFMIP and AerChemMIP so far.

RFMIP and AerChemMIP received output from 103 experiments leading to 204 publications to date. We separate the RFMIP and AerChemMIP experiments here into three classes, namely experiments with full coupling between the atmosphere and ocean (*hist-X)*, with prescribed sea-surface temperatures and sea-ice at pre-industrial level (*piClim-X)*, and with prescribed transient changes in sea-surface temperatures and sea-ice from a *historical* experiment (*histSST-X)*. Inter-comparing these classes, *piClim-X* experiments were performed the most with a total of 50 contributing models followed by *hist-X* with 36 experiments. However, hist-X is used three times more often in scientific publications (146) compared to *piClim-X (*52*)*. The higher computational demand of *hist-X,* therefore, seems justified by the much larger scientific output compared to the experiments without a coupled ocean (*histSST-X* and *piClim-X*), measured by the number of published articles.”

- Which protocols and experiments were not popular or successful : easy-aerosol? Double calls for IRF calculations? Long perturbed historical simulations? Why?

There were less contributions to ERF-IRF and ERF-SpAer than to ERF-ERF, presumably due to the necessity to implement diagnostics and replacing the host model’s anthropogenic aerosol parameterisation with MACv2-SP, respectively. It meant time commitment of personnel at the modelling centres to carry out these works at a time when many experiment requests from diverse MIPs were pending. Double and triple calls were available for a useful number of experiments for inter-comparing effective radiative forcing of anthropogenic aerosols, e.g., following the method by Ghan (2013) to identify direct and cloud-mediated effects or separating the instantaneous radiative forcing

from the net contribution of adjustments. Long perturbed historical simulations were computationally demanding and still done by several models, at least in parts since they were requested by several MIPs (e.g., RFMIP and DAMIP). Easy-aerosol (https://www.wcrp-climate.org/gc-clouds-circulation-activities/gc4-clouds-initiatives/368-gc-clouds-inititative3-easy-aerosol) was not part of the MIPs discussed here.

We add a paragraph in Section 3.1: “Simpler experiments are certainly always easier to perform and have the advantage that no expert at a modeling center is needed to enable the experiment and output, e.g., for implementing requested diagnostic output that is not yet available in the standard variable list of models, e.g., for RFMIP-IRF. Another example is an experiment design that needs to implement a different parameterization, e.g., for RFMIP-SpAer, which requires manpower at the modeling center for carrying out the work including coding, testing, and performing the experiments. In this case, it takes longer to finish the experiments and the associated scientific exploitation, e.g., in the case of RFMIP-SpAer several years after the work began (Fiedler et al., 2023), which is long compared to easy experiments that modelers can quickly set up via a simple change in the run script, e.g., for RFMIP-ERF. A rule of thumb for experimental design in MIPs could be choosing a setup as complex as necessary, but as simple as possible.” We further add in Section 2.2: “(...) and to separate direct and cloud-mediated effects following the method by Ghan (2013) in RFMIP experiments (e.g., Fiedler et al., 2023).”

Ghan, S. J.: Technical Note: Estimating aerosol effects on cloud radiative forcing, Atmos. Chem. Phys., 13, 9971–9974, https://doi.org/10.5194/acp-13-9971-2013, 2013.

- Why did certain protocols appear easier or harder to follow? E.g., PDRMIP was partially driven in emission-driven and partially in concentration-driven mode, ...

It was probably a question of resources when the decision was made for participating in certain protocols in addition to the model capabilities to perform the requested simulation. We add in Section 3.1: “MIPs already have a specific class of models in mind. For AerChemMIP, emission-driven models were targeted, whereas RFMIP was also including contributions from models with less complex representations of aerosols, e.g., those using prescribed aerosol optical properties. Hence, RFMIP had more participation than for instance AerChemMIP. RFMIP and AerChemMIP were endorsed by CMIP6 and hence had a different structural organization with formal experiment protocols. PDRMIP started earlier and was in comparison more self-organized and dynamic in the MIP life cycle. Hence, PDRMIP comprises an ensemble of models of different complexity. Specifically, some of the models in PDRMIP had the capability to perform experiments with prescribed emissions whereas others needed concentrations resulting in an ensemble of experiments partially driven by emissions and partially driven by concentrations of climate forcers. ”

- Did the use of Tiers help?

We add in Section 3.1: “ Setting priorities with tiers was useful to the extent that it highlighted the priority of experiments from the MIP’s perspective. In so doing, the tiers guided the participating modelers to set a focus on some experiments to have a larger model ensemble where the MIPs wanted contributions the most. However, in retrospect, some of the Tier 2 experiments may have been more useful than Tier 1. An example here is *piClim-histaer* (Tier 2) from RFMIP, which quantified the spread in magnitude and timing of historical aerosol forcing in CMIP6 models, was informationally rich, and a contributing factor in deriving the aerosol ERF time series for AR6 WG1.”

- Maybe some vision on whether MIPs and their experimental demand should maybe remain limited. Should these MIPs merge into one MIP? Are there benefits in keeping (small) separate MIPs?

Very good question that is also on our mind for the next CACTI initiatives, specifically for RFMIP2 and AerChemMIP2. The advantage of keeping the MIPs as two separate endeavours is that we can leverage on their familiarity in the community due to their endorsement during CMIP6. Scientists already have certain research themes, class of

models, and experiment setups in mind when they see the MIP's names. More general reasons against merging (small) MIPs into larger ones are the risk of reduced clarity on the science questions because of more diverse foci, long and potentially less clear motivations for individual experiments in lengthy protocols, the loss of focus on a specific model class, as well as more management and coordination works for large MIPs, which is usually not directly funded. CMIP could for instance be seen as a great large coordinated MIP, which is well structured and financially supported. We think combining smaller MIPs that are more bottom-up driven into larger MIPs would be more difficult to do due to financial and management constraints.

We add in the revised conclusion: "The planning of the second phases of RFMIP and AerChemMIP are currently in preparation as community MIPs for CMIP7 (https://wcrp-cmip.org/model-intercomparison-projects-mips/). The advantage of keeping the MIPs as two separate and comparably smaller endeavours is that we can leverage on their familiarity in the community due to their endorsement during CMIP6, clarity in the science questions because of specific foci of the experiment request from a certain class of models, as well as acceptable workloads for the management and coordination. "

- Is the multi-model approach put into question? Should models be selected (even more) on their key performance before they can go into an assessment? Does the model spread in the results represent our current uncertainty in understanding, or is it partially caused by lacking model selection?

We do not question the fundamental approach of MIPs. We add in Section 3.2:

"There is value in multi-model inter-comparisons to shed light on where the physical understanding is still limited based on the current representation of processes and where we have accomplished a satisfying advancement in our scientific understanding from such model experiments. An open and not restricted inclusion of models by key performance indicators allows broad participation of suitable ESMs in MIPs. Scientists can for their specific assessment decide which model experiments they include since not all experiments are equally suited for all questions, e.g., some models might miss processes and interactions that might be crucial to address the research question. Results of MIPs alone can not fully characterize the uncertainty, if it is at all possible since scientific knowledge might unfold in ways that can not be foreseen at present. This is what we call the process understanding abyss (Figure 2), which limits our ability in advancing the field with our available models. Other evidence should be considered in parallel or ideally in synergy with MIPs to gain new knowledge - may it be observational data from different sources or completely different models that are not suitable for participation in MIPs."

