Letter to GMD editors

To the Editors of GMD,

Attached is our point-by-point response to the reviewer reports of our article, *Reduced complexity model intercomparison project phase 1: Protocol, results and initial observations (gmd-2019-375)*. The latexdiff is included at the end of this document. We would like to thank the reviewers for the time taken to review our paper.

The reviewers comments were very helpful. However, they have resulted in a substantial change to the manuscript. Specifically, we focussed on the MIP description. In order to keep the paper scope manageable, this has meant that we have removed much of the detail on RCMs.

We feel that we have been able to respond to each comment made by the reviewers. In our point-by-point responses we have highlighted the original text and new text/sections and hope this will transparently show how we addressed each individual reviewer comment.

In the responses below, the original reviewer reports are in black, while all our comments are in blue. We have also numbered all the reviewer comments and our replies for clarity. We have *quoted text from the manuscript in grey italics*.

We thank you and the reviewers for the time invested into our manuscript and hope that it now reaches the high standards of *Geoscientific Model Development*.

Best regards,

Zebedee Nicholls and Robert Gieseke (corresponding authors)

In the responses below, the original reviewer reports are in black, while all our comments are in blue. We have also numbered all the reviewer comments and our replies for clarity. We have *quoted text from the manuscript in grey italics*.

Reviewer 1 Comment 1
Dear authors, I appreciate your paper that incorporates and compares a wide variety of different RCMs with different qualities. I would like to contribute to the progress of your RCM inter-comparison project and provide a review on the manuscript submitted to Geoscientific Model Development.

Reviewer 1 Reply 1
Thank you for taking the time to review our paper, it is greatly appreciated by all of the co-authors.

One general comment: We really appreciate the various suggestions for improvement and interesting cross-comparisons of the various modelling group's results. We wholeheartedly agree about the importance of these investigations. However, in response to reviewer 2 and after consulting with the editor, we have had to sharpen the paper's focus and accordingly turned it into a MIP description paper. We do really appreciate the suggestions and hope to be able to respond to them in future work, but we feel we cannot do all of them justice in the confined space we now have.

Reviewer 1 Comment 2
As far as comments on the content are concerned, the question arises how do different RCMs introduce nonlinearities of the temperature response. Your paper is about quantifying the temperature response and does not discuss different concepts that provide conceptual understanding. However, the (equilibrium) temperature response does not always scale linearly with CO2 forcing, and explaining the reader why we have nonlinearities of the temperature response (e.g. explicit feedback temperature dependence, among others) might be helpful for the reader to understand different or common model behavior.

Reviewer 1 Reply 2
Thank you for the comment. We agree that we have not discussed the many different reasons for model differences in any detail. For reasons of scope, we do not feel that we have room to do so in this paper, especially not after the comments of the other reviewers who have asked for further details on the project protocol. We hope to do so in a separate paper and hope that this choice of presenting the manuscript in the style of a MIP description paper is agreeable.

Reviewer 1 Comment 3
Another aspect that is important for an unexperienced reader and related to the former comment is why are different RCMs fitted to different numbers of CMIP models. For instance, some models are likely to runaway in the case of high forcing input, and this runaway can be attributed to different model parameters.

Reviewer 1 Reply 3

Thank you for your comment. In the revised manuscript, we clarify that each model is fitted to a different number of CMIP models due to different calibration choices by different modelling teams. In other words, calibrations depend on each RCM development team's individual capacity.

New text (see lines 437-438 of diff)

> *Instead, the CMIP models to which each RCM is calibrated depends on each RCM development team's capability and the time at which they last accessed the CMIP archives.*

We have also added a clarification of how differences in model parameters have been handled at this very early stage of RCMIP (see new text at beginning of revised Model Configuration Section 3.1, lines 161-170 of diff, not included here for brevity).

Reviewer 1 Comment 4

Further, I can hardly imagine that a parameter which represents feedback temperature dependence is well constrained by the observational record. I wonder how strong model parameters vary between fits to the reference period/observations and abrupt CO2 experiments. Adding brief, explicit paragraphs would be helpful.

Reviewer 1 Reply 4

Thank you for the comment. We agree that we have not discussed the nuances of model constraining at all. For reasons of scope, as in Reply 2 we do not feel that we have room to do so in this MIP description paper. We hope that such work can take place in future research such as https://www.earth-syst-dynam-discuss.net/esd-2019-82/.

Reviewer 1 Comment 5

This also holds true for the discussion on probabilistic projections. You mention very important aspects but how do the different models actually compare?

Reviewer 1 Reply 5

Thank you for the comment. We agree this is an important question but feel it is beyond the scope of the MIP description (see Reply 4).

Reviewer 1 Comment 6

I've a specific comment on the understanding of time-and state-dependent feedback (lines 417-427). It is said that models with time or state-dependent feedback avoid the problem that linear models predict an equally large amplitude to negative radiative forcing as positive radiative forcing. This holds true for state-dependent feedback or the combination of time- and

state-dependent feedback but the temperature response of purely time-dependent feedback scales linearly with forcing.

Reviewer 1 Reply 6

Thank you for the comment. We agree this is an important question but have had to remove this more detailed discussion in response to other review comments because it is beyond the scope of the MIP description. We hope this decision is understandable and that future work can consider this question in more detail.

Reviewer 1 Comment 7

As a short technical note, please revise your plotting routines in the supplementary material.

Reviewer 1 Reply 7

Thank you for your comment, we have updated the plots.

In the responses below, the original reviewer reports are in black, while all our comments are in blue. We have also numbered all the reviewer comments and our replies for clarity. We have *quoted text from the manuscript in grey italics*.

Reviewer 2 Comment 1
I recommend rejecting this paper for three main reasons:

1. The purpose of this paper remains unclear
2. The robustness of the scientific results remains unclear, and there is too little information in this paper to understand the analyses carried out
3. The logic of a substantial number of sentences remains unclear

These issues are the more surprising given the scientific expertise of the large number of co-authors listed on the title page. Given the importance of the results hinted at here, I encourage a re-submission of this manuscript.

In the following, I provide examples for the three overarching issues. I trust that a detailed listing of all issues is unnecessary given the expertise of the panel of authors.

Reviewer 2 Reply 1
Thank you for taking the time to review our paper. We appreciate the thought and consideration that have gone into your review. We believe we can address these major issues and have done so in our revised manuscript.

In response to your reason 1 and after consulting with the editor, we have now re-written the manuscript to make the purpose clearer. Specifically, we have made the paper a MIP description paper, removing discussion of other non-essential ideas. Whilst we think these other ideas are worthy of attention, we agree that such attention belongs in a separate paper in order to keep the key idea of this paper (i.e. the introduction of a new systematic effort to compare reduced complexity climate models) clear.

In response to reason 2, as part of the revisions we have turned our results section into a sample results section (which are more appropriate for MIP description papers). Accordingly, we have significantly softened the language related to any conclusions to make clear that the results are preliminary only and that further research is required to make robust conclusions.

In response to reason 3, we have significantly revised the paper and hope that the logic now makes much more sense.

Reviewer 2 Comment 2
1. According to the title, this paper provides the protocol, results and initial observations of RCMIP. However, the protocol is described on only about half a page, and the results are listed on only about three pages. In fact, much of these three pages describe possible future research

rather than providing actual results. In contrast, half the paper consists of a description of individual RCMIP models. I encourage the authors to more clearly define the purpose of this paper, and to have the text more directly reflect such purpose.

Reviewer 2 Reply 2

Thank you for the suggestion. We have altered the paper to focus on the MIP description and dedicated much more space for this purpose accordingly. Whilst we feel that a discussion of the state of RCMs is important, we acknowledge that it is too much for this paper and after discussing with the editor have accordingly removed it. We plan to cover study of the participating models in separate future research.

Reviewer 2 Comment 3

2. I was unable to follow how the evaluation of RCMIP models has been carried out, and which conclusions one can draw from any such analysis. Which observational datasets were used? What is their uncertainty? Which CMIP6 models were used for the comparison? Which degree of agreement can one expect given, for example, observational uncertainty and natural variability? Which degree of agreement can one expect given the tuning of RCMIP models? How is the statistical significance of model agreement or disagreement calculated? What is actually shown in the figures for individual RCMIP models? How is the result obtained that "46 % of the difference between CMIP5 and CMIP6 is scenario dependent"? Why is there no uncertainty attached to this number? Which assumptions went into its calculation? etc. etc. etc.

Reviewer 2 Reply 3

Thank you for the comment. We agree that our evaluation section was not as clear as it should have been. Given the request for improved clarity, particularly on the MIP description, we no longer feel we have the space to provide the evaluation requested. Accordingly, we have altered our results section so that it is now a sample results section, softened the language related to all conclusions to make clear that they are only preliminary and not comprehensive and leave further evaluation for future work.

As an example of the change, the previous text read (line 543 of diff)

> *When run with the same model, warming projections are higher in the SSPs than the RCPs*

It now reads (lines 452-459 of diff)

> *Finally, we present initial results from running both CMIP5 and CMIP6 generation scenarios (`RCP' and `SSP-based' scenarios respectively) with the same models (Figure 5).*
> *In the small selection of models which have submitted all RCP, SSP-based scenario pairs, the SSP-based scenarios are 0.21\degree C (standard deviation 0.10\degree C across the models' default setups) warmer than their corresponding RCPs (Figure 5(b)).*

This difference is driven by the 0.42 \unit{W m^{-2}} ± 0.26 \unit{W m^{-2}} larger effective radiative forcing in the SSP-based scenarios (Figure 5(d)), which itself is driven by the larger \chem{CO_2} effective radiative forcing in the SSP-based scenarios. As noted previously, these are only initial results, not a comprehensive evaluation and should be treated as such.

Reviewer 2 Comment 4

3. Just some example of unclear logic/grammar/style:

l.23: RCMs do not exchange limited resolution for computational efficiency. They have limited resolution, and are therefore computationally efficient.

Reviewer 2 Reply 4

Thank you for this comment. We agree that there are many ways to describe the trade-offs in RCM design, e.g. one could equally say, 'RCMs are designed to be computationally efficient exploratory tools and hence must have limited resolution'. We will revise the text for clarity.

Old text (see lines 35-37 of diff)

RCMs typically exchange limited spatial and temporal resolution for computational efficiency.

New text (see lines 35-37 of diff)

RCMs are designed to be computationally efficient tools, allowing for exploratory research and have smaller spatial and temporal resolution than complex models.

Reviewer 2 Comment 5

l.32: If it was "unfeasible to perform climate assessments with ESMs", no IPCC reports would exist

Reviewer 2 Reply 5

Thank you for this comment. We agree that taken by itself, this statement would clearly be wrong. We have revised the text to make clear that it would be unfeasible to perform climate assessment of 100s of IAM scenarios with ESMs given the much longer run-times of ESMs and the tight deadlines of the IPCC assessment process.

Old text (see lines 48-53 of diff)

Given there are hundreds of emission scenarios submitted by Integrated Assessment Models in IPCC AR5 and AR6 (available at \url{https://secure.iiasa.ac.at/web-apps/ene/AR5DB} and

\url{https://tntcat.iiasa.ac.at/SspDb}, hosted by the IIASA Energy Program) it is unfeasible to perform climate assessment with the world's most comprehensive models.

New text (see lines 53-59 of diff)

For the IPCC's forthcoming Sixth Assessment (AR6), it is anticipated that the number of scenarios will be in the several hundreds to a thousand (for example, see the full set of scenarios based on the SSPs at \url{https://tntcat.iiasa.ac.at/SspDb}). Both the number of scenarios and the tight timelines of the IPCC assessments render it infeasible to use the world's most comprehensive models to estimate the climate implications of these IAM scenarios.

Reviewer 2 Comment 6
l.40: What is "observationally consistent"?

Reviewer 2 Reply 6
We agree this text is unclear. We have clarified that we are talking about assessing the extent to which model output agrees with the observational record.

Old text (see lines 65-66 of diff)

The resulting projections provide a plausible, observationally consistent range of projections which is large enough to provide useful statistics.

New text (see lines 68-72 of diff)

Probabilistic climate projections are derived by running parametric ensembles of RCM simulations which capture the range of responses consistent with our understanding of the climate system \citep{Meinshausen_2009_b9j8fj,Smith_2018_gdrwm6,Goodwin_2016_gft5nc}.
The resulting ensemble is designed to capture the likelihood that different warming levels are reached under a specific emissions scenario (e.g. 50\% and 66\%) based on the combined available evidence hence is quite different from an ensemble emulating multiple model outputs, which have been produced independently with no relative relationship or probabilities in mind.

Reviewer 2 Comment 7
l.40: Why does only a "large range of projections" provide useful statistics? Later in the paper it is shown that the range of CMIP6 projections is larger than that of RCMIP projections. Does this imply that CMIP6 simulations provide more useful statistics?

Reviewer 2 Reply 7

Thank you for the comment, we agree that the text is unclear. We have updated the text to clarify that the probabilistic distributions derived from RCMs are designed to capture the likelihood that different warming levels are reached under a specific emissions scenario (e.g. 50% and 66%) based on the combined available evidence hence are quite different from an ensemble of multiple model outputs, which have been produced independently with no relative relationship in mind (see Reply 6 for text changes).

Reviewer 2 Comment 8
l.46: Style: "The first is a comparison with observations. These comparisons provide the most direct comparison of model response with the world around us today." What is a 'direct comparison with the world around us'? What is compared with observations? Which observations? etc.

Reviewer 2 Reply 8
We agree this was not as clear as it should have been. We have updated the text for clarity. In particular, to specify that a necessary condition for an SCM is to reproduce historical trends in at least observed global-mean temperature but ideally also ocean heat uptake and carbon content in the atmosphere, land and oceans.

Old text (see lines 124-125 of diff)

> *These comparisons provide the most direct comparison of model response with the world around us today.*

New text (see lines 119-122 of diff)

> *Before using any model, the most obvious question to ask is whether it can reproduce observations of the climate's recent evolution.*
> *For RCMs, the key observation is changes in air and ocean temperatures \citep{Morice_2012_q4f,Cowtan_Way_2014}.*
> *Beyond this, RCMs should also be evaluated against observed changes in ocean heat uptake \citep{Zanna_2019,von_schuckmann_2020} and estimates of carbon content in the air, land and oceans \citep{Friedlingstein_2019}.*

Reviewer 2 Comment 9
l.82: What is "projected warming uncertainty"?

Reviewer 2 Reply 9
We agree this was not as clear as it should have been. We have removed this text and narrowed the focus of our new Science Themes Section 2.