- In consecutive CMIP rounds, one sees the number of ESM models increase. Is this an efficient way to progress science, or would reducing the number of different models, and trying to build one exceptionally high resolution and very competent model, be a better way? Would such a uniformity block/hamper scientific creativity? (CERN of climate science)

There should always be freedom of science. Promising ideas, i.e., those that have proven that there might be potentially great progress if they were pursued further, should find sufficient support in the interest of all without the expense of suppressing ideas of others. If we accomplish to retain this status quo for science, which might require to add more resources rather than redistributing existing ones, we would see no risk for hampering scientific creativity for an improved ability for advancing our understanding of and for tackling climate change.

- How is the activity rate over the lifespan of a MIP: when do most model results come in? How long does analysis go on?

We add in Section : "Some key articles based on the experiments were written and submitted close to the IPCC WG1 AR6 deadline. Submission of model output and analyses continued thereafter and are partly still ongoing at the time of writing. We expect this development to continue for several years, although with a decline in new CMIP6 model output, until a quorum of CMIP7 model output come online. Looking at the history of the use of CMIP data, we would expect that also output of RFMIP and AerChemMIP will be re-used later for documenting progress across their phases, e.g., for the effective radiative forcing, which is also often done for tracking progress across CMIP phases."

- Is the use of more observations to constrain the models the way forward? This is slightly mentioned when discussing "sophisticated methods". However, this point is maybe worth more focus and ideas.

It is one important aspect. We add: “Constraining ESMs with observations are a key in advancing our understanding. Although many observations and reanalysis data are already well used, more could be done in the future. Specifically, instead of comparing to single observational or reanalysis datasets, using several observational data sources would allow to quantify the observational uncertainty against which model results can be better evaluated, e.g., a good performance might mean that model results fall within the observational uncertainty. Moreover, new combined observational products could help to evaluate model output, which may include translating observables into modelled variables. In the past, some approaches have been taken to translate simulated data into the satellite observable space. In the future, machine learning seems promising to develop new and easier ways for exploiting and combining observational data suitable for comparison to model output, e.g., methods for filling observational gaps in some satellite products due to the presence of clouds have been used. Such ideas could be explored more to unfold the new potential to evaluate and constrain model results in the future in ways we have not done in the past.”

- In the manuscript nothing is said about emerging constraints. Is that a way forward?

Emergent constraints help and might be most fruitful in combination with novel observational products (see previous comment). We add: “ Future work could also expand on the use of emergent constraints for responses including feedback mechanisms. In the past, an emergent constraint approach was for instance used to address the present-day forcing of halocarbons leading to a reduced spread in the forcing estimate (Morgenstern et al., 2020). Another example is adopting the approach to constrain anthropogenic aerosol forcing from ESMs (McCoy et al., 2020).”

McCoy, I. L., McCoy, D. T., Wood, R., Regayre, L., Watson-Parris, D., Grosvenor, D. P., ... & Gordon, H. (2020). The hemispheric contrast in cloud microphysical properties constrains aerosol forcing. *Proceedings of the National Academy of Sciences*, *117*(32), 18998-19006.

Morgenstern, O., O'Connor, F. M., Johnson, B. T., Zeng, G., Mulcahy, J. P., Williams, J., et al. (2020). Reappraisal of the climate impacts of ozone-depleting substances. *Geophysical Research Letters*, 47, e2020GL088295. https://doi.org/10.1029/2020GL088295

(7)
Also section 4.2 might benefit from tables containing all the suggested new diagnostics (e.g., IRF, ...) and experiments (e.g., fixed land surface temperature experiments, additional experiments for the impact community, …).

Good idea, we include a new summary Table 5 for the diagnostics and experiments in the revised manuscript:

Method	**Usage**
Improved diagnostic for PM	Air quality assessments and impact studies for health sector
Improved diagnostic for O3	Air quality assessments and impact studies for health sector
Diagnostics for hourly 100 m winds	Wind power studies with associated impact studies for energy sector
Diagnostics for hourly direct and diffuse irradiance	Solar power studies with associated impact studies for energy sector
Hourly output of surface shear stress and near-surface soil moisture	Dust emission studies with impact studies for health and energy sector
Diagnostic from multiple calls to the	Calculation of IRF and better understanding of process contributions to

radiation transfer scheme	model diversity in ERF
Experiments with fixed sea ice, land and sea surface temperatures	Calculation of ERF free of artefacts from land-temperature adjustments for more precise model intercomparisons on radiative forcing
Experiments for short-term climate change mitigation	Information for stakeholders on climate penalties and benefits from air pollution emission changes

Table 5: Proposed new and improved diagnostics and experiments for the three MIPs.

Included in the manuscript in Section 4: “Opportunities arising from novel capabilities and diagnostics are listed in Table 5 and elaborated on in the following sections.”

(8)
Would this be a good paper to introduce the 3 MIPs to a reader? Possibly not, there is very little description of the experiments suggested in the 3 different. Maybe explain and characterize the MIPs better, possibly in a table/matrix.

We include a concise text that explains the rationale of the MIPs and highlight their similarities and differences to characterize them in the new Section 2.1. The cited MIP protocols completely describe the details of the MIPs along with the experiment lists.

[Specific comments]

Below, you can find specific comments on the text of the manuscript. I have tried to indicate as well as possible the line numbers.

Thank you.

ABSTRACT
- line 3 : "in climate model experiments" : this gives the impression that it refers to the "setup" of experiments, whereas I assume it should refer to the difference in results between models.
Change to: “uncertainty in climate model results”

- line 4 : "this article" : I would not use the word article in the text (or abstract)
Rephrased throughout.

- line 9 : "of varying complexity" : is it mentioned because it played an important role in advancing science? I think the varying complexity was not an item/issue in itself. However, the MIPs were designed in such a way that even contributions from less complex models could contribute to certain parts of the analysis. E.g., models containing interactive aerosol but no interactive ozone could still contribute to the analysis of aerosol forcing, but not to the analysis of stratospheric ozone forcing. In addition, that aspect is not so much related to "the partial overlap between the MIPs". But it is true, estimating the feedback from natural emissions (which was an AerChemMIP activity), requires models to contain those interactive emissions.

Agreed, varying complexity of models is not an issue in itself, although it can be perceived as a limiting factor for participating in some experiments like stated. We think the varying complexity plays a useful role in advancing science. We intend to include the following at the beginning of Section 3.2 for a more balanced perspective:

„Although varying model complexity can be a difficulty in understanding differences between model results in a MIP, varying complexity helps in advancing our understanding of climate change. Model simulations with different complexity for instance help in quantifying contributions from feedback mechanisms to climate responses. Moreover, additional model components and representations of processes have been incorporated in Earth system models over time in addition to improvements of previously existing physical parameterization schemes and boundary data. Such model developments allowed new insights into the role of processes including feedback mechanisms for climate change, although the overall progress is possibly not as fast as one would hope for all aspects. Clouds and circulation are for instance outstanding challenges that have not been

resolved through the development of CMIP-class models to date."

- line 9-11 : "It specifically ... for estimating effective radiative forcing." This seems rather limited if this is the main synergy coming from the 3 MIPs.
We remove the example of estimating effective radiative forcing in the abstract.

- line 12 : "... that have specific biases ..." : I don't have the impression that this gets much attention in the text later. Therefore I don't know whether it should be mentioned in the abstract.
Change to: “We discuss the challenges of gaining insights from highly complex models that face computational and process representation limits and provide guidance from our lessons learned.”

- line 13 : are global kilometer-scale experiments in view in the next 5 to 10 years in the context of the relation between composition and climate change? In the last decade, global climate models only doubled their resolution (e.g., from 2x2 degrees horizontal resolution to 1x1 degrees). Do the authors expect to arrive at a kilometer-scale resolution in 5 to 10 years on a MIP-wide scale?

We now provide details of our thoughts on kilometer-scale experiments in Section 4.1.2 (see reply aloft).

- line 16 : "can" be evaluated -> "should" be evaluated
Change as suggested.

- line 16-18 : although I think this is important, "sophisticated methods" is a bit vague. It is also used in a few other places in the manuscript.
Change to “observational constraints” in the revised manuscript, e.g., here “Future experiments should be evaluated and improved with observational constraints that leverage multiple datasets, and thereby, help to advance the understanding of climate change and its impacts.”

1. INTRODUCTION
- line 25, 26 : why "concentrations" for GHGs but "burdens" for aerosols?
Revise to: “Radiative forcing may be caused by changes in atmospheric composition, including for instance aerosols and their precursors, greenhouse gases such as carbon dioxide and methane, as well as changes in surface albedo or irradiance.”

- line 29 : "... direct impact ... instantaneous radiative forcing ..." : to a novice in the field, this should possibly be slightly better explained. Now it seems more like defining one expression by another expression.
We add: “For example, changes in atmospheric composition and land-use perturb the Earth’s radiation balance, as quantified by the radiative forcing. (...) Radiative forcing is measured in units of power density (W m-2). (...) Instantaneous radiative forcing (IRF) is the change in radiation fluxes that arise from a climate forcer, e.g., a perturbation in the atmospheric composition.”

- line 33, 34 : "... can take several hundreds years depending on the magnitude" : as long as one stays within a linear regime, perturbations (whether small or large) disappear with the same timescale. If there is interannual variability, however, the smaller perturbation will sooner disappear behind the detection limit. Maybe one can be more precise.
Change to: “(...) can take several hundred years because of the slow response of ocean temperature. Smaller forcing and responses are more quickly masked by the internal-variability than larger perturbations.”

- line 38 : "due to changes in emissions of reactive trace gases" : I assume one refers, e.g., to DMS and biogenic VOC emissions which are precursors of (radiatively active) O3. It seems to exclude "emissions of species which have a direct impact on radiation", e.g., dust - however, I think they fall into the same category.
We add: “Moreover, changes in wind-dependent emissions of aerosols that occur due to circulation adjustments can be interpreted as chemical adjustments, although changes in aerosol emissions can occur with surface-temperature responses and would fall into the category of chemical feedback in that case.”

- line 40-41 : The second part of the sentence contains twice "response" and twice "changes". Cannot this be said in an easier way?
Thanks, change to: “Effective Radiative Forcing (ERF), measured at the atmosphere's boundaries, encompasses both the IRF and the contributions from rapid adjustments, that refer to flux-modulating changes in the system driven by IRF in the absence of surface temperature changes.”

- line 41 : "steps" : is "steps" the appropriate word in the context of this paradigm?
Steps might suggest a sequential behaviour which is not necessarily the case, e.g., for responses and feedbacks. We change “steps” to “segments” throughout the manuscript.

- line 43-44 : "Understanding and quantification ... derived" : "derived" does not go well together with "quantification".
True, change to ”assessed”.

- line 43-44 : "typically derived" : some parts in the paradigm can possibly be derived by other models than ESMs. I think line 51 is on the contrary correct : for climate response and feedbacks one needs the ESMs.
Add: “(...), although other methods for some of the segments exist, e.g., radiation transfer models to compute IRF.”

- line 44 : Heavens et al. (2013) : I have the impression that the text of Heavens et al. (2013) is more about running and verifying the ESMs, and not so much about disentangling the perturbation-response paradigm.
Yes, it may be misleading to include the reference in the long sentence here and we remove it.

- line 46-47 : "... simulate [aerosol and their precursor] emissions, [?] transport, and deposition [of aerosols]" : when one reads this sentence, one gets the feeling that something is missing on the location of the [?]. One could maybe change it into : "... simulate aerosol and their precursor emissions, and transport and deposition of aerosols".
Changed as suggested.

- line 44, 50 : What is meant by "design"? I would think that "process complexity" (line 49) is part of it.
Design was meant to cover different parameterization schemes, dynamical cores, spatial girds, numerical integration, tuning, boundary data etc.. Change to: “Modern ESMs vary in their design, e.g., concerning different parameterization schemes, dynamical cores, spatial grids, numerical integration, tuning, and boundary data. They also vary in their level of complexity for representing physical, chemical, and biological processes, and how represented processes interact. (...) The simulated aerosols may interact with the radiation transfer, formation of cloud droplets and ice, or just a part of it.” And change in line 49: “(...) for instance be due to differences in process complexity and interactions within the respective ESMs.”

- line 47 : "collaborates regularly" : "collaborates" gives the impression of a continuous process, whereas "regularly" gives the impression of a process with several breaks.
Change to: “ regularly produces”

- line 47 : "multi-model ensembles of a common set of experiments" : this is clear and referts to multi-model. However, on line 50 in "ensembles of ESM experiments", "ensembles" probably refers to a group of different experimental setups. Maybe this should be formulated more clearly to avoid confusion.
Change to: “Ensembles of different ESM experimental setups following the same protocol (...)”

- line 49 : "This diversity in response may be due to differences in process complexity within the respective ESMs, and/or may be due to the design and coupling of different model components" : aren't there more reasons for model diversity? E.g., different parameterisations (without difference in complexity), different parameter values, different but equally complex dynamical cores, ...
We clarify that these things are covered by “design” (see above) and add here “for instance” since also other things may explain model diversity.

- line 57 : "The principle idea of MIPs" : maybe change into "The principle idea of a MIP"

Changed as suggested.

- line 57 : "The principle idea" : what is the principle idea? Do MIPs exist in other sciences?
Different MIPs exist in climate science, e.g., here RFMIP and AerChemMIP. They share the same basic idea. We change “principle” to “basic”.