Deleted text (see lines 214-215 of diff)

How do probabilistic ensembles from RCMs compare to each other and observations and what can they tell us about projected warming uncertainty under different scenarios?

Reviewer 2 Comment 10
l.104: "This ensures consistency with CMIP6, albeit at the expense of using the latest data sources". Why is it an expense to use the latest data sources?

Reviewer 2 Reply 10
Our apologies, 'at the expense of' is colloquial english. We have rephrased to 'rather than' for clarity.

Old text (see lines 324-325 of diff)

This ensures consistency with CMIP6, albeit at the expense of using the latest data sources in some cases.

New text (see lines 324-325 of diff)

This ensures consistency with CMIP6, although it means that we do not always use the latest data sources.

Reviewer 2 Comment 11
l.119: "Given their focus on global-mean, annual mean variables we request a range of output variables from each RCM." The logic of this sentence is not clear to me.

Reviewer 2 Reply 11
We have removed this text from the manuscript and have clarified in a new, comprehensive Output Specifications Section 4.

Deleted text (see line 342 of diff)

Given their focus on global-mean, annual mean variables we request a range of output variables from each RCM

Reviewer 2 Comment 12
l.129: 'In the climate response to radiative forcing, the models range from two-box impulse response models to..." Probably should read "In their representation of the climate response to radiative forcing."? etc.etc.

Reviewer 2 Reply 12
Thank you for the suggestion. In response to Reply 2, we have significantly narrowed the focus of the manuscript. As a result, our discussion of the different models is now extremely limited

hence we have removed this sentence. We agree that a discussion of the different types of RCMs is an important one, but it is now clear to us that we do not have the space to do a sufficient job in this paper.

Reviewer 2 Comment 13
The description of models in 2.3 should be harmonized (including the level of details provided) to allow the reader to quickly compare characteristics of different models.

Reviewer 2 Reply 13
Thank you for this recommendation. Given the wide scope of RCMs, we do not feel we can provide a sufficient description of all the different models within this paper and simultaneously describe the MIP. We have removed the large discussion of the different model types and will leave such a discussion for future work. We now only present a very brief overview of the models which have participated to date (see revised Table 1).

Reviewer 2 Comment 14
I do not comment in detail on section 3 as this section needs to be entirely re-written in my view.

Reviewer 2 Reply 14
We have significantly revised this section (now split into Results Section 5 and Extensions Section 6) and hope that provides a much clearer representation of the results of this study.

Reviewer 2 Comment 15
I am sorry that I cannot provide a more positive review at this point. The important results hinted at here are potentially so important that they deserve a more rigorous analysis and description. All the best for revising this study.

Reviewer 2 Reply 15
Thank you for the time taken to do the review. It has been very helpful for us, in particular pointing out where we can improve our manuscript. We hope the revised manuscript better communicates the science we have undertaken, the novelty of our study and the most obvious next steps.

In the responses below, the original reviewer reports are in black, while all our comments are in blue. We have also numbered all the reviewer comments and our replies for clarity. We have *quoted text from the manuscript in grey italics*.

Reviewer 3 Comment 1

The paper aims to introduce the motivation and rational for a model intercomparison of reduced-complexity models. These models are commonly used to interpret (mostly global mean) temperature observations and complex model simulations. The paper introduces scientific questions which can be answered with this typ of models, the experimental design and diagnostics, participating models, and shows first analyses of modeled temperatures for the historical period and scenarios for the next century.

Reviewer 3 Reply 1

Thank you for your review of our paper. We greatly appreciate the time you have put in and have found your comments very helpful, particularly to better define the scope of our paper. We have produced an updated manuscript which we feel is greatly improved thanks to your suggestions. We hope it is a useful first step to helping the 'the big group of people who are confronted with RC output but don't know how to evaluate them' (as well as us as model developers).

Reviewer 3 Comment 2

One important contribution to the ongoing discussion of differences between CMIP5 and CMIP6 is the finding, that about 46% of the additional warming at the end of the 21st century in CMIP6 compared to CMIP5 models stems from differences in the radiative forcing in SSPs and RCPs.

Reviewer 3 Reply 2

Thank you for your comment. In response to Reviewer 2 (specifically their Comments 3 and 14) and after discussion with the editor, we have removed this discussion from this paper. Following your Comment 6 as well, we have removed this discussion and will save it for a paper which has the room to explore it in the detail it deserves. Instead, we have simply presented the difference between the SSPs and RCPs in the sample results section and stated that there is a difference in the results submitted to date but further evaluation is required to fully understand why because these are only preliminary results (see lines 471-479 in diff).

Reviewer 3 Comment 3

I applaud the endeavor to conduct an RCMIP. This will be very useful, both for people using single RCs but also the big group of people who are confronted with RC output without knowing how to evaluate them (both the impact user side and the GCM modeler side). However, the paper in its current format is weak and obscure and does not allow me to draw clear conclusions. It should not be too much effort to improve the paper, as it is mostly "just" improving the presentation, explanation, arguments, clarity. No new simulations are necessary.

Reviewer 3 Reply 3

Thank you for your positive comments. We agree that the paper required significant updates and have done so in the revised manuscript. In particular, we have focussed solely on the presentation of the MIP, leaving comprehensive evaluation of the results for future study.

Reviewer 3 Comment 4

Major comments 1) The differentiation between possible/future research questions and the ones addressed (and answered?) in this paper is unclear. I think possible questions do not belong into a paper. Anybody can come up with some vage questions. I read a paper to learn about what has been done and how I can used this for my own research. What somebody (who?) might be doing/planning/considering can be discussed in conferences etc. not in a scientific paper.

Reviewer 3 Reply 4

Thank you for your comment. We agree that these comments do not belong in a paper and have removed most of them, moving the relevant ones to the revised Extensions Section 6.

Reviewer 3 Comment 5

2) From a model intercomparison paper, I'd like to learn how I can use the output, which criteria have been used to select the models, which experiments have been conducted, . . . technical parameters, what can I learn from your effort. There are a lot of MIP-explaining papers out there. I suggest to imitate one of these in the structure and focus of the paper.

Reviewer 3 Reply 5

Thank you for your comment. We have updated the paper to read like a MIP-explaining paper (most closely following the structure of the RCEMIP paper) and hope that this makes the purpose of RCMIP and how it can be used much clearer (particularly revised Science Themes Section 2, Simulation Design Section 3 and Output Specifications Section 4) .

Reviewer 3 Comment 6

3) The interpretation of the differences between CMIP5 and CMIP6 scenarios is a major scientific contribution to the ongoing discussion. It is extremely relevant for writing the IPCC report. As such it belongs into a more visible, less technical journal and it needs to be highlighted. Here, this finding is buried towards the end of the paper and I get the sense that this is because the science behind this finding is actually not really well understood, at least I don't from reading the paper. What's the relationship of this finding with the paper of Forster et al. 2019, who's estimate for the impact of the different scenarios (in CMIP5 vs 6) to surface temperature is much smaller.

Reviewer 3 Reply 6

Thank you for your comment. We agree that such a discussion requires a much more in depth discussion and have accordingly removed it, saving it for future research (see also Reply 2).

Reviewer 3 Comment 7

Minor comments Title: “initial observations” - these are modeling results, reformulate

Reviewer 3 Reply 7
Thank you for your comment. We have updated the title to, *Reduced complexity model intercomparison project phase 1*, to better reflect the revised manuscript’s focus on MIP description.

Reviewer 3 Comment 8
line 12: output - in what

Reviewer 3 Reply 8
This was indeed unclear. We have reworded the sentence for clarity and to remove the emphasis on the RCP-SSP difference as requested in Comment 6.

Old text (see lines 18-20 of diff)

> *Comparing our results to the difference between CMIP5 and CMIP6 output, we find that the change in scenario explains approximately 46\% of the increase in higher end projected warming between CMIP5 and CMIP6.*

New text (see lines 14-22 of diff)

> *We present illustrative figures comparing model output with historic global surface air temperature (GSAT) observations, showing probabilistic projections, demonstrating different calibrations with CMIP model output as well as temperature change against cumulative emissions, and exploring differences between CMIP5’s Representative Concentration Pathways (RCPs) and CMIP6’s SSP-based (Shared Socioeconomic Pathways based) scenarios.*

Reviewer 3 Comment 9
line 12: change in scenario - please explain much more thorough throughout the paper: Why has the scenario been changed? Were they not supposed to be traceable (i.e. SSP8.5 approx RCP8.5? How does this fit with the 46% additional warming due to different scenarios

Reviewer 3 Reply 9
Thank you for the comment. Following your comments and Reviewer 2 we have removed the (attempted) discussion of the percentage difference between the SSPs and RCPs and instead simply presented raw results from the models along with the caveat that these results are not

comprehensive. We agree that exploring this change in more detail is required and hope future study can do so.

Reviewer 3 Comment 10
line 15: “as first anticipated” - by whom and why?

Reviewer 3 Reply 10
This expression was indeed vague and has been removed (line 25 of diff).

Reviewer 3 Comment 11
line 16: “provide results available . . .” which results? (the authors can pick/will find the results they need themselves. It’s not a scientific finding to plan to provide results.)

Reviewer 3 Reply 11
Thank you for highlighting this odd formulation, we have removed this line from the revised manuscript (line 26-27 of diff).

Reviewer 3 Comment 12
line 28: “exploring interacting uncertainties” - explain which parts of the climate system can interact in these models - mostly they only contain surface temperature and some forcing agents and parameterized ocean heat uptake?

Reviewer 3 Reply 12
Thank you for this suggestion. Many of the models contain parameterised representations of the carbon cycle, non-CO2 gas cycles and land surface, all of which can interact and represent (in a parameterised way) many of the key feedbacks in the Earth System. We now clarify this with some examples in the revised manuscript and hope to provide a follow up paper with much more detail on the models' different structures in future (given we do not have the space to do so here).

New text (see lines 77-83 of diff)

> *RCMs also play the role of `integrators of knowledge', examining the combined response of multiple interacting components of the climate system.*
> *The most comprehensive RCMs will include (highly parameterised) representations of the carbon cycle, permafrost, non-\chem{CO_2} gas cycles, aerosol chemistry, temperature response to radiative forcing, ocean heat uptake, sea-level rise and all their interactions and feedbacks.*
> *More complex models cannot include as many interactive components without the computational cost quickly becoming prohibitive for running multiple century-long simulations.*
> *As a result, RCMs are able to examine the implications of the Earth System's feedbacks and interactions in a way which cannot be done with other techniques.*

Reviewer 3 Comment 13
line 41: useful statistics - useful for what?

Reviewer 3 Reply 13
Thanks for pointing out this unclear sentence, it has been removed (line 66 of diff).

Reviewer 3 Comment 14
line 60: . . . to understand their strengths, weaknesses and limitation so that we can make more confident, informed conclusions from their quantitative results" - yes, great, it would be good if all these points were indeed discussed in the conclusion in a clear manner.

Reviewer 3 Reply 14
Thank you for the comment. We have expanded the discussion in the conclusion (see revised Section 7) and put the areas for further examination (as there are many we have not yet covered) in the new Extensions Section 6.

Reviewer 3 Comment 15
line 75: what's a lifetime of an RCMIP?

Reviewer 3 Reply 15
Thank you for the comment, we have removed this unclear phrasing (line 201 of diff).

Reviewer 3 Comment 16
line 75 following: For this paper, specify the questions you are actually answering. Here is mixed list is given of what is and could be done. Could/should/would is used much more in this paper than in usual scientific literature. For a MIP paper, a discussion of future/possible questions is fine, but these should be listed in one concise place and not dominate the paper.

Reviewer 3 Reply 16
Thanks for pointing this out, during the restructuring of the manuscript we have made a clear Science Themes (Section 2) and consolidated the most important of the future questions into a single extensions section.

Reviewer 3 Comment 17
line 77: "some aspects will receive less attention here than others" - and later: more precise language would help to make this a scientific and useful paper and not an opinion piece.

Reviewer 3 Reply 17
Thank you for the comment. We have removed this line (line 203 of diff) and clarified the scope of the paper, pushing all the extensions into a single extensions section which we hope helps to restore the scientific tone of the paper.

Reviewer 3 Comment 18
line 78: “what can they tell us about” ?

Reviewer 3 Reply 18
We agree this is an awkward, colloquial phrasing, it has been removed (line 214-215 of diff).

Reviewer 3 Comment 19
line 84: Experimental design: This section is not actually describe the experimental design fully. Or at least, after the section, I wouldn’t be able to replicate what you did. From the title, I expect a list of experiments, their input, assumptions, rational, in a clear understandable fashion. Right now the section is a collection of random issues (a lot of detail about emission some specific scenarios, non about others)

Reviewer 3 Reply 19
Thank you for the comment. We have overhauled the experimental design so it actually describes what we did in a clear, reproducible fashion (see revised Sections 3 and 4).

Reviewer 3 Comment 20
line 96: What’s the standard set of inputs from CMIP5 and CMIP6?

Reviewer 3 Reply 20
Thank you for the comment. We agree this is not sufficiently clear and have clarified in much more detail in the revised experimental design section (see diff lines 296 and 312 for removed text and revised Sections 3 for new description).

Reviewer 3 Comment 21
line 120: Diagnostics: I expect to learn how I can use this data. What’s the available output? What’s the rational for it?

Reviewer 3 Reply 21
Thank you for the comment, we have updated the manuscript to include a standalone output and diagnostics section which outlines the available output and why it is requested (see new Section 4).

Reviewer 3 Comment 22
line 128: How is an RC model defined? What’s the criteria to be included in your comparison? Table one is a nice overview. I suggest to move even more information from the text into the table: What are the input variables and assumptions about them? On what data are they tuned? Add a paragraph about similarities and differences among the models. What do they all share? Some of them seem to use the same basic equations. Are there classes of RCs? Could you draw a genealogy? Which ones are structurally more similar? Which ones are fully independent? From the text, I e.g. do not get a good sense of the difference between FaIR and CICERO-SCM.

Reviewer 3 Reply 22
Thank you for your comment. We have updated the text to provide clarity (see lines 110-114 of diff).