- line 64 : "both the MIPs" -> "both MIPs"
Changed as suggested.

- line 67 : "of the three MIPs" -> "of three MIPs"
Removed.

- line 67 : the summing up after ":" is later followed by a continuation of the sentence with ",". This is a bit strange.
Change to: “(...) connecting the scientific communities of AerChemMIP, RFMIP, and PDRMIP under one umbrella named TriMIP”

- line 67 : what do the authors mean by "diagnostic tools"? I think one could be more specific.
We mean for instance double calls to the radiation transfer calculation. Change to “diagnostic request” for clarity at the start of the paragraph and in line 67 removed for brevity.

- line 66-68 : three aims are mentioned in this sentence as if they constitute a complete set. However, a few sentences later some extra aims appear.
We move it up to align it better with the stated aims: “In so doing, we discuss the challenges of understanding multi-model climate responses and identify potential opportunities to make further advances in the research areas of these MIPs.”

- line 73 : "... in this area" : this is a bit vague. Does it refer to "understanding multi-model climate responses"?
Changed to: “(...) in the research areas of these MIPs.”

2. ADVANCEMENT THROUGH MIP'S CROSS-LINKAGES
- line 75-76 : "considering structural differences between ESMs" : how should "structural" be interpreted? Isn't it more often on the level of parameterizations that differences arise? Having a process in or not, should that be seen as a "structural" difference? For me, "structural" refers to the broad technical design choices of an ESM (order of processes when numerically solving, how do components technically interact, is a coupler like OASIS used (or something else), is the ESM one model or an assembly of models, ...), whereas I do not think that those differences contribute most in the end.
Change to: “(...) considering structural differences concerning the design and the level of complexity between ESMs.”

- line 77 : "components" in the paradigm. I would rather say that ESMs have components. Possibly use a different word for what constitutes the paradigm.
Change to “segment” throughout.

- line 77-78 : "RFMIP focuses on an improved understanding of the (role of) radiative forcing diversity (for the climate response)" : leaving out a few words, I thought it better described the aims of RFMIP.
Changed as suggested.

- line 79 : "... on precipitation response to idealized atmospheric composition ..." : I would maybe skip "idealized". I think the focus was on "precipitation response to atmospheric composition change", and the tool was indeed "using idealized atmospheric composition changes".
Remove “idealized”.

- line 80 : "earlier" : this gives the impression that there is a time dimension in the paradigm approach.
Yes, change to: “addresses all segments”

- line 80-83 : "... making these models more complex ..." : this gives the impression that AerChemMIP uses a specific class of models, which is confusing. E.g., I think that in RFMIP a large portion of the models and in PDRMIP at least half of the models also start from emissions.
We add: “(...) than is necessary for participation in the other two MIPs (...)”

- line 83 : "inspired by each other" : PDRMIP was set up before AerChemMIP and RFMIP, so it was maybe unidirectional.
Yes, change to: “PDRMIP began earlier and to some degree inspired the experimental protocols of AerChemMIP and RFMIP.”

- line 84 : "with a certain class of model in mind" : were they aiming for very different types of models? It is true that the natural feedback quantification of AerChemMIP needs interactive natural emissions, but in general the models were possibly reasonably similar?
Added: “(...), i.e., CMIP-class models in all three MIPs and specifically AerChemMIP required more interactive processes than the other two MIPs.”

- line 84-85 : "ensembles of ESM experiments of different complexity ...." : this is a bit confusing. Did the protocols differ in their demand for complexity, resolution, ..., or did the results just finally appear to be like that?
AerChemMIP required more interactive processes, thus more complexity, but a specific resolution of the model experiments was not requested in any of the MIP protocols. We elaborate more on these free choices at the modeling centers at the end of Section 3.1. Change to: “ Taken together, there are ensembles of ESM experiments of different complexity, model resolution, number, and length in the three MIPs.”

- line 86 : "A major advancement from the synergy between the three MIPs was the widespread adoption ..." : Is meant adoption outside the three MIPs mentioned here, or only within the framework of these 3 MIPs?
Add: “within and outside of the three MIPs ”

- line 88 : "consistent calculation" -> "consistent diagnosis"
Changed as suggested.

- line 97-98 : "experiments where the atmospheric composition represents the values in 1850." : I think this formulation is not very nice.
Change to: “experiments with an atmospheric composition as of 1850”

- line 99 : "diagnostic calls" -> "additional diagnostic calls".
Changed as suggested.

- line 100 : "in CMIP6 models" : this sounds a bit sloppy, so maybe it is better to write : in the ESMs used in CMIP6.
Changed as suggested.

- line 107 : concerning the practice of estimating ERF : add "already mentioned"
Added.

- line 113-114 : "more relevant" : If one uses "more", I would think one needs to mention what it is compared with.
Change to: “tailored to the needs”

3. CHALLENGES IN THE MIPs RESEARCH
- line 118 : "components along the perturbation-response paradigm" : what is meant by "components"? On line 41 and 43, the word "steps" was used in the perturbation-response paradigm ... ("component" is also used on line 122). What is further meant by "model differences"?
We replace “components” and “steps” with “segments” throughout the text. Change to: “A major challenge to further advancing the understanding of climate change with ESMs is that differences in their results for individual segments of the perturbation-response paradigm are not independent from other segments.”

- line 119-121 : it is not clear what difference one wants to stress here : (1) difference in forcing for the same composition change, (2) different climate response for same forcing, or (3) different feedbacks? Possibly write it clearer.
Change to: “Examples are an inter-model spread in forcing for the same change in atmospheric composition, and model-dependent climate responses to the same forcing involving different types and magnitudes of feedbacks.”

- line 123 : What is meant by "joint strength" : common approach? (which then facilitates comparison)
Replace with “common approach”.

- line 127 : Is here carefully the word "provided" used (leaving it open whether the data were actually used), as 50% of the models still used their own aerosol emission in PDRMIP?
Rephrase to: “PDRMIP asked for prescribing the same aerosol information in models (...)”

- line 131 : This is a bit confusing as the other MIPs also did atmosphere-only experiments. Possibly "atmosphere-only" is not the correct terminology, as probably the landmodel is also active in these simulations.
Change to: “(...) and removed feedbacks by performing experiments with prescribed sea-surface conditions. (...) To that end, RFMIP requested experiments with prescribed sea-surface conditions similar to AerChemMIP to obtain precise model estimates of ERF. ”

- line 131 and 133 : "atmosphere-only" versus "prescribed sea-surface conditions" : I would use one terminology.
We use “prescribed sea-surface conditions” throughout the revised text.

- line 135 : "in a complementary manner" : I am certainly aware that there has been a lot of synergy between the MIPs. However, this gives the impression that together, the MIPs covered almost (all) the topical research questions.
Add: “(...) for addressing their specific research questions.”

3.1 COMPUTATIONAL CAPACITY ABYSS
- In general I think that Section 3.1 is not so well written, and should be improved.
We substantially revise the section. Please see below for our specific replies.

- The three axis approach is nice and illustrative, but I have some concerns: (i) I would think that the triangle area is not a correct representation of the computational needs. I would rather think that, given a specific configuration, the product of the distances along the 3 axis is representative for the computational need (geometrically that would correspond to the volume of a block). Possibly, the volume behind the triangle (tetrahedron formed by the triangle and part of the 3 axes) could also be seen as proportional with the computational needs (and could be a better quantification of the computational needs than the area of the triangle). (ii) If one wants the volume to be representative for the computational needs, I would choose the axis linear. E.g., one places a 1x1 resolution simulation 4 times further from the origin than a 2x2 simulation. (iii) Setups with the same computational cost, will not lie on flat surfaces but rather on hyperbolic type of surfaces (I would think).
Thanks for the suggestion. Using a tetrahedron (three-sided pyramid) is a good idea for making the figure quantitatively more meaningfu, although the triangle is indicative of the growing computational need as we move out along the axesl. We change in the text: “The volume of the tetrahedron between the origin and the marked triangle indicates the computational need for the experiments. The computational need scales non-linearly. (...) To account for the non-linearity in the computational need, the volume of the tetrahedron would be calculated on scaled values, i.e., an experiment with twice as fine resolution would be marked four times further away from the origin on the resolution axis. ” And change in the caption of the figure: “The triangles mark potential choices along these axes with the volume of the tetrahedron filling the space between the origin and the colored triangle indicating the computational need.”

- line 156-157 : "For some research questions, the complexity of ESMs can be reduced to a large degree" : please give an example.
Add: “For instance, concentrations of well-mixed greenhouse gases can be prescribed instead of being simulated from emissions, if one is interested in computing the forcing and response to a given change in the atmospheric composition. ”

- line 137 : "Available computational capacity ..." -> "Limited (available) computational capacity ..."
Change to: “Limited (available) computational capacity (...)”

- line 137 : "modeling center" : maybe specify what this is for people not in the field.
Add: “Modeling centers perform the requested experiments with the ESM which they support. They contribute to the decision for which community-driven MIPs experiments of the ESM will be conducted.”

- line 138 : "on short timescales" -> "in a short period of time"
Changed as suggested.

- line 138 : "choice" and "defined" is a strange combination. It is also used on line 139.
Change to: “Not all experimental settings are explicitly defined by the MIP's experiment protocols, giving modelers room to make their own choices.”

- line 138 : where is freedom left in the experimental setup? Isn't it that rather some models do not have interactions? E.g., in some models in an abrupt-4xCO2 experiment, vegetated area can reduce/increase which can have an impact on dust emissions (in addition to the impact of dryer/wetter/windier conditions). In other models, the vegetated area is not allowed to change.
It is the process complexity like you say, the resolution of experiments, and the number of simulations in an ensemble, e.g., some modeling centers produce more simulations to better sample internal variability than others. We add: “Such choices may be regarding the capability of the model to specify a certain resolution ”

- line 139 : "exact experimental design" : what is meant by exact?
Change “exact” to “final”.

- line 139 : "Taken together, there were inevitable tradeoffs in the exact experimental design." -> "... there are ..."
Changed as suggested.

- line 142 : "in an ensemble of experiments" : unclear whether "experiments" refers to identical setup or not.
Change to: “(...) in an ensemble of different experimental setups per ESM.”

- line 143 : "area of triangle" -> "volume of tetrahedron"
Changed as suggested.

- line 145 : does not scale linearly : I would still put the real cost on a linear scale; e.g., 2x2 degree horizontal resolution corresponds with 1, 1x1 degree horizontal resolution corresponds with 4, such that the volume calculation is still correct.
Yes, we explain it in the text with: “To account for the non-linearity in the computational need, the volume of the tetrahedron would be calculated on scaled values, i.e., an experiment with twice as fine resolution would be marked four times further away from the origin on the resolution axis.”

- line 146 : "doubling the simulation length or number" -> doubling the simulation length or number of simulations
Changed as suggested.

- line 147 : "but this is not true for the model resolution" : I would think that it can be made true (see above)
Yes, added as above.

- line 148 : "for instances" -> "for instance", or "e.g."
Change to "for instance".

- line 149 : "has become available" -> "continues to grow"
Change to: "(...) computing power continues to grow, (...)".

- line 150 : I agree with "interactive chemistry" but I find "competition for priority of experiments" strange. I would say that "chemistry" competes with "resolution" or with the "number of simulations", but not with the "priority or experiments".
Change to: “This is for instance the case in light of the computational cost of interactive chemistry against the resolution and the number of simulations. Additionally, all model experiments, irrespective of whether the models have interactive chemistry, compete for the priority at modeling centers due to limited computing resources.”

- line 151, line 154 : I would not say "the most complex ESMs" -> "complex ESMs"
Changed as suggested.

- line 157-159 : I assume it is true, but it is so general that it would be nice to have an example.
We add: “For instance, an ESM could simulate changes in vegetation cover due to increased greenhouse gases that in turn have an impact on dust-aerosol emissions in addition to potential changes in soil moisture and winds. In less complex models, the vegetation cover is for instance prescribed such that the number of interactive physical processes is smaller. ”
- line 159-161 : such that "model-internal variability" can be separated from the "mean radiative forcing", "climate response" and "impacts on air-quality" : although I think I understand what is meant, I think it lacks some better description. I think one wants to split, e.g., the TOA imbalance in a "mean radiative forcing" and a contribution from "internal variability", and the same for the "climate response" and "impacts on air-quality".
Change to: “It makes creating large ensembles of ESM experiments possible that are needed to split for instance the imbalance in the radiation budget at the top of the atmosphere into a mean radiative forcing and contributions from internal variability. Similarly, a separation of the response in temperature or air quality into a forced signal and a contribution from internal variability is neccessary. The required ensemble size for sufficiently reducing the influence of model-internal variability on the global mean radiative forcing (e.g., Forster et al., 2016, Fiedler et al., 2017), climate responses (e.g., Maher et al., 2019, Deser et al., 2020), and impacts on air quality (e.g., Garcia-Menendez et al., 2017, Fiore et al., 2022) depends on the magnitude of the forced signal against the magnitude of the internal variability.”

- line 158 versus line 162 : what is written here seems to contradict itself : "high complexity can be reduced" (line 158) <-> "high process complexity, ... and needed" (line 162)
Add: “needed for specific research questions”

- line 162 : "but also poses" -> "also poses"
Changed as suggested.

- line 162 : what is meant by "needed"?
It depends on the research question what is needed. Now: “needed for specific research questions”.

- line 162-163 : in the first sentence, one mentions apparently one "challenge". In the next sentence, two challenges appear.
Plural now in first sentence: “challenges”.

- line 164 : "the number of interacting processes" : I don't know whether one can express this in a number for an ESM.
Change to: “intricacy and fidelity of represented processes”.