> *In the RCMIP community call (available at rcmip.org) RCMs were broadly defined as follows: ``[...] RCMIP is aimed at reduced complexity, simple climate models and small emulators that are not part of the intermediate complexity EMIC or complex GCM/ESM categories."*
> *In practice, we encouraged (and encourage) any group in the scientific community who identifies with the label of RCM to participate in RCMIP (see Table \ref{tab:rcmip-model-overview} for an overview of the models which participated in RCMIP Phase 1).*

We agree that the topic of differences and similarities between RCMs is an important one. However, we do not feel that we have sufficient space within this MIP description paper to do it justice hence, after consulting with the editor, have removed all but the most important details from this paper. We hope to provide a follow up paper which does discuss model details and genealogies in far more detail (explaining the difference between FaIR and CICERO-SCM, for example, is not a task which can be accomplished in a single paragraph).

Reviewer 3 Comment 23
line 156 what does it imply to have two or three timescales?

Reviewer 3 Reply 23
We have removed discussion of the intricacies of different model setups due to space constraints. We hope to provide a follow up paper which can cover this topic with the detail it deserves in future.

Reviewer 3 Comment 24
line 433-436 I do not understand the sentence “These probabilistic . . . “

Reviewer 3 Reply 24
Thank you for the comment, we have removed this unclear sentence (line 502 of diff). We have added a clearer discussion of the role of the probabilistic distributions in the new introduction.

New text (see lines 68-72 of diff)

> *Probabilistic climate projections are derived by running parametric ensembles of RCM simulations which capture the range of responses consistent with our understanding of the climate system \citep{Meinshausen_2009_b9j8fj,Smith_2018_gdrwm6,Goodwin_2016_gft5nc}.*

The resulting ensemble is designed to capture the likelihood that different warming levels are reached under a specific emissions scenario (e.g. 50\% and 66\%) based on the combined available evidence hence is quite different from an ensemble emulating multiple model outputs, which have been produced independently with no relative relationship or probabilities in mind.

Reviewer 3 Comment 25

line 442-444 “Given that . . . “ is pure speculation. Somebody might be doing these experiments, who knows, maybe not, . . . what’s the purpose of this “information” here? Is this a call to the community that these experiments should be done? Are you planning to do them? Can I expect the results in phase 2? Maybe this is all about precision of formulation only? It’s so vague, I don’t know what to do with this information.

Reviewer 3 Reply 25

Thank you for the comment. We agree it was far too vague. We have moved it to the new extensions section and made clear that we are calling on the community to do such experiments in future.

Old text (see lines 518-524 of diff)

Given that variations in both model structure and calibration technique influence probabilistic projections, an area for future research could be to try and disentangle the impact of these two components.
Such an experiment could involve constraining models with the same constraining technique or constraining a single model with two different techniques.

New text (see lines 518-524 of diff)

Following this, there is clearly some variation in probabilistic projections.
However, what is not yet known is the extent to which variations in model structure, calibration data and calibration technique drive such differences.
Investigating this `known unknown' would help understand the limits of probabilistic projections and their uncertainties.
Experiments could involve constraining two different models with the same constraining technique and data, constraining a single model with two different techniques but the same data or constraining a single model with a single technique but two different datasets.

Reviewer 3 Comment 26

line 448: Developing a method . . .” “Such results would enhance ...“ same as the point above: Are you suggesting to do this? Are you doing this? Should I do it? Why don’t you do the research first and then tell me about the outcome?

Reviewer 3 Reply 26
Thank you for the comment, like for Comment 25 we have moved the text to the new Extensions Section 6 (lines 537-542 of diff).

Reviewer 3 Comment 27
line 464 following: Isn't this way too important to be buried in the Supplemental Material?

Reviewer 3 Reply 27
Thank you for the comment. As discussed in Reply 2, we have revised the manuscript to present the raw data here and caveated the results by acknowledging that these are raw outputs and further evaluation is required to make strong conclusions. We have also provided revised Figure 5 which moves

Reviewer 3 Comment 28
line 469: monotonic relationship?

Reviewer 3 Reply 28
We have removed this text from the manuscript. For clarity, in this context monotonic simply means that if CO2 concentrations increase, CO2 effective radiative forcing increases.

Removed text (lines 560-561 of diff)

> *However, given the monotonic relationship between \chem{CO_2} concentrations and effective radiative forcing \citep{IPCC_2013_WGI_Ch_8}, it is likely that the same mechanisms are driving at least part of the increase between CMIP5 and CMIP6 projections.*

Reviewer 3 Comment 29
line 472 "At this stage, this residual is most likely explained . . ." and in two months you might change your mind or interpretation? Why not waiting with writing a paper until clear results and their interpretation materialize?

Reviewer 3 Reply 29
Thank you for the comment. We agree that the speculation was not helpful so have presented only raw results in the revised manuscript with appropriate caveats (see also Reply 2).

Removed text (lines 563-566 of diff)

> *At this stage, this residual is most likely explained by a change in the models submitting results to CMIP, which appear to be more sensitive to changes in atmospheric GHG concentrations in CMIP6 than in CMIP5 \citep{wyser_2019_scojd2,voldoire_2019_98cjk3,voosen_2019_9sc8df}.*

However, CMIP6 analysis is ongoing and should be considered before making strong conclusions about the robustness of these findings.

Reviewer 3 Comment 30
line 476 “A number of experiments have not been discussed here. . .” . . .?

Reviewer 3 Reply 30
Thank you for the comment. We agree this is vague and not helpful. We have revised our results section into an ‘Illustrative results’ section to make clear that the presented results are to illustrate the usefulness of the MIP, rather than being intended as detailed evaluation.

Old text (see lines 567-569 of diff)

A number of experiments have not been discussed here which would shed light on the differences between the RCMs in a number of other components.

New text (see lines 416-418 of diff)

The groups which have participated have submitted a number of results.
We provide a brief overview of these here to give an initial assessment of the diversity of models which have submitted results to date.
However, this is not intended as a comprehensive comparison or evaluation.

Reviewer 3 Comment 31
line 480 I can’t follow. Your “Conclusion” is an Outlook. Both, a conclusion and an outlook would be useful. I suggest to re-write the entire paper and discuss solely the models and (clarified) experimental set up and the *results* and then have one dedicated Outlook section with all your if/when/could/should/might items and maybe a clear plan for phase 2.

Reviewer 3 Reply 31
Thank you for the recommendation, we have found it very helpful. We have re-written the entire paper as suggested to make it much more focussed on the MIP description (particularly splitting out separate Sample Results Section 5, Extensions Section 6 and Conclusions Section 7). After discussions with the editor, we decided to remove the discussion of the models as there is not sufficient space to discuss both the models and the MIP to the level of detail required. We hope to provide a detailed model description paper in future.

Reviewer 3 Comment 32
Fig.1 suggestions: Shade CMIP5 and CMIP6 models? There’s too much information in this plot, I can’t differentiate the lines. Maybe add panels with each RC to the SM?

Reviewer 3 Reply 32
Thank you for the suggestion, we have updated the plots to make clear the different lines.

Reviewer 3 Comment 33
Fig.2 a again, I can't see which information is relevant here. e.g. why are there not "hector" for SSP126.

Reviewer 3 Reply 33
Thank you for the comment, we have updated the plots to highlight the key story (i.e. the degree to which RCMs reproduce the target CMIP6 model's behaviour). In the revised manuscript, we clarify that the calibrations depend on each RCM development team's individual capacity hence there is no Hector output for ssp126.

New text (see lines 435-438 of diff)

> *Each RCM is calibrated to a different number of CMIP models (some RCMs provide no calibrations at all) because there is no common calibration data resource.*
> *Instead, the CMIP models to which each RCM is calibrated depends on each RCM development team's capability and the time at which they last accessed the CMIP archives.*

Reviewer 3 Comment 34
Stretch all plots into the horizontal.

Reviewer 3 Reply 34
Thank you for the suggestion, we have updated the plots.

Reviewer 3 Comment 35
Why do the RCM lines stop earlier than the GCM line in panel c)? Use year 1, 2, 3 instead of 1850, 2000, . . . this is very confusing for idealized experiments.

Reviewer 3 Reply 35
Thank you for the comment. We only requested the abrupt-4xCO2 experiment be run until 2500 in our experiment protocol but the CMIP6 run has continued longer. We agree this is confusing and have updated the plot to use standard axis limits to remove this confusion as well as to use a more standard time axis.

Reviewer 3 Comment 36
Table 2 "[TO DO . . .]"

Reviewer 3 Reply 36
Thank you for picking this up, we have fixed the reference.

Reviewer 3 Comment 37

Fig.4 It's hard to see the point here (SM fig. 3 and 4 are much clearer). maybe change shading/colors? The information of the historical is not needed at this point anymore, the figures could start at year 2000 or so and then stretched. Maybe this would help to make the information more digestible? Adding versions of SM Fig. 3 and 4 could help to make this point stronger in the paper.

Reviewer 3 Reply 37

Thank you for the comment. We have updated the figure as suggested (and in line with Reply 2 and our softening of conclusions). The difference between the RCP and SSP-based scenario pairs is now much clearer (see revised Figure 5).

Reduced ~~complexity model intercomparison project phase~~ Complexity Model Intercomparison Project (Phase 1~~: Protocol, results and initial observations~~)

Zebedee R. J. Nicholls[1,2], Malte Meinshausen[1,2,3], Jared Lewis[1], Robert Gieseke[4], Dietmar Dommenget[5], Kalyn Dorheim[6], Chen-Shuo Fan[5], Jan S. Fuglestvedt[7], Thomas Gasser[8], Ulrich Golüke[9], Philip Goodwin[10], Corinne Hartin[6], Austin P. Hope[11], Elmar Kriegler[3], Nicholas J. Leach[12], Davide Marchegiani[5], Laura A. McBride[13], Yann Quilcaille[8], Joeri Rogelj[8,14], Ross J. Salawitch[11,13,15], Bjørn H. Samset[7], Marit Sandstad[7], Alexey N. Shiklomanov[6], Ragnhild B. Skeie[7], Christopher J. Smith[8,16], Steve Smith[6], Katsumasa Tanaka[17,18], Junichi Tsutsui[19], and Zhiang Xie[5]

[1]Australian–German Climate and Energy College, The University of Melbourne, Parkville, Victoria, Australia
[2]School of Earth Sciences, The University of Melbourne, Parkville, Victoria, Australia
[3]Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, Potsdam, Germany
[4]Unaffiliated
[5]Monash University, School of Earth, Atmosphere and Environment, Clayton, Victoria 3800, Australia
[6]Joint Global Change Research Institute, Pacific Northwest National Laboratory, College Park, MD, USA
[7]CICERO Center for International Climate Research, Oslo, Norway
[8]International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria
[9]BI Norwegian Business School, Nydalsveien 37, 0484 Oslo, Norway
[10]School of Ocean and Earth Science, University of Southampton, Southampton, UK
[11]Department of Atmospheric and Oceanic Sciences, University of Maryland-College Park, College Park, 20740, USA
[12]Department of Physics, Atmospheric Oceanic and Planetary Physics, University of Oxford, United Kingdom
[13]Department of Chemistry and Biochemistry, University of Maryland-College Park, College Park, 20740, USA
[14]Grantham Institute for Climate Change and the Environment, Imperial College London, UK
[15]Earth System Science Interdisciplinary Center, University of Maryland-College Park, College Park, 20740, USA
[16]Priestley International Centre for Climate, University of Leeds, UK
[17]National Institute for Environmental Studies (NIES), Tsukuba, Japan
[18]Laboratoire des Sciences du Climat et de l'Environnement (LSCE), Commissariat à l'énergie atomique et aux énergies alternatives (CEA), Gif sur Yvette, France
[19]Central Research Institute of Electric Power Industry, Abiko, Japan

Correspondence: Zebedee Nicholls (zebedee.nicholls@climate-energy-college.org)

Abstract.

~~Here we present results from the first phase~~ Reduced complexity climate models (RCMs) are critical in the policy and decision making space, and are directly used within multiple Intergovernmental Panel on Climate Change (IPCC) reports to complement the results of more comprehensive Earth System Models. To date, evaluation of RCMs has been limited to a few independent studies. Here we propose a systematic evaluation of RCMs in the form of the Reduced Complexity Model Intercomparison Project (RCMIP). ~~RCMIP is a systematic examination of reduced complexity climate models (RCMs), which are used to complement and extend the insights~~ We have performed Phase 1 of RCMIP with two scientific themes: examining how RCMs compare to observations and how RCMs compare to results from more complex ~~Earth System Models (ESMs), in~~

particular climate models such as those participating in the Sixth Coupled Model Intercomparison Project (CMIP6). In Phase 1 of RCMIP, with 14 participating models namely ACC2, AR5IR (2 and 3 box versions), CICERO-SCM, ESCIMO, FaIR, GIR, GREB, Hector, Held et al. two layer model , MAGICC, MCE, OSCAR and WASP, we highlight the structural differences across various RCMs and show that RCMs are capable of reproducing global-mean We also present our standardised data formats, experiment protocols and output specifications. So far 15 models have participated and submitted results for over 50 experiments. We present illustrative figures comparing model output with historic global surface air temperature (GSAT) changes of ESMs and historical observations. We find that some RCMs are capable of emulating the GSAT response of CMIP6 models to within a root-mean square error of 0.2(of the same order of magnitude as ESM internal variability) over a range of scenarios. Running the same model configurations for both RCP and SSP scenarios, we see that the SSPs exhibit higher effective radiative forcing throughout the second half of the 21st Century. Comparing our results to the difference between CMIP5 and CMIP6 output, we find that the change in scenario explains approximately 46% of the increase in higher end projected warming observations, showing probabilistic projections, demonstrating different calibrations with CMIP model output as well as temperature change against cumulative emissions, and exploring differences between CMIP5's Representative Concentration Pathways (RCPs) and CMIP6's SSP-based (Shared Socioeconomic Pathways based) scenarios. Further research on these and other questions can build on the open data and open source processing code provided with this paper. This suggests that changes in ESMs from CMIP5 to CMIP6 explain the rest of the increase, hence the higher climate sensitivities of available CMIP6 models may not be having as large an impact on GSAT projections as first anticipated. A second phase of RCMIP will complement RCMIP Phase 1 by exploring probabilistic results and emulation in more depth to provide results available for the IPCC's Sixth Assessment Report author teams.

1 Introduction

In an ideal world, sufficient computing power would Sufficient computing power to enable running our most comprehensive, physically complete climate models for every application of interest is not available. Thus, for many applications less computationally demanding approaches are used. However, computational limits do exist so we must sometimes turn to other approaches . One common approach is the use of reduced complexity climate models (RCMs), also known as simple climate models (SCMs).