- line 165 : "specification of the model resolution" : does this refer to the model center's choice of resolution, or the MIP-imposed resolution?
Add: “(...) by the modeling centers.”

- line 167 : "prescribeid aerosols such as the spatial distribution" : this should be better formulated
Change to: “(...) while other models prescribe spatial distributions of aerosol optical properties”

- line 168 and 169 : this sentence refers to both types of differences (model capabilities and experimental setup).

However, my impression is that the former sentences only refer to model capabilities (not about experimental setup).
Remove here: “and experimental setup”

- line 170 : what is "model diversity" in the design of a MIP?
Revise to: “The second challenge comes from the consideration of model diversity in the level of complexity already in the process of designing a MIP protocol, since for instance a few models can simulate processes that most others cannot.”

- line 170-171 : some models can simulate processes that others cannot : isn't that the same as "diversity in the level of complexity"?
Yes, now explicitly stated, see previous comment.

3.2 PROCESS UNDERSTANDING ABYSS
- line 176 : I would not use "most complex", but just "complex".
Changed as suggested.

- line 177 : I don't think that advancing climate science is limited because not "all" processes are represented. One can never know or represent them all, but one can make progress by adding the ones which we think are relevant. What the authors possibly want to say : we cannot reproduce observed climate change and simulate future climate change because we miss or did not represent some processes. (This corresponds probably more with what is said in the second sentence of this section, line 178.)
Change to: “There are some limits to advancing climate science with today's complex ESMs since we miss or did not represent some processes that are thought to be relevant to reproduce observed and project future climate change. ”

- line 179 : "are not represented represented differently" -> "are not represented or represented differently"
Changed as suggested.

- line 183 : "primary organic aerosols" : does this refer to (interactive) marine primary organic aerosols?
Yes, we add: “marine”.

- line 183 : "can be represented" -> "are represented"
Changed as suggested.

- line 186 : "... with potential health impacts." : this makes one think that no other impacts will be mentioned, but in the next sentence also the impact on clouds is mentioned.
We add: “Moreover, (...)” for a better coherence of the text.

- line 199 : ".. model consensus and smaller in magnitude might suggest that they are irrelevant ..." : maybe formulate in a different way.
Change to: “And of those feedbacks that are simulated, model consensus or small magnitudes for a feedback might lead to a misleading conclusion that these feedbacks are not important. ”

- line 200 : "dust trends" : possibly add "over the historical period"
Added as suggested.

- line 201 : "so much that they are of opposite sign" : a change in sign from a small negative to a small positive value is not automatically a dramatic change.
Change to: “(...) but the dust trends differ in sign across ESMs.”

- line 201 : "so much so that" : maybe formulate differently
We remove "so much so that" (see previous comment).

- line 201 : I would think that a small dust feedback does not imply automatically that the dust emission changes are small.

True, especially on regional scales. Change to: “The CMIP6 models show trends of different signs and magnitudes for desert-dust aerosols over the historical time period (Bauer et al., 2020, Thornhill et al., 2021), and there is no ESM that reproduces the magnitude of the reconstructed dust increase from the pre-industrial to the present-day (...)” and add “(...), which has implications for the understanding and quantification of the radiation imbalance”.

- line 203-223 : This section seems to focus on natural processes. It is however not sure, as it is mixed with information which is maybe not only related to natural aerosol. E.g., mentioning that trends in aerosol and ozone do not fit the observations can also be related to errors in anthropogenic emission estimates; the effect that optical properties or size distribution of aerosol is biased can possibly also apply to anthropogenic aerosol. In general, I find this paragraph a bit difficult to follow, and it should be improved.
We split the paragraph into two of which one is addressing natural processes and challenges in constraining historical trends for different aerosol species and their precursors:

“Of those processes that are simulated, a large driver in model diversity for atmospheric composition is thought to stem from the representation of natural processes (e.g., Seferian et al. 2020, Zhao et al., 2022). In particular, a better understanding of natural aerosols in the rapidly warming Arctic may be a key factor in resolving the puzzle of Arctic amplification (Schmale et al., 2021), where diversity across ESMs for short-lived climate forcers is large (Whaley et al., 2022). Another example of the crucial role of representing natural processes is the ability of ESMs to simulate aerosol properties and weather on regional scales. Circulation is a grand challenge for ESMs (Bony et al., 2015), affecting the spatiotemporal distribution of aerosols. Desert-dust aerosols are, for instance, emitted and transported by winds, with a persistently large diversity across ESMs (e.g., Evan et al., 2014, Checa-Garcia et al., 2021, Zhao et al., 2022, Kok et al., 2023). The ability to accurately simulate atmospheric circulation is also relevant to the challenge of realistically simulating clouds and rainfall, including their regional trends due to atmospheric composition changes (e.g., Sperber et al., 2013, Stevens et al., 2013, Fiedler et al., 2020, Wilcox et al., 2020). The simulated clouds influence how aerosols can affect them and rainfall determines when and where aerosols are removed from the atmosphere.

There are a number of challenges in better understanding historical trends in aerosol species and their precursors from different natural and anthropogenic sources. Such knowledge would help to unravel model diversity in the evolution of aerosol forcing over time, and how it is related to time-dependent temperature biases in CMIP6 models (Flynn et al., 2020, Smith et al., 2021a, 2021b, Zhang et al., 2021). ESMs simulate, for instance, different historical trends for O3 and aerosols (Mortier et al., 2020, Griffiths et al., 2021). Even for present-day conditions, outstanding challenges for simulating aerosols persist. AerChemMIP points to model differences in the concentrations of secondary organic aerosols (Turnock et al., 2020). Moreover, aerosol optical properties are partially biased (e.g., Brown et al., 2021), the size distributions of different aerosol species are not sufficiently understood (Mahowald et al., 2014, Croft et al., 2021), and inter-model differences in aerosol optical depth persist across different phases of CMIP and AeroCom (Wilcox et al., 2013, Vogel et al., 2022).”

- line 206 : secondary organic aerosols : mainly natural?
We add: “(...), which have natural and anthropogenic origins (Fan et al., 2022)”

Fan, W., Chen, T., Zhu, Z., Zhang, H., Qiu, Y., & Yin, D. (2022). A review of secondary organic aerosols formation focusing on organosulfates and organic nitrates. *Journal of Hazardous Materials*, *430*, 128406.

- line 214 : this sentence discusses dust again, whereas some aspects of dust had already been mentioned in line 198-202. Maybe this could be combined.
Yes, we move: “The CMIP6 models show trends of different signs and magnitudes for desert-dust aerosols over the historical time period (Bauer et al., 2020, Thornhill et al., 2021), (...) ” to the previous paragraph.

- line 219 : Although it is true, I don't know whether the tuning is relevant in this context and should be mentioned here.
Removed for better text coherence.

4. METHODOLOGICAL OPPORTUNITIES

- line 230 : "and and finally" -> "and finally"
Changed as suggested.

- line 231 : "to understand the causes of model diversity" : I don't think one needs observations to understand the differences. However, observations might constrain the models.
Change to: “to constrain models”.

- line 233 : "the further development of the method for radiative forcing calculations" : maybe reformulate
Change to: “ further improve radiative forcing calculations”.

- line 232-234 : I suggest to improve the sentence
Change to: “Moreover, there is the opportunity to further improve radiative forcing calculations, and diagnostic requests for ESM experiments to allow more synergies with impact assessments.”

4.1 AUGMENTED ESMs
4.1.1 EMULATORS WHERE INFORMATIVE
- line 238 : "are informed" : this is a strange expression, and not the same as on line 243; is the first one similar to "trained" (as "training" on line 244)?
Change to: “are trained on output”

- line 237-242 : "reduces computational demand", "fast calculation", "massively reduced computational cost" : the same concept is repeated several times. Maybe avoid repetition.
We remove the repetitions.

- line 245-248 : (informed by CMIP, idealized experiments, or PPEs) is broader than line 238-239 (informed by CMIP) : maybe it should be consistent.
Added e.g., in line 238: “(...) e.g., from the MIPs and several CMIP phases.”

- line 253-254 : "and explore climate responses to different forcing agents" : seems rather similar to what is mentioned on line 243
Agreed, Change in 253-254: “Also, the difficulty of accounting for non-parametric biases of CMIP models in emulators remains (Jackson et al., 2022). Nevertheless, emulators have already been proven useful to sample parametric differences and to study climate change (e.g., Tabaldi and Knutti, 2018).”

- line 254 : what is meant by "different" forcing agents? Is it the same as on line 243 ("different forcings")?
Remove in line 254 (previous comment) and change to “radiative forcing” in line 243.

4.1.2 KILOMETER-SCALE EXPERIMENTS WHERE POSSIBLE
- line 258 : "which can be enabled" : this gives the impression that this process/evolution is not difficult. Isn't that an underestimation?
Perhaps it needs to be tried more to say how difficult or easy the process is. We change: “can be enabled” to “would require” which is maybe neutral in tone.

- line 264 : "Simpkins (2018)": why not referring to the real paper, i.e., McCoy et al. (2018)?
We add McCoy et al. (2018).

- line 267 : "that involve" -> "that involves"
Changed as suggested.

- line 272 : "periods of a weeks to years" : "a week" or "weeks"
Change to: “time periods of a few weeks to years”

- line 272 : "they are promising to better simulate clouds, precipitation and circulation" : already mentioned on line 259-261.

We remove it here.

4.1.3 MACHINE LEARNING WHERE NEEDED
4.2 IMPROVED DIAGNOSTICS AND ANALYSIS
4.2.1 RADIATIVE FORCING CALCULATIONS
- line 299 : "e.g. Sherwood et al., 2015" -> "e.g.,"
Comma added.

- line 299-307 : Some of this (IRF) has been mentioned earlier in the text (Section 1, line 20-42). Maybe some connection should be made to those earlier mentions.
We improve the definitions of IRF and ERF in the introduction, and remove the definitions here.

- line 315 : "double calls" : possibly explain this
We explain it in Section 2.2 : " (...) additional diagnostic calls to the radiation schemes, also known as double and triple radiation calls, that enabled calculations of the IRF" and add here "(Section 2.2)"

- line 322-324 : these two methods have been mentioned earlier in the text (Section 2, line 90-95). At least mention that in the text.
We add: "As mentioned earlier (...)"

- line 326 : "atmosphere-only" : in such simulations the land model is also active, together with possibly parts of the sea-ice model. So maybe one should use another term to describe this setup.
Yes, we remove the phrase "atmosphere-only" throughout, although it is often used in the community, and now use: "model experiments using prescribed sea-surface temperatures and sea ice".

- line 334-336 : is it realistic to expect the fixed land-surface temperature method to be implemented in many ESMs soon (and thus on a MIP-wide scale)?
It is difficult to say because we have a limited view based on a few modelling centers which are involved here. We change to: "If adopting the fixed sea- and land-surface temperature method (Figure 3) in a MIP becomes feasible, (...)".

- line 337-344 : this last paragraph is a bit different from the rest of the text in this section. It is not a radiative forcing calculation, but appears in a section with the title "Radiative forcing calculation". It probably has its place in this section, but it should be better integrated/introduced.

True, we start with: "The radiative forcing of anthropogenic aerosols depends on the optical properties and the effects on clouds.", and add at the end of the paragraph: "RFMIP experiments point to overestimated aerosol absorption from anthropogenic black carbon and a relatively small share of natural aerosol absorption which leads to direct radiative effects of anthropogenic aerosols in some CMIP6 models which is implausible in light of other lines of evidence (Fiedler et al., 2023). That multi-model assessment was not as broad as it could have been due to the limited availability of requested output for aerosol properties and diagnostic calls to the radiation transfer scheme for aerosol effects in the CMIP6 models. If more such output is available from the next phases of RFMIP and AerChemMIP, we would learn more about the reasons for model differences in the radiative forcing of anthropogenic aerosols. "

- line 337 : "yet another area" : this is not a nice expression
Change to: "Improved diagnostics and observational constraints in the output analysis for aerosol burden and optical properties would be useful for better understanding the model diversity in the associated radiative forcing and the climate response."

- line 338, 342 : "sophisticated ways (of analysis)" : this should be more specific.
Yes, change to: "observational constraints", and "Analysis of relevant and correlated model diagnostics together with observational constraints (...)", respectively.

- line 339-340 : "diversity ... limit" -> limits
Changed as suggested.

4.2.2 SYNERGIES WITH IMPACT ASSESSMENTS
- line 346-363 : although I certainly see the value of this, implementation of these extra diagnostics should not only happen in the RFMIP/PDRMIP/AerChemMIP simulations, but also in other CMIP simulations. There is maybe a task for this community, to promote the importance of these diagnostics to the wider climate change modelling community.
It would certainly be useful to more broadly think about output that better supports impact assessments. There are communities that are better suited to have a complete overview of needs for impact assessments, e.g., VIACS and ISIMIP. We are glad to support these communities with our expertise where we can. To that end, we add ideas for output that would facilitate easier assessments of impacts that are part of solving the problem of future climate change, and are connected to our interests. Specifically, output for assessments of potentials for renewable power generations from solar and wind energy in future climate change experiments could help to accelerate the knowledge transfer to the economic sector for the politically fostered energy and mobility transition to mitigate climate change. We add these along with the previously existing ideas on model output from the text in the new Table 5 (see aloft).