RCMs typically exchange limited are designed to be computationally efficient tools, allowing for exploratory research and have smaller spatial and temporal resolution for computational efficiency. Specifically, they usually focus on than complex models. Typically, they describe highly parameterised macro properties of the climate system. Usually this means that they simulate the climate system on a global-mean, annual-mean quantitiesscale although some RCMs have slightly higher spatial

and/or temporal resolutions. As a result ~~, they are usually~~ of their highly parameterised approach, RCMs can be on the order of a million or more times faster than more complex models (in terms of simulated model years per unit CPU time).

The computational efficiency of RCMs means that they can be used where computational constraints would otherwise be limiting. ~~Such cases include performing climate assessment for a large number of scenarios, exploring~~ For example, some applications of Integrated Assessment Models (IAMs) require iterative climate simulations. As a result, hundreds to thousands of climate realisations must be integrated by the IAM for a single scenario to be produced. RCMs also enable the exploration of interacting uncertainties from multiple parts of the climate system or the constraining of unknown parameters by combining multiple lines of evidence in an internally consistent setup. In the ~~IPCC context~~ context of the assessment reports of the Intergovernmental Panel on Climate Change (IPCC), a prominent example is the climate assessment of ~~the WGIII socioeconomic scenarios . Given there are hundreds~~ socioeconomic scenarios by IPCC Working Group 3 (WGIII). Hundreds of emission scenarios ~~submitted by Integrated Assessment Models in IPCC~~ were assessed in the IPCC's Fifth Assessment Report (AR5~~and AR6 (~~, see **?**) as well as its more recent Special Report on Global Warming of 1.5°C (SR1.5, see **??**). (Scenario data is available at https://secure.iiasa.ac.at/web-apps/ene/AR5DB and ~~,~~ https://data.ene.iiasa.ac.at/iamc-1.5c-explorer/ for AR5 and SR1.5 respectively, both databases are hosted by the IIASA Energy Program)~~it is unfeasible to perform climate assessment with the world~~. For the IPCC's ~~most comprehensive models. Instead, reduced complexity models are used.~~

~~Beyond their computational efficiency, RCMs also offer conceptual simplicity. This gives them a second use: aiding in interpreting results from higher complexity models or observations. While we think this second use is also valuable (see e. g. **?** and **?**), it is beyond the scope of this paper~~forthcoming Sixth Assessment (AR6), it is anticipated that the number of scenarios will be in the several hundreds to a thousand (for example, see the full set of scenarios based on the SSPs at https://tntcat.iiasa.ac.at/SspDb). Both the number of scenarios and the tight timelines of the IPCC assessments render it infeasible to use the world's most comprehensive models to estimate the climate implications of these IAM scenarios.

~~When RCMs are used to overcome computational constraints, they are typically used in one of two ways. 'Emulation~~There are two key modes of use which are relevant for the assessment of a large number of IAM scenarios. The first is 'emulation' mode, where the RCMs are run in a setup which has been ~~tuned~~ calibrated to reproduce the behaviour of a Coupled Model Intercomparison Project (CMIP) (**??**) model as closely as possible over a range of scenarios. ~~'Probabilistic~~The second is 'probabilistic' mode, where the RCMs are run with a parameter ensemble which captures the uncertainty in ~~historical observations . The resulting projections provide a plausible, observationally consistent range of projections which is large enough to provide useful statistics. In some cases they are also used in a combination of the two.~~ estimates of specific Earth system quantities, be it observations of historical global mean temperature increase, radiative forcing, ocean heat uptake, or cumulative land or ocean carbon uptake. Probabilistic climate projections are derived by running parametric ensembles of RCM simulations which capture the range of responses consistent with our understanding of the climate system (**???**). The resulting ensemble is designed to capture the likelihood that different warming levels are reached under a specific emissions scenario (e.g. 50% and 66%) based on the combined available evidence hence is quite different from an ensemble emulating multiple model outputs, which have been produced independently with no relative relationship or probabilities in mind. The two approaches, emulation of complex models and historically constrained probabilistic mode, can also be combined, e.g. where historical constraints

are very weak. For example, the MAGICC6 probabilistic setup used in ~~the IPCC's Fifth Assessment Report (?)~~ AR5 (?) used randomly drawn emulations for the carbon cycle response whilst using a probabilistic parameter ensemble for the climate response to radiative forcing (?).

~~The validity of both approaches~~ RCMs also play the role of 'integrators of knowledge', examining the combined response of multiple interacting components of the climate system. The most comprehensive RCMs will include (highly parameterised) representations of the carbon cycle, permafrost, non-CO_2 gas cycles, aerosol chemistry, temperature response to radiative forcing, ocean heat uptake, sea-level rise and all their interactions and feedbacks. More complex models cannot include as many interactive components without the computational cost quickly becoming prohibitive for running multiple century-long simulations. As a result, RCMs are able to examine the implications of the Earth System's feedbacks and interactions in a way which cannot be done with other techniques.

1.1 Evaluation of reduced complexity climate models

The validity of the RCM approach rests on the premise that RCMs are able to replicate the ~~response characteristics~~ behaviour of the Earth ~~System and~~ system and response characteristics of our most complete models. ~~This abilityis generally quantified in two different ways. The first is a comparison with observations~~Over time, multiple independent efforts have been made to evaluate this ability. In 1997, an IPCC Technical Paper (?), investigated the simple climate models used in the IPCC Second Assessment Report and compared their performance with idealised Atmosphere-Ocean General Circulation Model (AOGCM) results. Later, ? compared the climate components used in IAMs, such as DICE (?), FUND (?) and the RCM MAGICC (version 4 at the time (?)), which is used in several IAMs. They focused on five CO_2-only experiments to quantify the differences in the behaviour of the RCMs used by each IAM. ? extended the work of ? to consider the impact of non-CO_2 climate drivers in the RCPs. Recently, ? proposed a series of impulse tests for simple climate models in order to isolate differences in model behaviour under idealised conditions.

Building on these efforts, an ongoing comprehensive evaluation and assessment of RCMs requires an established protocol. The Reduced Complexity Model Intercomparison Project (RCMIP) proposed here provides such a protocol (also see rcmip.org). We aim for RCMIP to provide a focal point for further development and an experimental design which allows models to be readily compared and contrasted. We believe that a comprehensive, systematic effort will result in a number of benefits seen in other MIPs (?) including building a community of reduced complexity modellers, facilitating comparison of model behaviour, improving understanding of their strengths and limitations, and ultimately also improving RCMs.

RCMIP focuses on RCMs and is not one of the official CMIP6 (?) endorsed intercomparison projects that are designed for Earth System Models. However, RCMIP does replicate selected experimental designs of many of the CMIP-endorsed MIPs, particularly the DECK simulations (?), ScenarioMIP (?), AerChemMIP (?), C4MIP (?), ZECMIP (?), DAMIP (?) and PMIP4 (?). Hence whilst RCMIP is not a CMIP6 endorsed intercomparison, its design is closely related in the hope that its results may be useful beyond the RCM community.

In what follows, we describe RCMIP Phase 1. In section 2, we detail the domain of RCMIP Phase 1 and its scientific objectives. In sections 3 and 4, we described the simulations performed and outputs requested from each model. In section 5

we present sample results from RCMIP Phase 1, before presenting possible extensions to RCMIP Phase 1 and conclusions in sections 6 and 7.

2 Science themes

In the RCMIP community call (available at rcmip.org) RCMs were broadly defined as follows: "[...] RCMIP is aimed at reduced complexity, simple climate models and small emulators that are not part of the intermediate complexity EMIC or complex GCM/ESM categories." In practice, we encouraged and encourage any group in the scientific community who identifies with the label of RCM to participate in RCMIP, see Table 1 for an overview of the models which participated in RCMIP Phase 1.

RCMIP Phase 1 focuses on evaluation of RCMs. Specifically, comparing them against observations of the Earth System and the output of more complex models from CMIP5 and CMIP6 within two scientific themes.

Theme 1: To what extent can reduced complexity models reproduce observed ranges of key climate change indicators (e.g. surface warming, ocean heat uptake, land carbon uptake)?

The first theme focuses on evaluating models against observations. Before using any model, one important question to ask is whether it can reproduce observations of the climate's recent evolution. For RCMs, the key observation is changes in air and ocean temperatures (**??**). Beyond this, RCMs should also be evaluated against observed changes in ocean heat uptake (**??**) and estimates of carbon content in the air, land and oceans (**?**).

These comparisons evaluate the extent to which the model's approximations cause its response to deviate from observational data. ~~These comparisons provide the most direct comparison of modelresponse with the world around us today.~~ However, ~~given the~~ most RCMs can be calibrated, i.e. have their parameters adjusted, such that they reproduce our best-estimate (typically median) observations. Hence, where available, we also evaluate the extent to which RCMs can be configured to reproduce the range of available observational estimates too. The handling of such observational estimates, particularly their uncertainties, is a complex topic in and of itself. In RCMIP we rely on published estimates and make basic assumptions about how their uncertainty estimates should be compared to model output ranges, each of which we detail when the comparison is performed.

Given the limited amount of observations available and the ease of calibration of RCMs, comparing only with observations leaves us with little understanding of how RCMs perform in scenarios apart from a historic one in which anthropogenic emissions are ~~causing the climateto warm. Given our range of plausible~~ heating the climate. Recognising that there are a range of possible futures, it is vital to also assess RCMs in other scenarios~~(e. g. reducing~~ . Prominent examples include stabilising or falling anthropogenic emissions, ~~instantaneous changes in atmospheric CO2 concentrations, modifications to the solar constant). For example, an RCM that exhibits effectively constant climate feedbacks might be able to replicate an ESM's point response under an idealised scenario, but might not compare well to the longer-term response under either higher forcing or lower overshoot scenarios (?).~~ strong mitigation of non-CO_2 climate forcers and scenarios with CO_2 removal. The limited observational set motivates RCMIP's second theme: evaluation against more complex models.

~~Whilst the results~~ **Theme 2: To what extent can reduced complexity models emulate the response of more complex models?**

Whilst the response of more comprehensive models may not represent the behaviour of the actual ~~earth system~~Earth System, they are the best ~~representation of our knowledge of the earth system's behaviour. Comparing RCM behaviour with more complex model behaviour in these 'non-historical' experiments allows us to~~ available representation of the Earth System's physical processes. By evaluating RCMs against more complex models, we can quantify the extent to which ~~RCMs can capture the range of responses~~ the simplifications made in RCMs limit their ability to capture physically-based model responses. For example, the extent to which the approximation of a constant climate feedback limits an RCM's ability to replicate ESMs' longer-term response under either higher forcing or lower overshoot scenarios (**?**).

In combination, these two research themes examine how well the reduced complexity approach can a) reproduce historical observations of the climate and b) respond to scenarios other than the recent past in a way which is consistent with our best understanding of the ~~earth system~~Earth system's physical and biogeochemical processes.

~~In~~

3 Simulation design

RCMIP Phase 1 ~~of RCMIP we benchmark current RCM performance by comparing them with each other, observations and CMIP results over a range of experiments. This allows us to understand their strengths, weaknesses and limitations so that we can make more confident, informed conclusions from their quantitative results. RCMIP focuses on RCMs and is not one of~~ includes over 50 experiments. To help modelling groups prioritise model runs and ensure comparibility of core experiments three tiers of model runs and output variables were defined. Ideally at least all Tier 1 scenarios and variables for a default model version should be submitted. The following describes the simulation design, model runs as well as data sources and format of RCMIP.

3.1 Model configuration

RCMs are usually highly flexible. Their response to anthropogenic and natural drivers strongly depends on the configuration in which they are run (i.e. their parameter values). To mitigate this as a cause of difference between models in RCMIP Phase 1, we have requested that all models provide one set of simulations in which their equilibrium climate sensitivity is equal to 3°C. While this does not define the entirety of a model's behaviour, it removes a major cause of difference between model output which is not related to model structure. On top of these 3°C climate sensitivity simulations, we have also invited groups to submit other default configurations, where each participating modelling group is free to choose their own defaults. In practice, these defaults are typically a group's most likely parameter values given their own expert judgement. Finally, where available, we have also requested probabilistic output i.e. output which quantifies the probable range of a number of output variables rather than a single timeseries for each output variable (see section 1).

3.2 RCM drivers

Depending on the experiment in RCMIP, the ~~CMIP6 (?) endorsed intercomparison projects that are designed for Earth System Models. However, RCMIP replicates the experimental design of many of the CMIP-endorsed MIPs, particularly the DECK simulations (?), ScenarioMIP (?), AerChemMIP (?) and others,~~ drivers of the RCMs will vary e.g. ~~ZECMIP (?), DAMIP (?) or PMIP4 (?).~~

~~RCMIP builds on previous efforts to compare and understand RCMs. In 1997, the IPCC Technical Paper (?) investigated simple climate models used in the IPCC Second Assessment Report and compared their performance against idealised AOGCM results. compared RCMs used in integrated assessment models (IAMs). They focussed on five~~ the RCMs might run with prescribed CO_2 ~~-only experiments to quantify the differences in the behaviour of each IAM's climate component (each of which is an RCM due to computational constraints). ? extended the work of ? to consider the impact of non-~~concentrations and calculate consistent CO_2 ~~climate drivers in~~ emissions or the opposite i.e. run with prescribed CO_2 emissions and calculate consistent CO_2 concentrations. Below we describe each of the different setups used in RCMIP. However, a model did not need to be able to run in all of these ways to participate in RCMIP Phase 1.

3.2.1 Concentration driven

The concentration driven setup can strictly better be described as 'well-mixed greenhouse gas concentration' driven. Here, 'well-mixed greenhouse gases' refers to CO_2, CH_4, N_2O, hydrofluorocarbons (HFCs), perfluorocarbons (PFCs) and hydrochlorofluorocarb (HCFCs). Depending on the ~~RCPs. Recently, ? proposed a series of impulse tests for simple climate models in order to isolate differences in model behaviour under idealised conditions. RCMIP expands on the scope of previous work to include a broader range of scenarios and to make the first systematic comparison with both observations and CMIP model output~~experiment, these simulations are also supplemented by aerosol emissions and natural effective radiative forcing (specifically solar and volcanic forcings). For models which do not include the aerosol emissions to effective radiative forcing step, prescribed aerosol effective radiative forcing can instead be used.

This setup mirrors the majority of experiments performed in CMIP5 and CMIP6 such as the historical, RCP/SSP-based scenario and one percent per year rise in atmospheric CO_2 concentration (1pctCO2) experiments. The key difference between the RCMIP experiments and the CMIP experiments is that some RCMs include more anthropogenic drivers than CMIP models. Specifically, CMIP models do not include the full range of HFC, PFC and HCFC species, instead using equivalent concentrations (??). In addition, some CMIP models will not include the effect of aerosol precursors such as nitrates, ammonia and organic carbon (?).

3.3 ~~Science questions~~

3.2.1 CO_2 emissions driven

~~Over the course of its lifetime, RCMIP intends to answer the following scientific questions. This Phase 1 paper sets out the protocols required to do so. However, given the scope of the questions, they cannot all be answered in a single paper and so some aspects will receive less attention here than others~~In the CO_2 emissions driven setup CO_2 emissions are amended with concentrations of non-CO_2 well-mixed greenhouse gases. Like the concentration-driven setup, these simulations are also supplemented by aerosol emissions (or aerosol effective radiative forcing) and natural effective radiative forcings.

This setup mirrors the CO_2 emissions driven experiments performed in CMIP5 and CMIP6 such as the esm-hist, esm-ssp/rcp and esm-1pctCO2 experiments. As above, a cause of difference between CMIP and RCMIP simulations is the number of climate drivers that are explicitly modelled.

~~How do existing RCMs vary in their response to changes in~~

3.2.2 Emissions driven

The emissions driven or rather 'well-mixed greenhouse gas ~~(WMGHG) emissions, WMGHG concentrations, anthropogenic aerosol precursor emissions~~ emissions' driven setup is, like the concentration-driven and ~~natural changes in effective radiative forcing? To what extent can RCMs emulate the response of more complex models from CMIP5 and CMIP6? How do probabilistic ensembles from RCMs compare to each other and observations and what can they tell us about projected warming uncertainty under different scenarios ?~~ CO_2 emissions driven setups, supplemented by aerosol emissions (or aerosol effective radiative forcing) and natural effective radiative forcings.

These experiments have no obvious equivalent within the CMIP protocol. However, for many climate policy applications they are the most relevant set of experiments, given that anthropogenic emissions and reduction targets are what climate policy is directly concerned with (rather than atmospheric concentrations of GHGs). In addition, these experiments are of particular interest to the Integrated Assessment Modelling Consortium (IAMC) community and their contribution in IPCC WGIII because they require climate assessment of socioeconomic scenarios that are described in terms of their corresponding emissions, not concentrations.

4 ~~Methods~~

3.1 Experimental design

~~Phase 1 of RCMIP~~RCMIP's experimental design focuses on a limited set of the CMIP6 experiment protocol ~~plus a few experiments from~~ (**?**) plus some CMIP5 ~~. On top of the experiments from CMIP6 and CMIP5 we also add other experiments which are~~ experiments (**?**). We then complement this CMIP-based set with other experiments of interest to the ~~community (a full list is available~~ RCM and IAMC communities.

Systematic intercomparison projects such as RCMIP require the definition of a clear input and output data handling framework (see Section 4 for output specifications). Historically, comparing RCMs required learning how to set up, configure and run multiple RCMs in order to produce results. This required significant time and hence, as previously discussed, has only

been attempted in standalone cases with a limited number of models (????). With a common framework, once a model has participated in RCMIP, it is simpler to run it again in different experiments and provide output in a common, standardised format. This allows researchers to design, run and analyse experiments with far less effort than was previously required. As a result, it becomes feasible to do more regular and targeted assessment of RCMs. This capacity improves our knowledge of RCMs, our understanding of the implications of their quantitative results and our ability to develop and improve them.

Our input protocol is designed to be easy to use and hence easily able to be extended within future RCMIP phases or in separate research. The full set of RCMIP experiments is described in Supplementary Table ??)and available at rcmip.org. ~~The first class of these are the 'esm-X-allGHG' experiments. These runs are driven by emissions of all greenhouse gases (, , HFCs, PFCs etc.), rather than only~~

3.1.1 Input format

All input data is provided in a text-based format based on the specifications used by the IAMC community (?). The computational simplicity of RCMs means that their input specifications are relatively lightweight and hence using an uncompressed, text-based input format is possible. Further, the format is explicit about associated metadata and ensures metadata remains attached to the timeseries. As the IAMC community is a major user of RCMs, as well as being the source of input data for many experiments run with RCMs, using their data format ensures that data can be shared easily and assessment of IAM emissions scenarios can be performed with minimal data handling overhead.

The inputs are formatted as text files with comma separated values (CSV), with each row of the CSV file being a timeseries (see rcmip.org). This format is also often referred to as 'wide' although this term is imprecise (?). The columns provide metadata about the timeseries, specifically the timeseries' variable, units, region, model and scenario. Other columns provide the values for each timestep within the timeseries.

Being simplified models, RCMs typically do not take gridded input. Hence we use a selection of highly aggregated socio-economic regions, which once again follow IAMC conventions (?). RCMIP's variables and units are described in Section 4.1. The regions used in RCMIP are described in Table ??. Scenarios are discussed in section 3.1.3 and summarised in Table ??.

One complication of using the IAMC format is that the 'model' column is reserved for the name of the integrated assessment model which produced the scenario. To enhance compatibility with the IAMC format, we don't use the 'model' column. Instead, as described in Section 4, we use the separate 'climate_model' column to store metadata about the climate model which provided the timeseries.

In general, we follow the naming conventions provided by the CMIP6 protocol (?). These typically specify CO_2 ~~as is typical for ESMs in CMIP6. These experiments are particularly useful to the WGIII community as they perform climate assessment based on emissions scenarios, hence need models which can run from emissions and do not require exogenous concentrations of greenhouse gases.~~ emissions driven runs by prefixing the scenario name with 'esm-', with all other scenarios being concentration-driven. Where it is not possible to follow CMIP6 naming schemes, we use our own custom conventions. For example, full greenhouse gas emissions driven runs are typically not performed in CMIP6 because of computational cost. RCMIP's convention is to denote all greenhouse gas emissions driven by prefixing the scenario name with 'esm-' as well as

suffixing the name with '-allGHG' (e.g. 'esm-ssp245-allGHG'). In addition, RCMIP includes a number of CMIP5 experiments, which sometimes have the same name as their CMIP6 counterpart (e.g. 'historical'). Where such a clash exists, we append the CMIP5 experiment with '-cmip5' to distinguish the two (e.g. 'historical-cmip5'). Finally, if an experiment is not a CMIP6-style experiment then we cannot use a CMIP6 name for it. In such cases, we choose our own name and describe it within Table **??**.

~~We also add one extra experiment, ssp370-lowNTCF-gidden, onto the ssp370-lowNTCF experiment from AerChemMIP. The ssp370-lowNTCF experiment explicitly excludes a reduction in methane concentrations.However, the ssp370-lowNTCF emissions dataset as described in **?** and calculated in **?** does include reduced methane emissions~~

3.1.2 Idealised experiments

The first group of experiments in RCMIP is idealised experiments. They focus on examining model response in highly idealised experiments. These experiments provide an easy point of comparison with output from other models, particularly CMIP output, as well as information about basic model behaviour and dynamics which can be useful for understanding the differences between models.

RCMIP's Tier 1 idealised experiments are: piControl, esm-piControl, 1pctCO2, 1pctCO2-4xext, abrupt-4xCO2, abrupt-2xCO2 and ~~hence atmospheric concentrations. We include a~~ abrupt-0p5xCO2 (Table **??**). The piControl and esm-piControl control experiments serve as a useful check of model type. Most RCMs are perturbation models and hence do not include any internal variability, so will simply return constant values in their control experiments. Deviations from constant values in the control experiments quickly reveals those models with more complexity. Apart from esm-piControl, all of the Tier 1 experiments are concentration driven.

After the control experiments, the other Tier 1 experiments examine the models' responses to idealised, CO_2-only concentration changes. They reveal differences in model response to forcing, particularly whether the RCM response to forcing includes non-linearities. In addition, these experiments also provide a direct comparison with CMIP experiments (i.e. more complex model behaviour) and are a key benchmark when examining an RCM's ability to emulate more complex models.

The idealised Tier 2 experiments add idealised CO_2 removal experiments, which complement the typically rising/abruptly changing Tier 1 experiments. Idealised Tier 3 experiments examine the carbon cycle response in more detail with idealised emissions driven experiments as well as experiments in which the carbon cycle is only coupled to the climate system radiatively or biogeochemically (the '1pctCO2-rad' and '~~ssp370-lowNTCF-gidden' scenario to complement ssp370-lowNTCF and examine the consequences of a strong reduction in methane emissions.~~ 1pctCO2-bgc' experiments (**?**)). In concentration-driven experiments, RCMs report emissions (often referred to as 'inverse emissions') and carbon cycle behaviour consistent with the prescribed CO_2 pathway. For brevity, we do not go through all Tier 2 and 3 experiments in detail here, further information can be found in Table **??**.

~~The standard set of inputs from~~

3.1.3 Scenario experiments

In addition to the idealised experiments, RCMIP also includes a number of scenario based experiments. These examine model responses to historical transient forcing as well as a range of future scenarios. The historical experiments provide a way to compare RCM output against observational data records, and are complementary to the idealised experiments which provide a cleaner assessment of model response to forcing. The future scenarios probe RCM responses to a range of possible climate futures, both continued warming as well as stabilisation or overshoots in forcing. The variety of scenarios is a key test of model behaviour, evaluating them over a range of conditions rather than only over the historical period. Direct comparison with CMIP output then provides information about the extent to which the simplifications involved in RCM modelling are able to reproduce the response of our most advanced, physically-based models.

RCMIP's Tier 1 scenario experiments are: historical, ssp119, ssp585, esm-hist, esm-ssp119, esm-ssp585, esm-hist-allGHG, esm-ssp119-allGHG and esm-ssp585-allGHG. We focus on simulations (historical plus future) which cover the highest forcing (ssp585) and lowest forcing (ssp119) scenarios from the CMIP6 ScenarioMIP exercise (**?**). These quickly reveal differences in model projections over the widest available scenario range which can also be compared to CMIP6 output.

The Tier 2 experiments expand the CMIP6 scenario set to include the full range of ScenarioMIP concentration-driven experiments (**?**), which examine scenarios between the two extremes of ssp585 and ssp119, as well as the CMIP5 historical experiments. The CMIP5 experiments are particularly useful as they provide a direct comparison between CMIP5 and CMIP6~~was translated into the RCMIP experiment protocol, using the WGIII format (?), so it could be used by the modelling groups~~, something which has only been done to a limited extent with more complex models (**?**). Finally, the Tier 3 experiments add the remaining emissions-driven ScenarioMIP experiments, the rest of the CMIP5 scenario experiments (the so-called 'RCPs') and detection and attribution experiments (**?**) designed to examine the response to specific climate forcers over both the historical period and under a middle of the road emissions scenario (ssp245).

3.1.4 Data sources

CMIP6 emissions projections follow **?** and are ~~taken from~~ available at https://tntcat.iiasa.ac.at/SspDb/dsd?Action=htmlpage&page=60 ~~,~~ (hosted by IIASA~~. Where WMGHG~~). Where well-mixed greenhouse gas emissions are missing, we use inverse emissions based on the CMIP6 concentrations from MAGICC7.0.0 (**?**). Where regional emissions information is missing, we use the downscaling procedure described in **?**. The emissions extensions also follow the convention described in **?**.

For CMIP6 historical emissions (year 1850-2014), we have used data sources which match the harmonisation used for the CMIP6 emissions projections. This ensures consistency with CMIP6, ~~albeit at the expense of using the~~ although it means that we do not always use the latest data sources~~in some cases~~. CMIP6 historical anthropogenic emissions for CO_2, CH_4, BC, CO, NH_3, NOx, OC, SO_2 and non-methane volatile organic compounds (~~VOCs~~NMVOCs) come from CEDS (**?**). Biomass burning emissions data for CH_4, BC, CO, NH_3, NOx, OC, SO_2 and ~~volatile organic compounds (VOCs)~~ NMVOCs come from UVA (**?**). The biomass burning emissions are a blend of both anthropogenic and natural emissions, which could lead to some inconsistency between RCMs as they make different assumptions about the particular anthropogenic/natural emissions split. CO_2 global land-use emissions are taken from the Global Carbon Budget 2016 (**?**). ~~Other CMIP6 historical~~ Emissions of N_2O and the regional breakdown of CO_2 land-use emissions come from PRIMAP-hist ~~(?)~~ Version 1.0 ~~(and land-use regional~~

~~information)~~(**?**, see https://doi.org/10.5880/PIK.2016.003). Where required, historical emissions were extended back to 1750 by assuming a constant relative rate of decline based on the period 1850-1860 ~~. While this means that our~~ (noting that historical emissions are ~~highly uncertain, all we require is~~ somewhat uncertain, we require consistent emissions inputs ~~so we leave improved quantifications of emissions in this period for other research~~in Phase 1, uncertainty in historical emissions will be explored in future research).

CMIP6 concentrations follow **?**. CMIP6 radiative forcings follow the data provided at https://doi.org/10.5281/zenodo.3515339). CMIP5 emissions, concentrations and radiative forcings follow **?** and are taken from http://www.pik-potsdam.de/~mmalte/rcps/.

3.2 ~~Diagnostics~~

4 Output specifications

~~Given their focus on global-mean, annual mean variables we request a range of output variables from each RCM (detailed in Supplementary Table **??**) . The output variables focus on the response of the climate system to radiative forcing (e. g. surface air temperature change, surface ocean temperature change, effective radiative forcing, effective climate sensitivity) as well as the carbon cycle (carbon pool sizes, fluxes between the pools, fluxes due to Earth System feedbacks). The~~ RCMIP Phase 1's submission template (see rcmip.org or https://doi.org/10.5281/zenodo.3593570) is composed of two parts. The first part is the data submission and is identical to the input format (see Section 3.1.1). This allows for simplified analysis with the same tools we used to develop the input protocols and exchange with the IAMC community as they can analyse the data using existing tools such as pyam (**?**). The second part is model metadata. This includes the model's name, version number, brief description, literature reference and other diagnostics (see Section 4.1). We also request a configuration label, which uniquely identifies the configuration in which the model was run to produce the given results.

Given the typical temporal resolution of ~~the models means that we request the data on~~ RCMs, we request all output be reported with an annual timestep~~but may increase the temporal resolution in Phase 2.~~

~~The output dataset represents a huge data resource. In this paper we focus on a limited set of variables, particularly related to the climate response to radiative forcing. The dataset extends well beyond this limited scope and should be investigated in further research. To facilitate such research, we have put the entire database~~ . In addition, to facilitate use of the output, participating modelling groups agree to have their submitted data made available under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license. All input and output data, as well as all code required to produce this paper, is available at gitlab.com/rcmip/rcmip and archived at https://doi.org/10.5281/zenodo.3593569.

4.1 Variables

5 ~~Results and Discussion~~

RCMIP has a large variable request (26 Tier 1 variables, 344 Tier 2 variables and 13 Tier 3 variables), reflecting the large number of climate components included in RCMs. Here we discuss the Tier 1 variables. Tier 2 and 3 variables, which go into more detail for various parts of the climate system, are described in Supplementary Table **??**.

RCMs agree moderately well with historical observations of global-mean surface air temperature (GSAT)(Figure 1). For the period 2000-2019, the RCMs project warming of 0.94with a standard deviation of 0.09relative to a reference period of 1850-1900 and a warming rate of 0.24/ decade with a standard deviation of 0.04/ decade. These projections agree with best-guess warming observations of 0.99but no RCM agrees with the observed warmingrate of 0.19/ decade. With the exception of the GREB model (which is a The Tier 1 variables focus on key steps in the cause-effect chain from emissions to warming. We request emissions of black carbon, CH_4, carbon monoxide, CO_2 -only model), all the RCMs demonstrate an ability to reproduce short-term cooling due to major volcanic eruptions. , N_2O, NH_3, nitrous oxides, organic carbon, sulphates and non-methane volatile organic compounds. These cover the major greenhouse gases plus aerosol pre-cursor emissions. In the case of emissions driven runs, these emissions are prescribed hence we only request that these variables are reported as outputs where the modelling groups have had to alter them (e.g. their model includes internal land-use calculations which cannot be exogenously overridden). In the case of concentration-driven runs, we request emissions compatible with the prescribed concentration pathway (where these can be derived). We also request cumulative emissions of CO_2 given their strong relationship with peak warming (**????**).

The RCMs do not exhibit the same high-frequency modes as observations due to their lack of internal variability, particularly representations of phenomena like El Nino, the Interdecadal Pacific Oscillation or Atlantic Meridional Variability. This lack of natural variability may explain why RCMs (with the exception of GREB which is a In Tier 1, we only request atmospheric concentrations of CO_2 -only model) appear to be too cool through the 1960's, 70's and 80's and then warm too quickly thereafter. Alternately, it could be that RCMs overestimate aerosol cooling over this period (although the spread in aerosol and CH_4. Many models are capable of reporting much more detail than this, and we encourage them to report this detail, however some models only focus on a limited set of concentrations hence we restrict our Tier 1 variables.

In addition to concentrations, we request total, anthropogenic, CO_2 and aerosol effective radiative forcing estimates across models is nonetheless fairly large, see Supplementary Figure **??**) or overestimate of the impact of the eruption of Mt Agung in 1963. and radiative forcing. These forcing variables are key indicators of the long-term drivers of climate change within each model as well as being a key metric for the IAMC community. Effective radiative forcing and radiative forcing are defined following **?**. In contrast to radiative forcing, effective radiative forcing includes rapid adjustments beyond stratospheric temperature adjustments thus is a better indicator of long-term climate change.

This suggests a clear area for further evaluation of RCMs and has important implications for remaining carbon budget estimates, which are highly sensitive to estimates of recent warming trends (**?**). Discrepancies in the carbon cycle response to emissions also impact remaining carbon budget estimates. Finally in Tier 1, we request output of total climate system heat uptake, ocean heat uptake, surface air temperature change and surface ocean temperature change. These variables are most directly comparable to available observations and CMIP output, with surface temperature also being highly policy-relevant.

Focusing on these key variables allows us to discern major differences between RCMs, with Tier 2 and ~~further analysis on~~ 3 variables then providing further points of comparison at a finer level of detail.

4.0.1 Probabilistic outputs

To reduce the total data volume, we request that groups provide only a limited set of percentiles from reporting probabilistic outputs, rather than every run which makes up the probabilistic ensemble. The 10th, 50th (median) and 90th percentiles are Tier 1, with the 5th, 17th, 33rd, 67th, 83rd and 95th percentiles being Tier 2. When calculating these percentiles, groups must take care to calculate derived quantities (e.g. Effective Climate Sensitivity) from each run in the probabilistic ensemble first and then calculate the percentiles in a second step. Doing the reverse (calculating percentiles first, then derived quantities from percentiles) will not necessarily lead to the same answer.

4.1 Diagnostics

On top of the variable request, we ask for one other diagnostic. This is the equilibrium climate sensitivity, defined as 'the equilibrium warming following an instantaneous doubling of atmospheric CO_2 concentrations'. Unlike more complex models, RCMs typically have analytically tractable equilibrium climate sensitivities. This means we do not need to include ten thousand year long simulations, which would allow the models to reach true equilibrium. In contrast to the equilibrium climate sensitivity, the ~~all-greenhouse gas emissions driven runs would highlight further model differences~~ more commonly used effective climate sensitivity, derived using the Gregory method (**?**), underestimates warming at true equilibrium in many models (**?**).

5 Illustrative results

15 models have participated in RCMIP Phase 1 (see Table 1 for an overview and links to key description papers). This is a promising start, demonstrating that the protocol is accessible to a wide range of modelling teams. We encourage any other interested groups to join further phases of the project.

~~The discrepancies between RCMsare of~~ The groups which have participated have submitted a number of results. We provide a brief overview of these here to give an initial assessment of the diversity of models which have submitted results to date. However, this is not intended as a comprehensive comparison or evaluation.

Firstly, we present a comparison of model best-estimates against observational best estimates (Figure 1). Such comparisons are a natural starting point for evaluation of all RCMs. We see that all the RCMs are able to capture the approximately 1 °C of warming seen in the historical observations compared to a ~~similar order of magnitude to observational uncertainties (**?**) . The spread of RCMs is also much smaller than the spread in available CMIP6 model results (warming of $1.13 \pm 0.3$and warming rate of 0.24 ± 0.08/ decade when only the first available ensemble member from each model group is used) . Given their simple, tuneable nature, it is no surprise that RCMs tend to agree more closely with observations than CMIP modelsand exhibit less spread.~~ pre-industrial reference period (**??**). We also see that all the RCMs include some representation of the impact of volcanic eruptions, most notably the drop in global-mean temperatures after the eruption of Mount Agung in 1963.

The exception is the CO_2-only model, GREB, which lacks the volcanic and aerosol induced cooling signals of the 19th and 20th Centuries.

~~In general, RCMs can emulate CMIP6 surface air temperature change relatively well (Figure 3). Given that RCMs do not include internal variability,~~ Another way to evaluate RCMs is to compare their probabilistic results to observational best estimates as well as uncertainties (Figure 2). Such comparisons are vital to understanding the limits of projected probabilistic ranges and their dependence on model structure. Here we see large differences in probabilistic projections despite the similarities in the models' historical simulations. Determining the underlying causes of such differences requires investigation into and understanding of how the probabilistic distributions are created.

RCMIP also facilitates a comparison of model calibrations and CMIP output (Figure 3). Each RCM is calibrated to a different number of CMIP models (some RCMs provide no calibrations at all) because there is no common resource of calibration data. Instead, the ~~lower bound for~~ CMIP models to which each RCM is calibrated depends on each RCM development team's capability and the time at which they last accessed the CMIP archives.

Examining multiple emulation setups (Figures **??** - **??**), RCMs can reproduce the temperature response of CMIP models to idealised forcing changes to within a root-mean ~~square error (RMSE) is of the same order of magnitude as internal variability in CMIP6 models . The best emulators are pushing this limit, with RMSE on the order of~~ square error of 0.2~~K (Table 2). As~~ °C (Table 2). In scenario-based experiments, it appears to be harder for RCMs to emulate CMIP output than in idealised experiments. We suggest two key explanations. The first is that effective radiative forcing cannot be easily diagnosed in SSP scenarios hence it is hard to know how best to force the RCM during calibration. The second is that the forcing in these scenarios includes periods of increase, sudden decrease due to volcanoes as well as longer term stabilisation rather than the simpler changes seen in the idealised experiments. Fitting all three of these regimes is a more difficult challenge than fitting the the idealised experiments alone.

We also present plots of the relationship between surface air temperature change and cumulative CO_2 emissions from the 1pctCO2 and 1pctCO2-4xext experiments (Figure 4). These can be used to derive the transient climate response to emissions (**?**), a key metric in the calculation of our remaining carbon budget (**?**). The illustrative results here demonstrate a range of relationships between these two key variables, from weakly sub-linear to weakly super-linear (see further discussion in **?**).

Finally, we present initial results from running both CMIP5 and CMIP6 ~~results have only recently become available, we expect further calibration efforts to reduce RMSE even further .~~ generation scenarios ('RCP' and 'SSP-based' scenarios respectively) with the same models (Figure 5). In the small selection of models which have submitted all RCP, SSP-based scenario pairs, the SSP-based scenarios are 0.21°C (standard deviation 0.10°C across the models' default setups) warmer than their corresponding RCPs (Figure 5(b)). This difference is driven by the 0.42 $\pm$0.26 Wm^{-2} larger effective radiative forcing in the SSP-based scenarios (Figure 5(d)), which itself is driven by the larger CO_2 effective radiative forcing in the SSP-based scenarios (Figure 5(f)). As noted previously, these are only initial results, not a comprehensive evaluation and should be treated as such. Nonetheless, they agree with other work (**?**) which suggests that even when run with the same model (in a concentration-driven setup), the SSP-based scenarios result in (non-trivially) warmer projections than the RCPs.

Despite their relatively good performance, resultsfor the different emulation setups have generally only been submitted for a limited set of scenarios (

6 Extensions

RCMIP Phase 1 provides proof of concept of the RCMIP approach to RCM evaluation, comparison and examination. The RCMIP Phase 1 protocol focuses on model evaluation hence is limited to experiments which are directly comparable to observations and CMIP output. In this section we present a number of ways in which further research and phases of RCMIP could build on the work presented in this paper.

The first is a deeper evaluation of the results submitted to RCMIP Phase 1. Here we have only presented illustrative results, however these can be evaluated and investigated in far more detail. For example, quantifying the degree to which different RCMs agree with observations, carefully considering how to handle observational uncertainties, natural variability (which many RCMs cannot capture) and model tuning.

Secondly, there is a wide range of RCMs available in the literature. This variety can be confusing, especially to those who are not intimately involved in developing the models. An overview of the different models, their structure and relationship to one another would help reduce the confusion and provide clarity about the implications of using one model over another.

The third suggested extension is an investigation into how different RCMs reach equilibrium in response to a step change in forcing. In RCMIP Phase 1, we only specified the equilibrium climate sensitivity value but temperature response is potentially further defined by linear and nonlinear feedbacks on different timescales. Further phases could investigate whether model structure is a driver of difference between model output or whether these differences are largely controlled by differences in parameter values.

Fourthly, emulation results have generally only been submitted for a limited set of experiments (see Supplementary Table **??** and Supplementary Figures **??** - **??**). **??**). Hence it is still not clear whether the good performance in idealised scenarios also carries over to projections, particularly for the SSPs. Having said this, results for MAGICC7.1.0, which has supplied projections for the SSPs for each emulation setup, are promising. MAGICC7.1.0's results suggest that RCMs should be close to the lower limit of RMSE as more emulation performance seen in idealised experiments also carries over to scenarios, particularly the SSP-based scenarios. As the number of available CMIP6 results become available and further calibration efforts are carried out.

From the available results, differences emerge between modelswith constant effective climate sensitivities and models with time or state dependent effective climate sensitivities. Models with constant effective climate sensitivities, such as the AR5IR implementations, struggle to capture the non-linear response of ESMs to abrupt changes in concentrations. Firstly, they predict an equally large response to negative radiative forcing as positive radiative forcing which isn't always the case in ESMs (Panels (d) and (e) of Figure 3). Secondly, in order to capture the long-term warming seen in many abrupt-4xCO2 experiments, one of the boxes often has a response timescale on the order of thousands of years. This is problematic because it leads to equilibration times on the order of thousands of years and large equilibrium responses in the abrupt-2xCO2 experiments (i.

e. large equilibrium climate sensitivities, see Supplementary Table ??). Models with time or state dependent effective climate sensitivities avoid both those problems. In particular, their temperature response to positive and negative radiative forcing need not be of equal magnitude and they can exhibit the long warming tail seen in ESM abrupt-4xCO2 runs whilst avoiding extremely long equilibration times and large equilibrium temperature perturbations in abrupt-2xCO2 runs.

Probabilistic projections from RCMs illustrate the large range of plausible temperature projections resulting from physical parameter sets which are consistent with observations (Figure 2). For example, under the very ambitious mitigation scenario ssp119, the models presented here have a best estimate of approximately 1.5for end of century warming. However, they also suggest that there is still approximately a 1 in 6 chance that warming would exceed 2.

These probabilistic projections extend the results of CMIP6, which do not include such large perturbed parameter ensemble plus constraining exercises. The 66% ranges presented here are, in general, significantly narrower than the results continues to grow, this area is ripe for investigation and will lead to improved understanding of the limits of the reduced complexity approach. A common resource for RCM calibration would greatly aid this effort because CMIP6 intermodel spread. There is no requirement that CMIP6 results lie within some range of historically observed temperature changes but the difference suggests that some caution should be used when inferring projection uncertainty from CMIP6 results alone. data handling requires specialist big data handling skills.

Four of the models (MCE, WASP, FaIR and OSCAR) provide remarkably similar median projections. On the other hand, Hector projects significantly smaller surface air temperature increases, likely due to its lower radiative forcing estimates (Figure ??). Fifthly, while RCMIP Phase 1 allows us to evaluate the differences between RCMs, the root causes of these differences may not be clear. This can be addressed by extending RCMIP to include experiments which specifically diagnose the reasons for differences between models e.g. simple pulse emissions of different species or prescribed step changes in atmospheric greenhouse gas concentrations. Such experiments could build on existing research (??) and would allow even more comprehensive examination and understanding of RCM behaviour.

On the other hand, there is a surprising amount of variation in probabilistic simulations of the historical period. The variation in ranges, from MCE with relatively large ranges, to WASP and Hector with much smaller ranges, likely reflects differences in constraining techniques. Given that variations in both model structureFollowing this, there is clearly some variation in probabilistic projections. However, what is not yet known is the extent to which variations in model structure, calibration data and calibration technique influence probabilistic projections , an area for future research could be to try and disentangle the impact of these two components. Such an experiment drive such differences. Investigating these questions would help understand the limits of probabilistic projections and their uncertainties. Experiments could involve constraining two different models with the same constraining technique or and data, constraining a single model with two different techniques but the same data or constraining a single model with a single technique but two different datasets.

One other area for Next, the current experiments can be extended to examine the behaviour of models' gas cycles, particularly their interactions and feedbacks with other components of the climate system. This will require custom experiments but is important for understanding the behaviour of these emissions driven runs. Such experiments are particularly important for the carbon cycle, which is strongly coupled to other parts of the climate system. It should be noted that, for ESMs, the suggestion

of extra experiments is limited by human and computational constraints. This constraint does not apply to RCMs because of their computational efficiency: adding extra RCM experiments adds relatively little technical burden.

One final suggestion for future research is the ~~impact of the~~ importance of the choice of reference period. Within the reference period, all model results and observations will be artificially brought together, narrowing uncertainty and disagreement within this period (**?**). This can alter conclusions as the reference period will become less important for any fitting algorithm (because of the artificial agreement), placing more weight on other periods. Developing a method to rebase both the mean and variance of model and observational results onto other reference periods would allow the impact of the reference period choice to be explored in a more systematic fashion.

~~Looking forward, it is clear that making probabilistic projections consistent with CMIP6 results requires structural model flexibility, such that models are shown to be able to reproduce CMIP6 results. Having achieved this, probabilistic parameter ensembles can then be derived while considering uncertainty in both and non-climate drivers. During RCMIP Phase 1, only Hector has been able to perform both these steps. However, we hope that more models will be able to perform these steps in further phases. Such results would enhance our understanding of the uncertainty in observationally consistent climate change projections and hence be of interest to the climate research community and beyond.~~

~~When run with the same model, warming projections are higher in the SSPs than the RCPs (Figure 5). Whilst historical warming estimates are very similar, if not slightly higher in the RCP-compatible historical runs, the scenarios separate over the course of the 21st Century.~~

7 Conclusions

~~For the RCMIP results, we can see that the increase in warming projections is due to the higher effective radiative forcing in the SSPs throughout the second half of the 21st Century (Supplementary Figure **??**). The higher effective radiative forcing results appears to be a result of~~ RCMs are used in many applications, particularly where computational constraints prevent other techniques from being used. Due to their importance in climate policy assessments, in carbon budget calculations, as well as applicability to a wide range of scientific questions understanding the behaviour and output from RCMs is highly relevant and requires continuous updating with the ~~SSPs agreeing more closely with their nameplate 2100 radiative forcing level than the RCPs, which were generally too low. The increased forcing is driven largely by increased effective radiative forcing (Supplementary Figure **??**), which itself is driven by increased emissions (**?**). Even though aerosol effective radiative forcing is also slightly more negative in the SSPs (Supplementary Figure **??**), the difference of approximately 0.1 is not enough to offset increased effective radiative forcing of approximately 0.5~~ latest science. Here we have presented the Reduced Complexity Model Intercomparison Project (RCMIP), an effort to facilitate the evaluation and understanding of RCMs in a systematic, standardised and detailed way. We hope this can greatly improve ease of use of and familiarity with RCMs.

~~At present, effective radiative forcings diagnosed from the CMIP6 models are not available (as such diagnosis is not a trivial task). However, given the monotonic relationship between concentrations and effective radiative forcing (**?**), it is likely that the same mechanisms are driving at least part of the increase between CMIP5 and CMIP6 projections.~~

Comparing warming projections between CMIP5 and CMIP6, our results suggest that around 46% of the increase is scenario driven. However, this still leaves 54% which is not explained by the change in scenarios. At this stage, this residual is most likely explained by a change in the models submitting results to CMIP, which appear to be more sensitive to changes in atmospheric GHG concentrations in CMIP6 than in CMIP5 (**???**). However, CMIP6 analysis is ongoing and should be considered before making strong conclusions about the robustness of these findings.

As discussed previously, the results from RCMIP can provide much more information than has been presented here. A number of experiments have not been discussed here which would shed light on the differences between the RCMs in a number of other components. In addition, RCMsalso offer the chance to explore more experiments than is planned in CMIP6 due to their computational efficiency. An experiment which is an example of both these points is the ssp370-lowNTCF scenario as quantified by **?**, which includes reductions in methane emissions. In contrast, the ssp370-lowNTCF as defined by AerChemMIP explicitly includes methane emissions reductions. RCMs can examine the impact of this difference by running an extra experiment, 'ssp370-lowNTCF-gidden', which follows the emissions quantified by **?**. Preliminary results are given in Supplementary Figure **??** and , unsurprisingly, show that surface air temperatures rise in ssp370-lowNTCF (relative to ssp370) whilst they fall in ssp370-lowNTCF-gidden. This fall in temperatures is driven entirely by reductions in methane emissions and users of CMIP6 data should be careful not to confuse the results of the ssp370-lowNTCF scenario with the emissions scenarios presented in **?**. They are two different experiments.

Phase We have performed RCMIP Phase 1of RCMIP has identified some of the strengths, weaknesses and limitations of various RCMs which exist in the literature today. This paper has focussed on surface air temperature (GSAT) changes but many more output variables are available in the Phase , which provides an initial database of experiments conducted with 15 participating models from the RCM community. RCMIP Phase 1 database. To facilitate further research, the entire database is available under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license.

We have found that RCMs are capable of reproducing broad-scale characteristics of observed historical GSAT changes as well as the response of ESMs under various experiments. Further work could focus on why RCMs exhibit relatively high recent warming rates compared to observations and using the ever growing body of CMIP6 results to improve RCM emulation capabilities. Nonetheless, there is clear evidence that the addition of time and state-dependent climate feedbacks in many RCMs has improved their ability to emulate the behaviour of more complex models under a range of forcing conditions.

Probabilistic projections from RCMs complement higher complexity model results by providing uncertainties which are by design consistent with historically observed temperature changes. Further evaluating these probabilistic distributions and the impact of different derivation techniques and model structures is a clear next step for the RCM community. Another next step is adding more models which are both calibrated to CMIP6 results and have probabilistic distributions as only the Hector model has managed this to date.

focused on basic evaluation and benchmarking of RCMs, providing some key starting points for all users of RCMs to examine when considering their model of choice. Here we have only presented illustrative results and further analysis is warranted to quantify the differences in behaviour (and associated uncertainty) between the different RCMs. Further work will examine the results from Phase 1 paves the way for further phases of RCMIP. Much of the work of defining community

~~standards, data handling practices and communication methods has been established and now only needs refining. Further phases of RCMIP could focus on many different themes, for example, considering a wider range of variables, probabilistic climate projections (something which cannot be done with more complex models due to computational expense), specific components of the earth system (e. g. ocean heat content, representation of aerosols, sea-level rise) or model development (e.g. adding new components to models). We would~~ and RCMs in more detail, improving evaluation, comparison and understanding of the implications of differences between models.

RCMIP aims to fill a gap in our understanding of RCM behaviour, in particular, in how different RCMs perform relative to each other as well as in absolute terms. This gap is particularly important to fill given the widespread use of RCMs throughout the integrated assessment modelling community and in large-scale climate science assessments. We welcome requests, suggestions and further involvement from throughout the climate modelling research community. With our efforts, we hope to increase understanding of and confidence in RCMs, particularly for their many users at the science-policy interface.

Code and data availability. RCMIP input timeseries and results data along with processing scripts as used in this submission are available from the RCMIP GitLab repository at https://gitlab.com/rcmip/rcmip and archived by Zenodo (https://doi.org/10.5281/zenodo.3593569).

The ACC2 model code is available upon request.

The implementation of the AR5IR model used in this study is available in the OpenSCM repository: https://github.com/openscm/openscm/blob/ar5ir-notebooks/notebooks/ar5ir_rcmip.ipynb

The model version of ESCIMO used to produce the RCMIP runs can be downloaded from http://www.2052.info/wp-content/uploads/2019/12/mo191107%202%20ESCIMO-rcimpfrom%20mo160911%202100%20ESCIMO.vpm. The vpm extension allows you to view, examine and run the model, but not save it. The original model with full documentation is available from http://www.2052.info/escimo/.

FaIR is developed on GitHub at https://github.com/OMS-NetZero/FAIR and v1.5 used in this study is archived at Zenodo (**?**).

The GREB model source code used is available, upon request, on Bitbucket: https://bitbucket.org/rcmipgreb/greb-official/src/official-rcmip/. The last stable versions are available on GitHub at https://github.com/christianstassen/greb-official/releases.

The Held two layer model implementation used in this study is available in the OpenSCM repository: https://github.com/openscm/openscm/blob/ar5ir-notebooks/notebooks/held_two_layer_rcmip.ipynb

Hector is developed on GitHub at https://github.com/JGCRI/hector. The exact version of Hector used for these simulations can be found at https://github.com/JGCRI/hector/releases/tag/rcmip-tier1. The scripts for the RCMIP runs are available at https://github.com/ashiklom/hector-rcmip.

MAGICC's Python wrapper is archived at Zenodo (https://doi.org/10.5281/zenodo.1111815) and developed on GitHub at https://github.com/openclimatedata/pymagicc/.

OSCAR v3 is available on GitHub at https://github.com/tgasser/OSCAR.

WASP's code for the version used in this study is available from the supplementary material of **?**: https://doi.org/10.1029/2018EF000889. See also the WASP website at http://www.waspclimatemodel.info/download-wasp.

The other participating models are not yet available publicly for download or as open source. Please also refer to their respective model description papers for notes and code availability.

Author contributions. ZN and RG conceived the idea for RCMIP. ZN, MM and JL setup the RCMIP website (rcmip.org), produced the first draft of the protocol and derived the data format. All authors contributed to updating and improving the protocol. ACC2 results were provided by KT and EK. AR5IR and Held et al. two layer model were provided by ZN. CICERO-SCM results were provided by JF, BS, MS and RS. ESCIMO results were provided by UG. FaIR results were provided by CS. GIR results were provided by NL. GREB results were provided by DD, CF, DM and ZX. Hector results were provided by AS and KD. MAGICC results were provided by MM, JL and ZN. MCE results were provided by JT. OSCAR results were provided by TG and YQ. WASP results were provided by PG. ZN wrote, except for the model descriptions, the first manuscript draft, produced all the figures and led the manuscript writing process with support from RG. All authors contributed to writing and revising the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We acknowledge the World Climate Research Program (WCRP) Coupled Model Intercomparison Project (CMIP) and thank the climate ~~modeling~~ modelling groups for producing and making available their model output. RCMIP could not go ahead without the outputs of CMIP6 nor without the huge effort which is put in by all the researchers involved in CMIP6 (some of whom are also involved in RCMIP).

We also thank the RCMIP Steering Committee, comprised of ~~Jan Fuglestvedt,~~ Maisa Corradi, ~~Malte Meinshausen,~~ Piers Forster, Jan Fuglestvedt, Malte Meinshausen, Joeri Rogelj and Steven Smith, for their support and guidance throught Phase 1. We look forward to their ongoing support in further phases.

ZN benefited from support provided by the ARC Centre of Excellence for Climate Extremes (CE170100023). RG acknowledges support by the German Federal Ministry for the Environment, Nature Conservation and Nuclear Safety (16_II_148_Global_A_IMPACT) while working at PIK in the beginning of RCMIP. KT is supported by the Integrated Research Program for Advancing Climate Models (TOUGOU Program), the Ministry of Education, Culture, Sports, Science, and Technology (MEXT), Japan.

Table 1. Overview of the physical components of the models participating in RCMIP Phase 1.

Model (acronym used in figures)	Spatial resolution	~~Temporal resolution~~Key references ~~Climate response to radiative forcing Other components~~
ACC2 (ACC2-v4-2)	Global land/ocean	~~Annual 1D ocean heat diffusion (DOECLIM) Land and ocean carbon cycle, ocean carbonate chemistry, parameterized atmospheric chemistry involving CH4, OH, NOx, CO, and VOC, radiative forcing of 29 halocarbons~~ ?? (also ???)
AR5IR (ar5ir-2box, ar5ir-3box)	Global	~~Annual Impulse response None~~ ?
CICERO-SCM (CICERO-SCM)	~~By hemisphere~~ Hemispheric	~~Annual~~ ? (also ????)
EMGC (EMGC)	~~Energy balance/upwelling diffusion model~~ Global	~~Land and ocean carbon cycle~~ ??
ESCIMO (ESCIMO)	Global	~~Annual Conserved flows of carbon, heat, albedo, permafrost, biome and biomass change. Driven by GHG emissions, the rest is endogenous. No complete water cycle, water is tracked as ocean, high and low clouds, ice (glacial, arctic, Greenland and Antarctic), and vapor.~~ ?
FaIR (FaIR-v1-5)	Global	~~Annual Modified impulse response Simple ozone, aerosol, greenhouse gas and land use relationships from precursor emissions~~ ??
GIR (GIR)	Global	~~Annual Modified Impulse Response Simple (typically state-dependent) ozone, aerosol and greenhouse gas relationships~~ ?
GREB (GREB-v1-0-1)	96 x 48 grid	~~Monthly Energy Balance model atmospheric transport of heat and moisture, surface and subsurface ocean layer. Hydrological cycle, sea ice.~~ ?
Hector (hectorl62381e71)	Global	~~Annual 1D ocean heat diffusion (DOECLIM) Land and ocean carbon cycle. Ocean carbonate chemistry and simplified thermohaline circulation. Atmospheric chemistry of CH4, OH, NOx, and halocarbons.~~ ??? (see also ??)

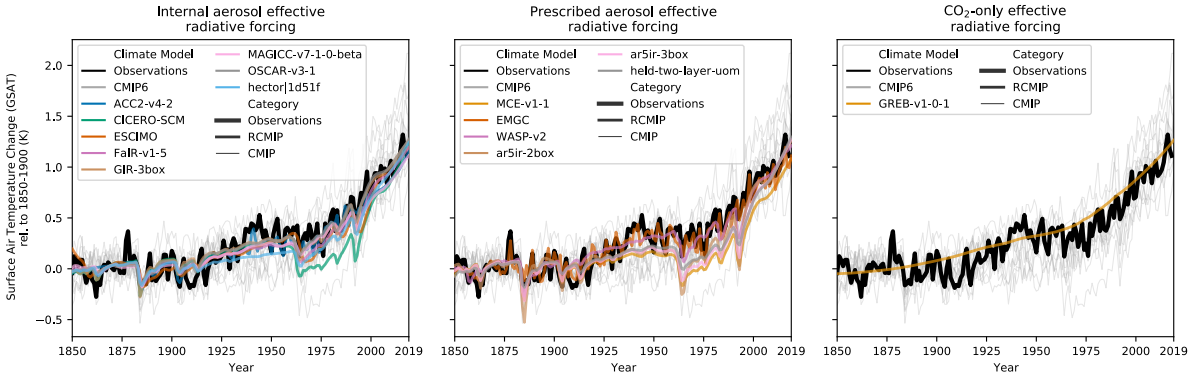


Figure 1. Historical global-mean annual mean surface air temperature (GSAT) simulations. Thick black line is observed GSAT (**??**). Medium thickness lines are illustrative configurations for RCMIP models. Thin grey solid lines are ~~CMIP models (~~CMIP6 ~~in dark blue, CMIP5 in grey)~~models. In order to provide timeseries up until 2019, we have used data from the combination of historical and ssp585 simulations for RCMIP and CMIP6 models and rcp85 data for CMIP5 models.

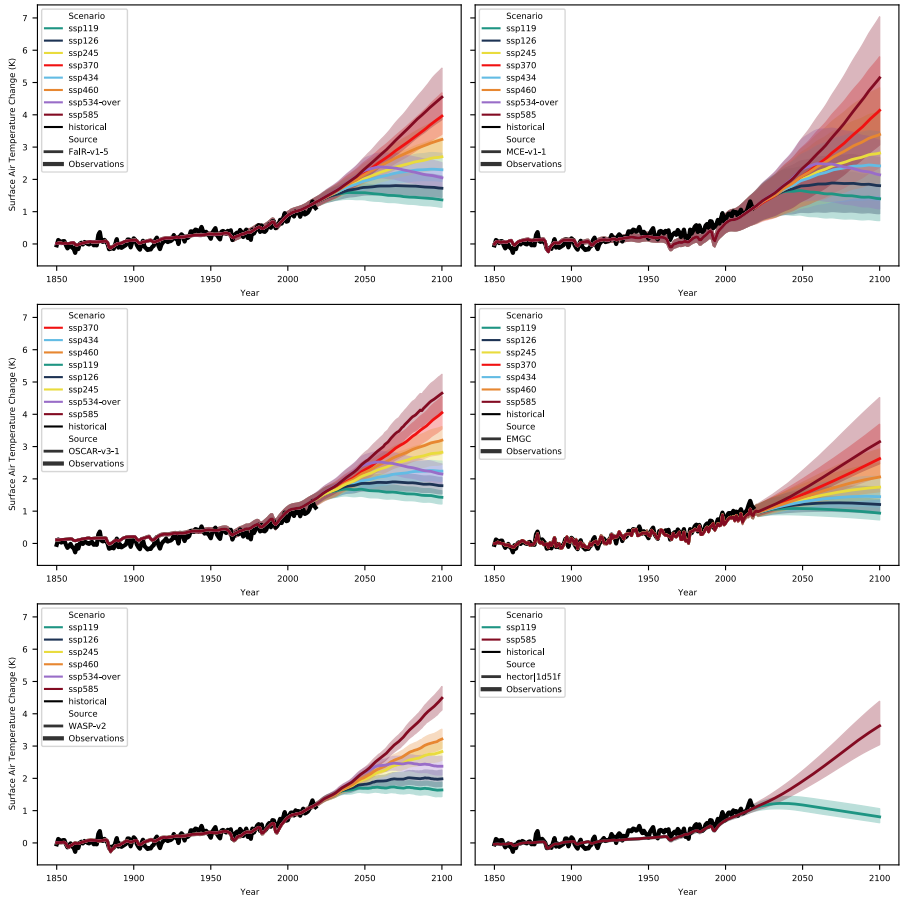


Figure 2. Probabilistic projections. Black line is observed GSAT (**??**). Coloured lines are results for different RCMs for the SSP-based scenarios (ranges are 66% ranges). Note that not all groups have been able to perform all simulations.

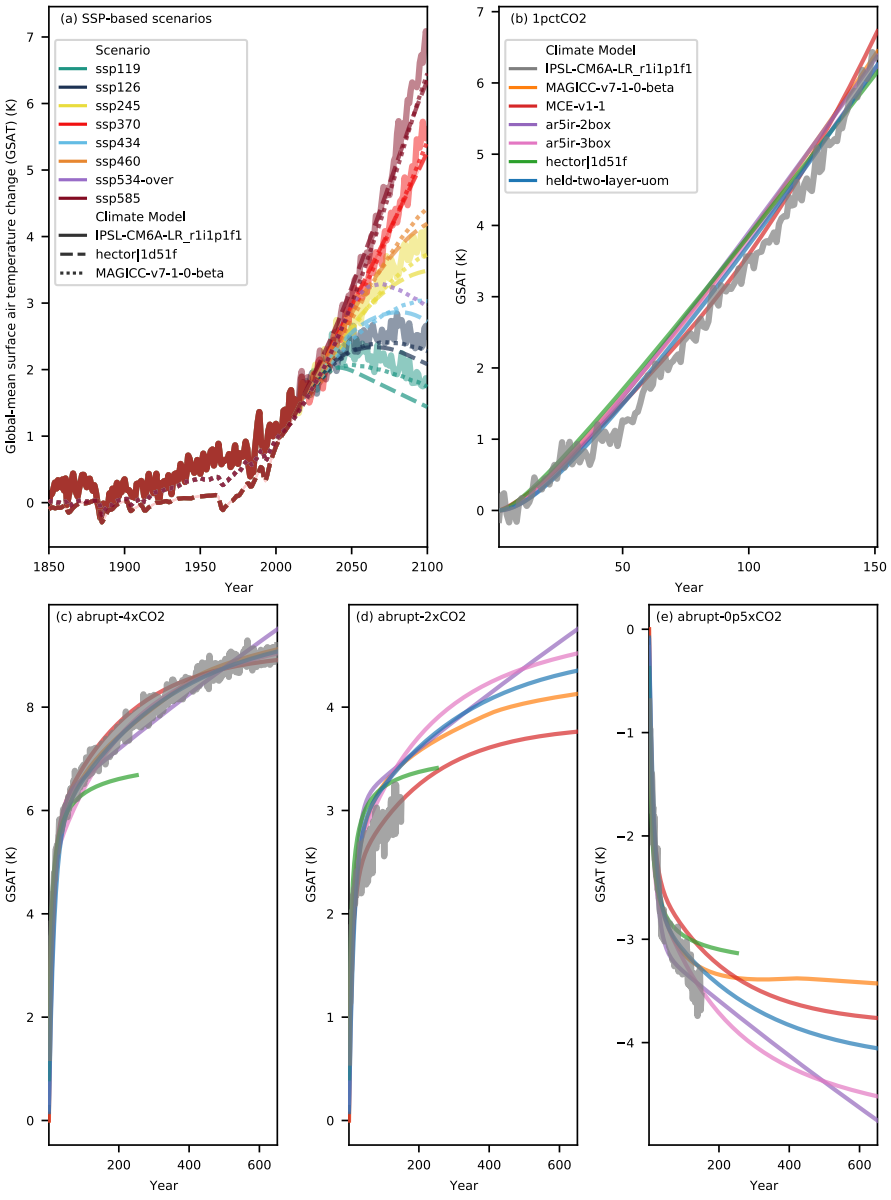
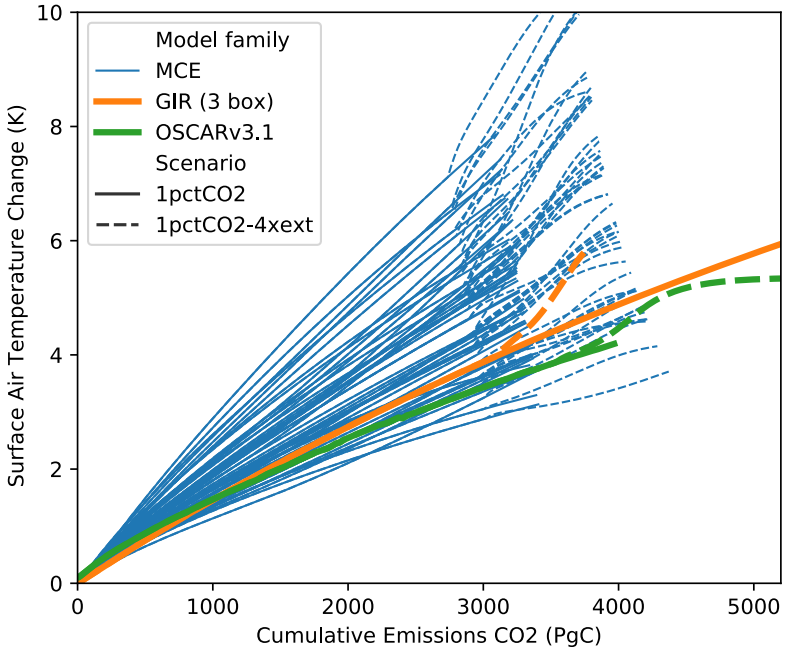


Figure 3. Emulation of CMIP6 models by RCMs. The thick transparent lines are the target CMIP6 model output (here from IPSL-CM6A-LR r1i1p1f1). The thin lines are emulations from different RCMs. Panel (a) shows results for scenario based experiments while panels (b) - (e) show results for idealised CO2-only experiments (note that panels (b) - (e) share the same legend). See the Supplementary Information for other target CMIP6 models.

Table 2. Model emulation scores over all emulated models and scenarios. Here we provide root-mean square errors over the SSPs plus four idealised CO2-only experiments (abrupt-2xCO2, abrupt-4xCO2, abrupt-0p5xCO2, 1pctCO2). As the models have not all provided emulations for the same set of models and scenarios, the model emulation scores are indicative only and are not a true, fair test of skill. For target model by target model emulation scores, see ~~TODO supplementary table reference~~Table **??**.

Model (number of emulated scenarios)	**Surface Air Temperature Change (GSAT ~~) i.e.~~ aka tas) root-mean square error (indicative only)**
~~hector|62381e71 (32~~MAGICC-v7-1-0-beta (131)	~~0.42~~ 0.21 K
MCE-v1-1 (44)	0.19 K
~~MAGICC-v7-1-0-beta (135~~ar5ir-2box (36)	~~0.21~~ 0.24 K
~~ar5ir-2box (40~~ar5ir-3box (36)	~~0.23~~ 0.28 K
~~ar5ir-3box (40~~hector|1d51f (64)	~~0.27~~ 0.28 K
held-two-layer-uom (~~38~~34)	~~0.17~~ 0.18 K



~~Probabilistic projections~~

Figure 4. Surface air temperature change against cumulative CO_2 emissions in the 1pctCO2 and 1pctCO2-4xext experiments. ~~Black line is observed GSAT (??)~~Thin lines are used for the MCE model's family of emulation setups. Thick lines are ~~RCMs~~ used for the GIR (~~error bars represent 66% range~~3 box) and ~~thin lines are CMIP6 results.~~ OSCARv3.1 default setups (aOSCARv3.1's probabilistic output is available but not shown)~~ historical period (1850-2025); (b) - projections (2000-2110)~~.

~~Comparison of temperature projections under~~

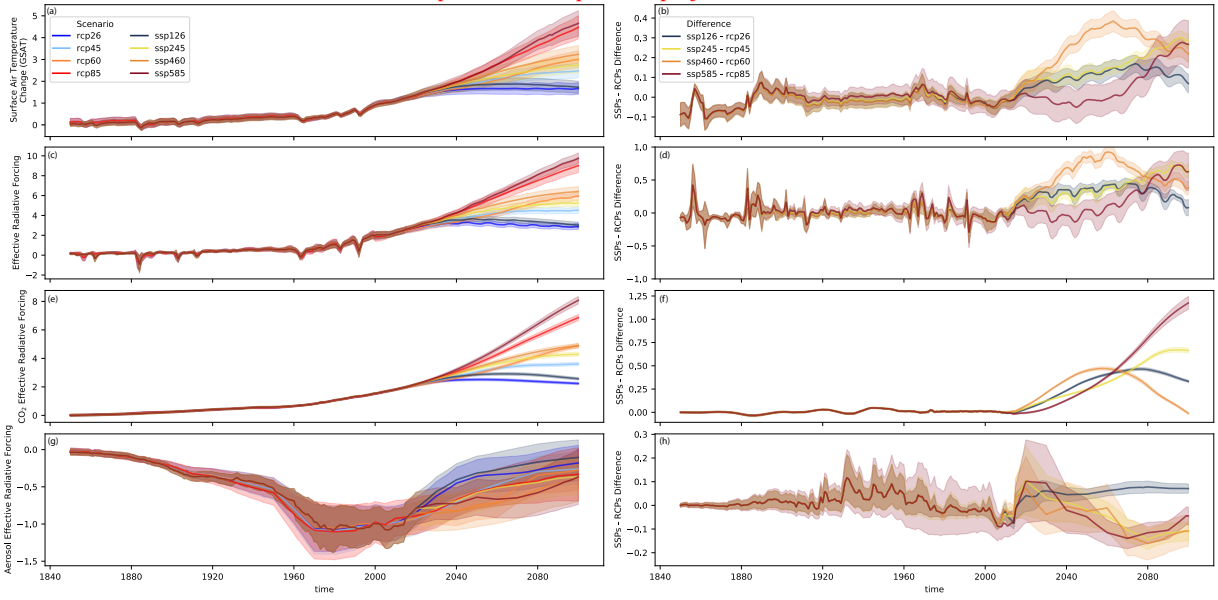


Figure 5. Output from the RCPs and ~~SSPs~~ SSP-based scenarios up until 2100. The ~~coloured solid lines are RCMIP output where the RCP/SSP pair has been run with the same~~ left-hand column shows raw model ~~in the same configuration~~output. ~~For comparison,~~ The right-hand column shows the ~~dotted lines show CMIP5 and CMIP6~~ difference between scenarios for a given model's output. The ~~plumes show~~ shaded range shows one standard deviation ~~of~~ about the ~~available model results whilst the~~ median (solid lines~~show the mean~~). Output is shown for surface air temperature change (GSAT, (a) and (b)), effective radiative forcing ((c) and (d)), CO_2 effective radiative forcing ((e) and (f)) and aerosol effective radiative forcing ((g) and (h)). The results here are illustrative and provided only for those models which have done RCP, SSP-based scenario pairs.