We revise the last paragraph: “Another opportunity to connect more with impact-oriented research can arise from ESM experiments for additional future socio-economic and mitigation-based pathways such that uncertainty in emission developments, including mitigation and associated impacts of atmospheric composition changes, can be systematically explored. In addition to new phases of AerChemMIP and RFMIP, examples are a MIP on future methane removal (Jackson et al., 2021) in support of potential climate solutions or on fire emission developments possibly accounting for the new capability to represent fire feedbacks (Teixeira et al., 2021). If these MIP communities would pursue stronger interactions with communities concerned with climate-change impacts, e.g., with the Vulnerability, Impacts, Adaptation and Climate Services (VIACS, Ruane et al., 2016) and the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP, Frieler et al., 2017) community, they could enhance the usage of output from MIPs for societally relevant problems. Such engagement could lead to a better-integrated understanding of links between climate change, extremes, air quality, and the impacts in different sectors, e.g., health, energy, and economics, for climate change preparedness. ”

- line 353 : "combination of species" -> "combinations of species"
Changed as suggested.

- line 363-364 : "considerations" : What is meant by "considerations"? What is meant by this sentence?
Replaced (see paragraph above).

- line 366 : "the usage of MIPs" : What is meant by this : the use of data from several MIPs? Or use the concept to start extra MIPs?
Change to: “usage of output from MIPs”

5. CONCLUSIONS
- line 373 : some ideas appear which have not been mentioned earlier, e.g., that the paradigm does not work for precipitation.
It works, but it seems less satisfying for precipitation responses due to model disagreement on regional precipitation trends due to composition changes. We add in the conclusion: “(...) , e.g., due to reduced model consensus on regional changes in precipitation compared to temperature for a given forcing.” Difficulties to represent rainfall was listed in Section 3.2. We revise the part for clarity: “The ability to accurately simulate atmospheric circulation is also relevant to the challenge of realistically simulating clouds and rainfall, including their regional trends due to atmospheric composition changes (e.g., Sperber et al., 2013, Stevens et al., 2013, Fiedler et al., 2020, Wilcox et al., 2020). The simulated clouds influence how aerosols can affect them and rainfall determines when and where aerosols are removed from the atmosphere.”

- line 371 : the understanding "with" Earth System Models : maybe the understanding of climate change with Earth System models

Changed as suggested.

- line 374-375 : "In part, this is related to the grand challenge of representing clouds and circulation, which can be addressed with newly evolving capabilities." : it is not clear what "this" refers to.
Change to: “the difficulty of simulating precipitation responses”

- line 394-396 : this last sentence (about GIANT) appears to be quite different from the rest of the paragraph (which was mainly about experimental design). I would suggest trying to integrate this better.
True, removed the sentence on GIANT.

- line 400-405 : I don't know whether an experiment involving 2 models should be mentioned in the conclusions. The conclusions should have a broad general view.
True, also removed.

SHORT SENTENCES :

I found a few short sentences, which broke the nice reading flow. I would try to modify them, and better integrate them in the text.

- line 80 : "AerChemMIP also focuses on quantifying radiative forcing and responses."
Change to: “AerChemMIP also focuses on quantifying radiative forcing and responses, but addresses all segments in the paradigm since all participating models simulate atmospheric composition based on emissions, transport, chemical transformations, and deposition, making these models more complex in their process representation and interactions than is necessary for participation in the other two MIPs (e.g., Thornhill et al., 2021).”

- line 163 : "AerChemMIP emphasized two such challenges."
Change to: “There are two challenges for reducing uncertainty that can be emphasized.”

- line 199 : "Dust is one such example."
Change to: “And of those feedbacks that are simulated, model consensus or small magnitudes for a feedback might lead to a misleading conclusion that these feedbacks are not important. Dust trends over the historical period is one such example.”

- line 281-282 : "Proofs of concept from single ESMs exist."
Change to: “Proofs of the concept of applying machine learning in our research field exist.”

AERCHEMMIP :

I have the impression that AerChemMIP stresses a bit more on its achievements than the other two MIPs. So I would suggest trying to reformulate a few sentences. They are :

- line 163 : "AerChemMIP emphasized two such challenges ..."
Change to: “There are two challenges for reducing uncertainty that can be emphasized.”

- line 181-183 : "AerChemMIP showed that including previously missing interactive sources of chemical species in an ESM ..."
Change to: “Including previously missing interactive sources of chemical species in an ESM has the potential for surprising results in estimates of forcing.”

- line 189 : "Of the three MIPs, AerChemMIP played a unique role ..."
Change to: “AerChemMIP played a role in the quantification of non-CO2 biogeochemical feedbacks (...)”.

- line 205 : "AerChemMIP further points to model difference ..."

Change to: “Even for present-day conditions, outstanding challenges for simulating aerosols persist, e.g., for the concentrations of secondary organic aerosol (Turnock et al., 2020), which have natural and anthropogenic origins (Fan et al., 2022).”

REFERENCES :

- line 482 : "JOURNAL OF CLIMATE" -> "Journal of Climate"
Corrected.

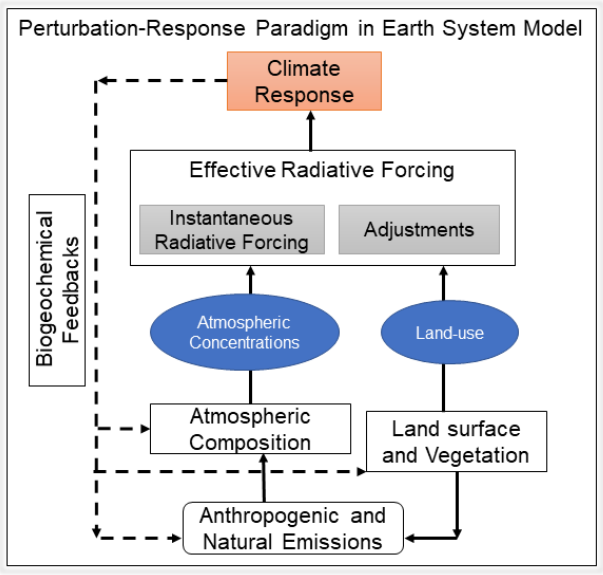
- line 633 : "GEOSCIENTIFIC MODEL DEVELOPMENT" -> "Geoscientific Model Development"
Corrected.

FIGURES :

- Figure 1 : does "Earth System Model" in the red box refer to the AOGCM (component of an ESM)? Maybe this figure can be improved.

This figure has been redrawn and is included below. The revised caption of Figure 1 is: “Schematic depiction of the perturbation-response paradigm in understanding and quantifying climate changes to perturbations using an Earth System model. Blue circles indicate options for simpler ESMs that prescribe perturbations in concentrations and land-use. Climate responses are simulated in the model configuration coupled to an ocean model.”.

Revised Figure 1:



- Figure 2 : "The main goals of AerChemMIP are ..." : "The main goal of AerChemMIP is ..."
Changed as suggested.

- Figure 2 : "... where the emissions or concentrations of the species of interest is perturbed" -> "are" perturbed
Changed as suggested.

- Figure 3 : some arrows have points in colours, some in black
Thanks, all points in black now.

- Figure 3 : in the upper right figure, the temperature in the troposphere should not change (I assume). However, the lines do not completely overlap in the troposphere, which is confusing.
Yes, the lines are overlapping in the toposphere now.

Revised Figure 3: