Reduced Complexity Model Intercomparison Project Phase 2: Synthesising Earth system knowledge for probabilistic climate projections

Z. Nicholls[1,2], M. Meinshausen[1,2,3], J. Lewis[1], M. Rojas Corradi [4,5], K. Dorheim[6], T. Gasser[7], R. Gieseke[8], A. P. Hope[9], N. J. Leach[10], L. A. McBride[11], Y. Quilcaille[7], J. Rogelj[12,7], R. J. Salawitch[11,9,13], B. H. Samset[14], M. Sandstad[14], A. Shiklomanov[15], R. B. Skeie[14], C. J. Smith[16,7], S. J. Smith[17], X. Su[18], J.Tsutsui[19], B. Vega-Westhoff[20]and D. Woodward[5]

[1]Australian-German Climate & 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]Department of Geophysics, University of Chile, Santiago, Chile
[5]Center for Climate and Resilience Research, CR2, Santiago, Chile
[6]Pacficic Northwest National Laboratory
[7]International Institute for Applied Systems Analysis, Laxenburg, Austria
[8]Independent researcher, Potsdam, Germany
[9]Department of Atmospheric and Oceanic Science, University of Maryland-College Park, College Park, 20740, USA
[10]Atmospheric, Oceanic, and Planetary Physics, Department of Physics, University of Oxford, United Kingdom
[11]Department of Chemistry and Biochemistry, University of Maryland-College Park, College Park, 20740, USA
[12]Grantham Institute, Imperial College London, London, UK
[13]Earth System Science Interdisciplinary Center, University of Maryland-College Park, College Park, 20740, USA
[14]CICERO Center for International Climate Research, Oslo, Norway
[15]NASA Goddard Space Flight Center, Greenbelt, MD, USA 20771
[16]Priestley International Centre for Climate, University of Leeds, United Kingdom
[17]Joint Global Change Research Institute, Pacific Northwest National Laboratory, College Park, MD, USA
[18]Research Institute for Global Change / Research Center for Environmental Modeling and Application / Earth System Model Development and Application Group, Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokohama, Japan
[19]Environmental Science Research Laboratory, Central Research Institute of Electric Power Industry, Abiko, Japan
[20]Department of Atmospheric Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Key Points:

- Reduced complexity climate models (RCMs) are key for making probabilistic climate projections because of their computational efficiency
- We evaluate how well RCMs' probabilistic setups can simultaneously reflect and emulate Earth system knowledge from multiple specialist research domains
- No model is able to capture all forcing, warming, heat uptake and carbon cycle metrics we evaluate, however some come very close, with deviations greater than 10% in only four metrics

---

Corresponding author: Zebedee Nicholls, `zebedee.nicholls@climate-energy-college.org`

Abstract

Over the last decades, climate science has branched out into many smaller expert communities across the carbon cycle, radiative forcings, climate feedbacks or ocean heat uptake domains. Our best tools to capture state-of-the-art knowledge are the increasingly complex fully coupled Earth System Models (ESMs). However, computational limitations and the structural rigidity of ESMs mean that the full range of uncertainties across multiple domains are difficult to capture with multi-model ESM ensembles and perturbed parameter single ESM ensembles alone. The tools of choice are hence more computationally efficient reduced complexity models (RCMs), which are structurally flexible and can span the response dynamics across a range of domain-specific models and/or ESM experiments. Here, we provide the first comprehensive intercomparison of multiple RCMs that are probabilistically calibrated to key benchmark ranges from specialised research communities. This exercise constitutes Phase 2 of the Reduced Complexity Model Intercomparison Project (RCMIP Phase 2). We find that even if RCMs perform similarly against historical benchmarks, their future projections can still diverge. Under the low-emissions SSP1-1.9 scenario, across the RCMs, median 2081-2100 warming projections range from 1.1 to 1.4°C while median peak warming projections range from 1.3 to 1.7°C (relative to 1850-1900, using an observationally-based historical warming estimate of 0.8°C between 1850-1900 and 1995-2014). Our findings suggest that users of RCMs should carefully evaluate the RCM they are using, specifically its skill against key benchmarks and consider the need to include future projections benchmarks either from ESM results or other assessments to reduce such divergence.

Plain Language Summary

Our best tools to capture state-of-the-art knowledge are complex, fully coupled Earth System Models (ESMs). However, ESMs are expensive to run and no single ESM can easily produce responses which represent the full range of uncertainties. Instead, for some applications, computationally efficient reduced complexity climate models (RCMs) are used in a probabilistic setup. An example of these applications is estimating the likelihood that an emissions scenario will stay below a certain global-mean temperature change (e.g. 2°C). Here we present a study (referred to as the Reduced Complexity Model Intercomparison Project (RCMIP) Phase 2) which investigates the extent to which different RCMs can be probabilistically calibrated to reproduce key benchmark ranges from specialised research communities. We find that the agreement between each RCM and the benchmarks varies, although the best performing models show good agreement with both the best-estimate and uncertainty ranges over the majority of benchmarks. Even though the models all used the same target benchmark ranges, their future projections still diverge. Under the low-emissions SSP1-1.9 scenario, across the RCMs, median peak warming projections range from 1.3 to 1.7°C (relative to 1850-1900, using an observationally-based historical warming estimate of 0.8°C between 1850-1900 and 1995-2014).

1 Introduction

Coupled Earth System Models (ESMs) have evolved for decades as primary climate research tools (Edwards, 2000). They represent the state of the art of complex Earth system modelling. Nonetheless, they are not the tool of choice to assess the full breadth of scenario and Earth system response uncertainty that has been identified in the scientific literature. It is infeasible to assess the climate implications of hundreds to thousands of emissions scenarios with the world's most comprehensive ESMs, such as those participating in the Sixth Phase of the Couple Model Intercomparison Project (CMIP6) (Eyring et al., 2016), because of ESMs' computational cost, the complexity in setting up input data and the sheer volume of output data generated. Yet, such assessments are vital for understanding the consequences of various policy choices and their residual climate hazards.

Similarly, while some ESMs perform large, perturbed physics experiments (e.g., Stainforth et al., 2005) that aim to explore the full range of potential Earth system long-term annual-average responses, the ability to capture full uncertainty ranges is limited. The ability to capture full uncertainty ranges is limited because these ESMs are relatively rigid in their structure - lacking a representation of uncertainties in vital components like the carbon cycle or effective radiative forcings.

An answer to both of these challenges, i.e. (a) limited computational resources and (b) structural scope and flexibility to represent long-term uncertainties in key metrics like global-mean surface air temperatures, are Reduced Complexity Models (RCMs), often also referred to as simple climate models (SCMs). RCMs can play the vital role of extending the knowledge and uncertainties from multiple domains, particularly a multitude of ESM experiments, to probabilistic long-term climate projections of key variables over a wide range of scenarios (see Section 2 in (Meinshausen et al., 2011) for other uses of RCMs).

Typically, RCMs achieve this computational efficiency and structural flexibility by limiting their spatial and temporal domains to global-mean, annual-mean quantities i.e the domains of relevance to long-term, global climate change. Rather than aiming to represent the physics of the climate system at the process level and high-resolution, RCMs use parameterisations of the system which capture its large-scale behaviour at a greatly reduced computational cost. This allows them to perform 350-year long simulations in a fraction of a second on a single CPU, multiple orders of magnitude faster than our most comprehensive ESMs which would take weeks to months on the world's most advanced supercomputers.

A key example of large-scale emissions scenario assessment, and the one we focus on in this paper, is the climate assessment of socioeconomic scenarios by the Intergovernmental Panel on Climate Change (IPCC) Working Group 3 (WG3). Hundreds of emission scenarios were assessed in the IPCC's Fifth Assessment Report (AR5, see Clarke et al. (2014)) as well as its more recent Special Report on Global Warming of 1.5°C (SR1.5, see Rogelj et al. (2018); Huppmann et al. (2018)). (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). 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 (an initial snapshot of scenarios based on the SSPs is available at `https://tntcat.iiasa.ac.at/SspDb`).

One further reason that the world's most comprehensive ESMs would have difficulty running WG3-type scenarios is because greenhouse gas cycles, atmospheric chemistry and dynamic vegetation modules would be required to run the WG3 emission scenarios. While some ESMs have the required components, they are rarely used for long-term experiments for reasons of computational cost. The most comprehensive RCMs include parameterised representations of the required components, enabling the exploration of interacting uncertainties from multiple parts of the climate system in an internally consistent setup.

In general, RCMs do not include the detail of ESMs across the emissions-climate change cause-effect chain, but they do tend to include uncertainty representations for more steps in the chain (i.e. RCMs tradeoff depth for breadth compared to ESMs). For example, many RCMs include the relationship between methane emissions and concentrations (including temperature and other feedbacks) whereas few ESMs do in their long-term experiments. On the other hand, few RCMs directly use land-cover information within their carbon cycles, and none consider it in the detailed way which ESMs do. In addition, there are clearly applications where RCMs are not a feasible tool. For example, near-term attribution studies, such as the World Weather Attribution project (Uhe et al., 2016). For this latter application, large-ensemble ESM runs are vital - as only they can reflect natural variability and weather patterns. Overall, there is no question that ESMs are by far the most important research tool to project future climate change. RCMs complement the ESM efforts. Within this paper, we focus on a very specific niche of this complementing role, i.e. synthesising multiple lines of evidence across the emissions-climate change cause-effect chain.

Within the IPCC, RCMs' synthesising niche facilitates the transfer of knowledge from Working Group I (WG1), which assesses the physical science of the climate system, to WG3, which assesses the socioeconomics of climate change mitigation. The knowledge transfer ensures that WG3's scenario classification is consistent with the physical science assessment of WG1 - a key precondition to have confidence that WG3's conclusions about the socioeconomic transformation required to mitigate anthropogenic climate change to specific levels are based on our latest scientific understanding. Here, we describe RCMs as 'integrators of knowledge' because they integrate (a relevant sub-section of) the assessment from WG1, providing WG3 with a tool that can be used for assessing the climate implications, particularly global-mean temperature changes, of a wide range of emissions scenarios.

Typically, RCMs perform this knowledge integration using probabilistic distributions, which are distinct from the emulator mode in which RCMs can also be run (see Nicholls et al. (2020) for a discussion of emulation with RCMs). These probabilistic distributions are derived by running an RCM with a parameter ensemble which captures the assessed ranges of specific Earth system quantities, e.g. historical global mean temperature increase, effective radiative forcing due to different anthropogenic emissions, ocean heat uptake, or cumulative land and ocean carbon uptake. The resulting distributions are designed to facilitate WG3's scenario classification e.g. 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 (in this case the WG1 assessment). As a result of their probabilistic nature, the ensembles resulting from RCMs are conceptually different from an ensemble of multiple model outputs (such as those from CMIP6) taken without constraining or any other sort of post-processing.

Due to their role in the IPCC assessment (and for analysing mitigation options in line with temperature targets more generally), understanding the degree to which RCMs can reflect a range of radiative forcing, warming, heat uptake and concentration assessments simultaneously is of vital importance. If RCMs are inherently biased in some way, this will affect the WG3 climate assessment and interpretation of the RCMs' outputs should be adjusted accordingly.

This study's scope, in terms of number of climate dimensions considered and number of climate models evaluated, is unique. There have been studies with single models which choose parameter sets that match various assessments of ECS and TCR (Meinshausen et al., 2009; Rogelj et al., 2012). Smith, Forster, et al. (2018) compared two models' probabilistic outputs.

Here, in the second phase of RCMIP, we evaluate the degree to which multiple RCMs are able to synthesise Earth system knowledge within a probabilistic distribution. We then examine the implications of differences in these probabilistic distributions for climate projections. We extend previous probabilistic evaluation work and build on the progress made in the first phase (Nicholls et al., 2020) and other RCM intercomparison studies (van Vuuren et al., 2011; Harmsen et al., 2015; Schwarber et al., 2019). We widen the first phase's scope both in terms of number of climate dimensions considered and the number of models evaluated. To our knowledge, this is the most comprehensive evaluation performed to date of the ability of RCMs to capture a broad range of climate metrics and key indicators, such as those assessed in by IPCC WG1.

2 Participating models

Nine models have participated in RCMIP Phase 2 (Table 1 and Supplementary Text S1). These models and their components range from simpler, regression-based approaches to more complex representations with detailed processes and regions. The models have been constrained in a number of different ways, using statistical techniques ranging in complexity from Monte Carlo Markov Chains to using pass/fail criteria to determine valid parameter values. As a result, they cover a wide range of the techniques in the literature and their results allow us to evaluate the implications of different choices.

3 Methods

In this study, the RCMs are run in a probabilistic setup. As discussed in the introduction, a probabilistic setup means that each RCM is run with an ensemble of parameters. Specifically, for a given experiment, each RCM is run multiple times, each time with slightly different parameter values. All of these different runs are then combined to form a probabilistic set of outputs. With these probabilistic sets, we can then calculate ranges of each output variable of interest (e.g. global-mean surface temperatures).

Modelling groups use a range of techniques to derive their parameter ensembles i.e. to constrain their models (Table 1). Typically, modelling groups will also use different data to derive their parameter ensemble. This can lead to differences in model projections which are simply based on choices made by the modelling groups and are not related to model structure or constraining technique at all. We remove the choice of data as a point of difference by ensuring that all modelling groups agree on a common set of target assessed ranges i.e. benchmarks.

In this study, our target assessment is a 'proxy assessment', which uses assessed climate system characteristics in line with IPCC AR5 as its starting point and updates key values using more recent literature (see Table 2). We explicitly use the name 'proxy assessment' throughout to make clear that we are not constraining to any ranges coming from the formal IPCC assessment, rather an approximation thereof.

We use surface air ocean blended temperatures from the HadCRUT.4.6.0.0 dataset (Morice et al., 2012). HadCRUT4.6.0.0 is a widely used observational data product and is representative of other observations of changes in surface air and ocean temperatures (Simmons et al., 2017). Our key metric for evaluating RCM temperature projections is the warming between the 1961-1990 and 2000-2019 periods (using the SSP2-4.5 scenario to extend the CMIP6 historical experiment to 2019). We choose a relatively recent period to match the increase in global observations since the 1960s.

For ocean heat content, we use the recent work of von Schuckmann et al. (2020). We focus on the change in ocean heat content between 1971 and 2018, when the largest set of observations are available.

We use the recent assessment of Sherwood et al. (2020) for equilibrium climate sensitivity (ECS). ECS is defined as the equilibrium warming which occurs under a doubling of atmospheric CO_2 concentrations relative to pre-industrial concentrations. The ECS assessment is combined with the constrained transient climate response (TCR) assessment of Tokarska et al. (2020). TCR is defined as the surface air temperature change which occurs at the time at which atmospheric CO_2 concentrations double in an experiment in which atmospheric CO_2 concentrations rise at one percent per year (a 1pctCO2 experiment). Carbon cycle behaviour is considered via the transient climate response to emissions (TCRE). TCRE is defined as the ratio of surface air temperature change

Table 1. Overview of the models and constraining approaches used in this paper. Detailed descriptions of each model are available in Supplementary Text S1.

Model	Constraining technique	Key references
Cicero-SCM	550 members sub-sampled from a posterior of 30 040 members to form a set that match the proxy assessment ECS distribution while reproducing surface air temperature change from 1850-1900 to 1985-2014	Schlesinger et al. (1992); Joos et al. (1996); Etminan et al. (2016); Skeie et al. (2017, 2018); Nicholls et al. (2020)
EMGC	160 000 sample members, retaining the 1 000 that minimize reduced-chi-squared between modeled and observed GMST and OHC from 1850-1999	Canty et al. (2013); Hope et al. (2017, 2020); McBride et al. (2020)
FaIRv1.6.1	3 000 sample members retaining the 501 that minimise RMSE between modelled and observed 1850-2014 GMST	Millar et al. (2017); Smith, Forster, et al. (2018)
FaIRv2.0.0-alpha	1 million member raw ensemble, constrained with 90% credible range of current level and rate of attributable warming (Haustein et al., 2017). 5000 members randomly drawn from the constrained ensemble for use here.	Millar et al. (2017); Haustein et al. (2017); Smith, Forster, et al. (2018); Leach et al. (2020)
Hectorv2.5.0	10 000 sampled ensemble from Markov chain Monte Carlo chains constrained with global surface temperature and ocean heat content	Vega-Westhoff et al. (2019)
MAGICCv7.4.1	~ 20 million member Monte Carlo Markov Chain, 600 member sub-sample selected to match proxy assessed ranges	Meinshausen et al. (2009, 2011, 2020)
MCE v1.2	600 members sampled with a Metropolis-Hastings algorithm through Bayesian updating to reflect an ensemble of complex climate models constrained with the proxy assessed ranges	Tsutsui (2017, 2020) (see also Joos et al. (1996); Hooss et al. (2001))
OSCARv3.1	10 000 Monte Carlo members, weighted using their agreement with a set of assessed ranges (Supplementary Text S1)	Gasser et al. (2017, 2018, 2020)
SCM4OPT v2.0	For each emission scenario, 2 000 sample members are used to reflect uncertainties resulting from carbon cycle, aerosol forcings and temperature change, while constrained by the historical mean surface temperature of HadCRUT.4.6.0.0 (Morice et al., 2012).	Su et al. (2017, 2018, 2020)

Table 2. The proxy assessed ranges used in this study. The assessed ranges are labelled as 'vll' (very-likely lower i.e. 5th percentile), 'll' (likely lower, 17th percentile), 'c' (central, 50th percentile), 'lu' (likely upper, 83th percentile) and 'vlu' (very-likely upper, 95th percentile). Sources are described in Section 3.

Metric	Assessed range Unit	vll	ll	c	lu	vlu
2000-2019 GMST rel. to 1961-1990	K	0.50	0.52	0.54	0.56	0.58
Equilibrium Climate Sensitivity	K	2.30	2.60	3.10	3.90	4.70
Transient Climate Response	K	0.98	1.26	1.64	2.02	2.29
Transient Climate Response to Emissions	K / TtC	1.03	1.40	1.77	2.14	2.51
2014 CO_2 Effective Radiative Forcing	W / m^2		1.69	1.80	1.91	
2014 Aerosol Effective Radiative Forcing	W / m^2		-1.37	-1.01	-0.63	
2018 Ocean Heat Content rel. to 1971	ZJ		303	320	337	
2011 CH_4 Effective Radiative Forcing	W / m^2		0.47	0.60	0.73	
2011 N_2O Effective Radiative Forcing	W / m^2		0.14	0.17	0.20	
2011 F-Gases Effective Radiative Forcing	W / m^2		0.03	0.03	0.03	

to cumulative CO_2 emissions at the time when atmospheric CO_2 concentrations double in a 1pctCO2 experiment. We use the TCRE assessment from Arora et al. (2020), which is based on the latest generation of Earth System Models which have participated in CMIP6 (Eyring et al., 2016). There is a potential inconsistency between our ECS, TCR and TCRE ranges, which arises because the TCR assessment is based on a constrained set of CMIP6 models, the TCRE assessment is based on unconstrained CMIP6 Earth System Models and the ECS assessment comes from a study which uses multiple lines of evidence. We discuss the importance of this inconsistency and its consequences in 4.

The other key metrics are related to effective radiative forcing (ERF, Forster et al., 2016). These values generally follow the AR5 assessment, except for aerosol, CO_2 and CH_4 ERF. For aerosol and CO_2 ERF, we use the more recent work of Smith et al. (2020). For CH_4 ERF, we increase the AR5 assessment following Etminan et al. (2016) although we note that this increase may be offset by an updated understanding of the impact of rapid adjustments following Smith, Kramer, et al. (2018).

At this point, we stress that our proxy assessed ranges are only one of a range of possible choices. Assessing all the available literature is a demanding task that is well undertaken by the IPCC. We do not attempt to reproduce this task here. Instead, the key is that our proxy assessed ranges are a) reasonable and b) available now so all modelling groups can use consistent benchmarks to constrain their models.

Following this intercomparison consortium's choice of proxy assessed ranges, modelling groups then had the opportunity to develop parameter ensembles which best reflected these assessed ranges. As a result, we have, for the first time, a set of models, all of which used the same 'constraining benchmarks' (with a number of different techniques being employed to consider the constraining benchmarks, see Table 1). We gain unique insights into the impact of differences in model structure and constraining techniques when RCMs are used as integrators of knowledge, free from a typical source of disagreement between the models, namely that they were constrained to reproduce different understandings of the climate.

The modelling groups submitted a range of concentration-driven, emission-driven and idealized scenarios for their chosen parameter subsets (see scenario specifics below). Subsequently, several metrics were calculated, such as TCR from the idealised CO_2-only

1pctCO2 experiment (in which atmospheric CO_2 concentrations rise at 1% per year from pre-industrial levels). Calculating derived metrics on each individual ensemble member ensures that all metrics are calculated from internally self-consistent model runs, which is of particular importance when the metric is based on more than one output variable from the model (e.g. TCRE, which relies on both surface air temperature change and inverse emissions of CO_2). If we instead calculated results based on percentiles of different variables, we would not be using an internally self-consistent set. Where modelling groups felt it was more appropriate (e.g. OSCAR), they performed their own weighting of ensemble members before submitting.

The one metric which is not easily calculated from model results is ECS because it is defined at equilibrium. Accordingly, modelling groups reported their own diagnosed ECS for each ensemble member, rather than performing experiments which would allow it to be calculated after submission had taken place.

When evaluating model performance, we are interested not only in how well a model can reproduce the best estimate, but also the range of a given quantity. A key part of any climate assessment is the uncertainty and it is critical that RCMs reflect the assessed likely and very likely ranges if they are to be used as integrators of knowledge. We assess the relative difference between the model and the assessed ranges at the very likely lower (5^{th} percentile, also referred to as 'vll'), likely lower (17^{th} percentile, ll), central (50^{th} percentile, c), likely upper (83^{th} percentile, lu) and very likely upper (95^{th} percentile, vlu). Assessing deviations using relative differences allows us to quickly evaluate how models perform over a range of metrics on the same scale.

The set of scenarios that each modelling group was asked to run follow the experimental protocols of CMIP6's ScenarioMIP (O'Neill et al., 2016). The SSPX-Y.Y experiments (e.g. SSP1-1.9, SSP2-4.5, SSP5-8.5) are defined in terms of concentrations of well-mixed greenhouse gases i.e. CO_2, CH_4, N_2O, hydrofluorocarbons (HFCs), perfluorocarbons (PFCs) and hydrochlorofluorocarbons (HCFCs), emissions of 'aerosol precursor species emissions' i.e. sulfur, nitrates, black carbon, organic carbon and ammonia and natural effective radiative forcing variations. As described in Nicholls et al. (2020), where required, models may use prescribed effective radiative forcing where they do not include the required gas cycles or radiative forcing parameterisations.

The esm-SSPX-Y.Y experiments are identical to the SSPX-Y.Y experiments, except CO_2 emissions are prescribed instead of CO_2 concentrations, following the CMIP6 C4MIP protocol (Jones et al., 2016). Finally, we also perform esm-SSPX-Y.Y-allGHG experiments. These are identical to the esm-SSPX-Y.Y experiments, except they are defined in terms of emissions of all well-mixed greenhouse gases, not only CO_2, rather than concentrations. There is no equivalent of these esm-SSPX-Y.Y-allGHG experiments in the CMIP6 protocol, however it is these experiments which are of most interest to WG3, given that WG3 focusses on scenarios defined in terms of emissions alone. We use the data sources described in Nicholls et al. (2020) to specify the inputs for each of these scenarios. The input dataset compilations, comprising emission, scenario and forcing data, as well as the protocols are available at `rcmip.org` (last accessed 28 October 2020) - and can contribute to scientific studies beyond this intercomparison as they largely reflect the CMIP6 experimental designs.

The protocol designed for this study requires that each RCM modelling group runs every probabilistic ensemble member once for each scenario and then submits their output for further analysis. With nine modelling groups participating, this intercomparison project compiled a database of results containing thousands of runs for each RCM, from which we can calculate different warming, effective radiative forcing or ocean heat uptake percentiles for a wide range of scenarios.

4 Results and discussion

4.1 Fit to assessed ranges

The ability of RCMs to match the assessed ranges varies (Table 3, Supplementary Table S1 and Supplementary Figures S1 - S10). In general, the RCMs capture the central assessed values better than the likely and very likely ranges. Historical warming and the TCRE are notable exceptions to this. For both these metrics, the very likely lower and likely lower assessed values are better captured by the RCMs than the central values.

Considering the variation between metrics, we see that the proxy assessment of the ECS and effective radiative forcing metrics is better captured by the RCMs than the other metrics (see multi-model median in Table 3). For ECS and all the effective radiative forcing metrics, the median multi-model difference is less than or equal to 10% for the central proxy assessed range. However, there is less close agreement with the very likely and likely proxy assessed ranges for the ECS and effective radiative forcing metrics, with median multi-model differences being up to 18% (CH_4 effective radiative forcing).

For the other metrics (historical warming, TCR, TCRE and historical ocean heat content changes), the median multi-model difference is greater than 20% for at least one of the assessed ranges. However, there is significant variation across the likelihood levels. For example, the multi-model median matches the very likely lower and likely lower historical warming (rows labelled '2000-2019 GMST rel. to 1961-1990' in Table 3) to within 2% and 6% respectively. However, the multi-model median differs from the central, likely upper and very likely upper historical warming by 11%, 25% and 44% respectively, indicating that the models are having greater difficulty capturing the upper-end warming estimates.

There is also significant spread in performance across the models. Two models perform better than the multi-model median across all metrics and assessed ranges (very likely lower, likely lower, central, likely upper, very likely upper) except for three metrics. Those models are MAGICC7 (worse than multi-model median for all assessed ranges of TCR, likely lower 2011 CH_4 effective radiative forcing and very likely lower TCRE) and MCE-v1-2 (worse than multi-model median for all assessed ranges of ECS, very likely lower and very likely upper TCR and likely lower, central, likely upper and very likely upper TCRE). However, all RCMs had at least one strength where they matched the proxy assessment at all likelihood levels to within 20%.

4.2 Projections

For each probabilistic setup, the RCMs also submitted projections of global-mean surface temperature, effective radiative forcing (split into total, aerosols and CO_2) and atmospheric CO_2 concentrations for the SSPX-Y.Y, ESM-SSPX-Y.Y and ESM-SSPX-Y.Y-allGHG experiments. Despite all being constrained with the same target distributions, there are considerable differences between the projections from various models.

4.2.1 Global-mean Surface Air Temperature

Under SSP1-1.9, median end of century (2081-2100) projections relative to 1995-2014 vary by 0.3°C across the models (from Cicero-SCM, EMGC and Hector with 0.3°C of warming to MAGICC7, FaIR1.6 and FaIRv2.0.0-alpha with 0.6°C, Figure 1 a)-c)). Variations in 5^{th} percentile warming show a similar range, from -0.1°C to 0.2°C. In contrast, upper-end, 95^{th} percentile warming shows far greater variation, from 0.4°C for OSCARv3.1 to 1.9°C for EMGC. For the SSP1-1.9 scenario, the spread in RCMs' probabilistic projections is similar to the spread in the CMIP6 multi-model ensemble. Nonetheless, the most extreme CMIP6 model projections are outside the range of most RCMs' 5-95^{th} per-

Table 3. Comparison of each model's probabilistic distribution with the proxy assessment. In each square, we show the relative difference between the model result and the proxy assessed value (Δ_m, calculated as $\Delta_m = \frac{m-a}{|a|}$ where m is the value from the model's probabilistic distribution and a is the proxy assessment value). If a row is completely empty for a model, this indicates that the model did not submit results which allowed that metric to be calculated. Empty cells within a row which is otherwise not completely empty for a model indicates that no proxy assessment at this likelihood level was available (e.g. we have proxy assessments for likely lower 2014CO$_2$ effective radiative forcing, but not for very likely lower 2014CO$_2$ effective radiative forcing). Only the magnitude of Δ_m from each model was used to calculate the multi-model median (to ensure that positive and negative values of Δ_m from different models would not cancel out). The assessed ranges are labelled as 'vll' (very-likely lower i.e. 5th percentile), 'll' (likely lower, 17th percentile), 'c' (central, 50th percentile), 'lu' (likely upper, 83th percentile) and 'vlu' (very-likely upper, 95th percentile). (Note, continues on next page.)

Climate model	Multi-model median of magnitude of relative differences				
Assessed range	vll	ll	c	lu	vlu
2000-2019 GMST rel. to 1961-1990	2%	6%	11%	25%	44%
Equilibrium Climate Sensitivity	15%	9%	8%	11%	16%
Transient Climate Response	37%	19%	9%	7%	8%
Transient Climate Response to Emissions	13%	11%	20%	20%	22%
2014 CO$_2$ Effective Radiative Forcing		5%	5%	2%	
2014 Aerosol Effective Radiative Forcing		16%	10%	12%	
2018 Ocean Heat Content rel. to 1971		9%	7%	24%	
2011 CH$_4$ Effective Radiative Forcing		10%	7%	18%	
2011 N$_2$O Effective Radiative Forcing		11%	3%	7%	
2011 F-Gases Effective Radiative Forcing		2%	3%	4%	

Climate model	Cicero-SCM					EMGC					FaIR1.6				
Assessed range	vll	ll	c	lu	vlu	vll	ll	c	lu	vlu	vll	ll	c	lu	vlu
2000-2019 GMST rel. to 1961-1990	-19%	-4%	14%	34%	48%	-36%	-28%	-11%	17%	44%	5%	11%	23%	36%	46%
Equilibrium Climate Sensitivity	-6%	-6%	-5%	-4%	-6%	-43%	-42%	-38%	-28%	-12%	-16%	-11%	-3%	11%	32%
Transient Climate Response	12%	4%	1%	5%	8%						43%	23%	10%	5%	8%
Transient Climate Response to Emissions											17%	-3%	-10%	-11%	-12%
2014 CO$_2$ Effective Radiative Forcing		15%	8%	2%			12%	6%	-0%			-2%	5%	14%	
2014 Aerosol Effective Radiative Forcing		33%	43%	80%			16%	16%	16%			2%	0%	-4%	
2018 Ocean Heat Content rel. to 1971		-16%	-7%	2%			-9%	2%	26%			-0%	12%	24%	
2011 CH$_4$ Effective Radiative Forcing		12%	-12%	-27%			23%	-3%	-20%			3%	-8%	-15%	
2011 N$_2$O Effective Radiative Forcing		13%	-6%	-20%			25%	4%	-12%			8%	-2%	-9%	
2011 F-Gases Effective Radiative Forcing												-1%	-3%	-4%	

Table 3. (Continued.)

Climate model	FaIRv2.0.0-alpha					Hector					MAGICC7				
Assessed range	vll	ll	c	lu	vlu	vll	ll	c	lu	vlu	vll	ll	c	lu	vlu
2000-2019 GMST rel. to 1961-1990	2%	12%	26%	39%	50%	-2%	6%	15%	25%	33%	1%	2%	2%	2%	3%
Equilibrium Climate Sensitivity	-15%	-9%	3%	13%	24%	-20%	-17%	-8%	0%	16%	-8%	-9%	-6%	-1%	2%
Transient Climate Response	29%	17%	9%	7%	7%	45%	25%	11%	3%	0%	54%	32%	20%	16%	14%
Transient Climate Response to Emissions	-5%	-18%	-20%	-20%	-22%						29%	9%	8%	8%	5%
2014 CO_2 Effective Radiative Forcing		2%	8%	13%								-1%	1%	2%	
2014 Aerosol Effective Radiative Forcing		-15%	-16%	-21%			51%	44%	29%			-5%	-8%	-10%	
2018 Ocean Heat Content rel. to 1971												-1%	0%	2%	
2011 CH_4 Effective Radiative Forcing		7%	0%	-3%								-12%	-7%	-3%	
2011 N_2O Effective Radiative Forcing		6%	-1%	-6%								-5%	-3%	1%	
2011 F-Gases Effective Radiative Forcing		3%	5%	7%								-1%	-1%	-0%	

Climate model	MCE-v1-2					OSCARv3.1					SCM4OPTv2.0				
Assessed range	vll	ll	c	lu	vlu	vll	ll	c	lu	vlu	vll	ll	c	lu	vlu
2000-2019 GMST rel. to 1961-1990	-2%	-1%	1%	1%	2%	0%	1%	0%	0%	0%	-37%	-15%	10%	33%	57%
Equilibrium Climate Sensitivity	-23%	-23%	-23%	-25%	-25%	3%	-9%	-18%	-16%	-25%	13%	4%	12%	6%	-3%
Transient Climate Response	24%	4%	-8%	-14%	-16%	50%	22%	-1%	-13%	-16%	32%	16%	-1%	-7%	1%
Transient Climate Response to Emissions	0%	-18%	-24%	-24%	-28%	13%	-11%	-22%	-26%	-30%					
2014 CO_2 Effective Radiative Forcing		-1%	-0%	1%			13%	6%	0%			-7%	4%	11%	
2014 Aerosol Effective Radiative Forcing		13%	6%	-12%			-20%	-10%	-2%			-21%	-8%	0%	
2018 Ocean Heat Content rel. to 1971		-2%	-0%	1%			-18%	9%	33%			-34%	36%	101%	
2011 CH_4 Effective Radiative Forcing		-0%	-1%	-2%			5%	-17%	-32%			-15%	-17%	-23%	
2011 N_2O Effective Radiative Forcing		-2%	-1%	-1%			26%	5%	-11%			27%	9%	-3%	
2011 F-Gases Effective Radiative Forcing		-1%	-0%	-1%			15%	4%	-6%			20%	12%	3%	

centiles, suggesting that such projections are incompatible with current observations of historical warming and ocean heat content as well as effective radiative forcing understanding (a similar conclusion to Tokarska et al. (2020)).

A similar spread is seen in peak temperature (Figure 1 f)-g)). Across the RCM ensemble, SSP1-1.9 median peak warming ranges from 0.6°C to 0.75°C while the 5th and 95th percentiles range from 0.2°C to 0.5°C and 0.7°C to 2.0°C, respectively. In contrast, the year of peak warming shows much more variation, particularly at the upper end (Figure 1 d)-e)). While the median peak year is fairly consistent across the RCMs' ensembles, around 2045, and the 5th percentile peak year varies from 2030 to 2040, the 95th percentile varies from 2050 to beyond the end of this century. In SCM4OPTv2.0, EMGC, FaIR1.6 and FaIRv2.0.0-alpha, there is a significant area of parameter space which results in ongoing warming even after CO_2 emissions have reached net zero. However, the warming rate is quite slow in these simulations because there is not an equivalently large spread in end of century temperature projections (see the relatively consistent 95th percentile end of century projections in Figure 1 f)-g)).

In the SSP1-2.6 scenario, median warming is 0.3-0.5°C higher than in SSP1-1.9 (Supplementary Figure S11). Median end of century warming (relative to 1995-2014) ranges from 0.6°C to 1.0°C. End of century 5th percentile warming ranges from 0.1°C to 0.5°C and 95th percentile warming ranges from 1.2°C to 1.9°C. A number of CMIP6 model projections lie above the upper end of the constrained RCMs for this SSP1-2.6 scenario.

Under SSP1-2.6, the RCMs diverge more in their peak temperature projections, both compared to end of century warming and compared to SSP1-1.9. Once again, the 5th percentile and median are fairly consistent (ranging from 0.3°C to 0.8°C and 0.7°C to 1.1°C respectively). However, 95th projections vary from 1.2°C to 2.8°C. The upper-end is driven by FaIR1.6, and appears to be the result of persistent warming after CO_2 emissions reach net zero given that its 83rd percentile peak warming year is after 2100. Across the models, peak warming year shows a similar range to SSP1-1.9, albeit occurring 25-30 years later in the median (ranging from 2065 to 2075). Once again, the 5th percentile (ranging from 2050 to 2060) shows a much smaller spread across the models than the 95th percentile (ranging from 2075 to beyond the end of the 21st Century).

The warmest RCMs in mitigation scenarios are also the warmest under the high-emissions, SSP5-8.5, scenario (Supplementary Figure S12). The exception is MAGICC7, which is about 0.5°C warmer by the end of the century than all other models in the median under SSP5-8.5, in contrast to the mitigation scenarios where it showed similar warming levels to both FaIR1.6 and FaIRv2.0.0-alpha. Under SSP5-8.5, median end of century warming ranges from 2.4°C to 4.0°C across the RCMs. Unlike the mitigation scenarios, there is a similar level of disagreement in 5th and 95th percentile warming, with the 5th percentile ranging from 1.8°C to 3.1°C and the 95th percentile ranging from 3.8°C to 5.5°C. MAGICC7 is the model closest to the CMIP6 projections, with most other RCMs showing warming projections well below the CMIP6 multi-model ensemble. Such a difference suggests a structural difference between CMIP6 models and RCMs, which most clearly emerges under high warming scenarios.

The difference between MAGICC7 and the other RCMs becomes even clearer if we consider long-term (2250-2300) warming under the SSP5-8.5 scenario (Figure 2, see Supplementary Figure S13 and Supplementary Figure S14 for long-term warming under SSP1-1.9 and SSP1-2.6 respectively). MAGICC7's median 2250-2300 warming relative to 1995-2014 of 9.5°C is only just below the 83rd percentile of FaIRv2.0.0-alpha and above this percentile for all other models (despite having quite similar long-term effective radiative forcing, see Supplementary Figure S15). There is a significant spread in such long-term projections across the models, with the median ranging from 4.5°C to 9.5°C, 5th percentile from 3°C to 7°C (excluding SCM4OPTv.20 which is a clear outlier) and 95th from 8°C to 14°C. Even these upper end projections are below the highest CMIP6 projections, which

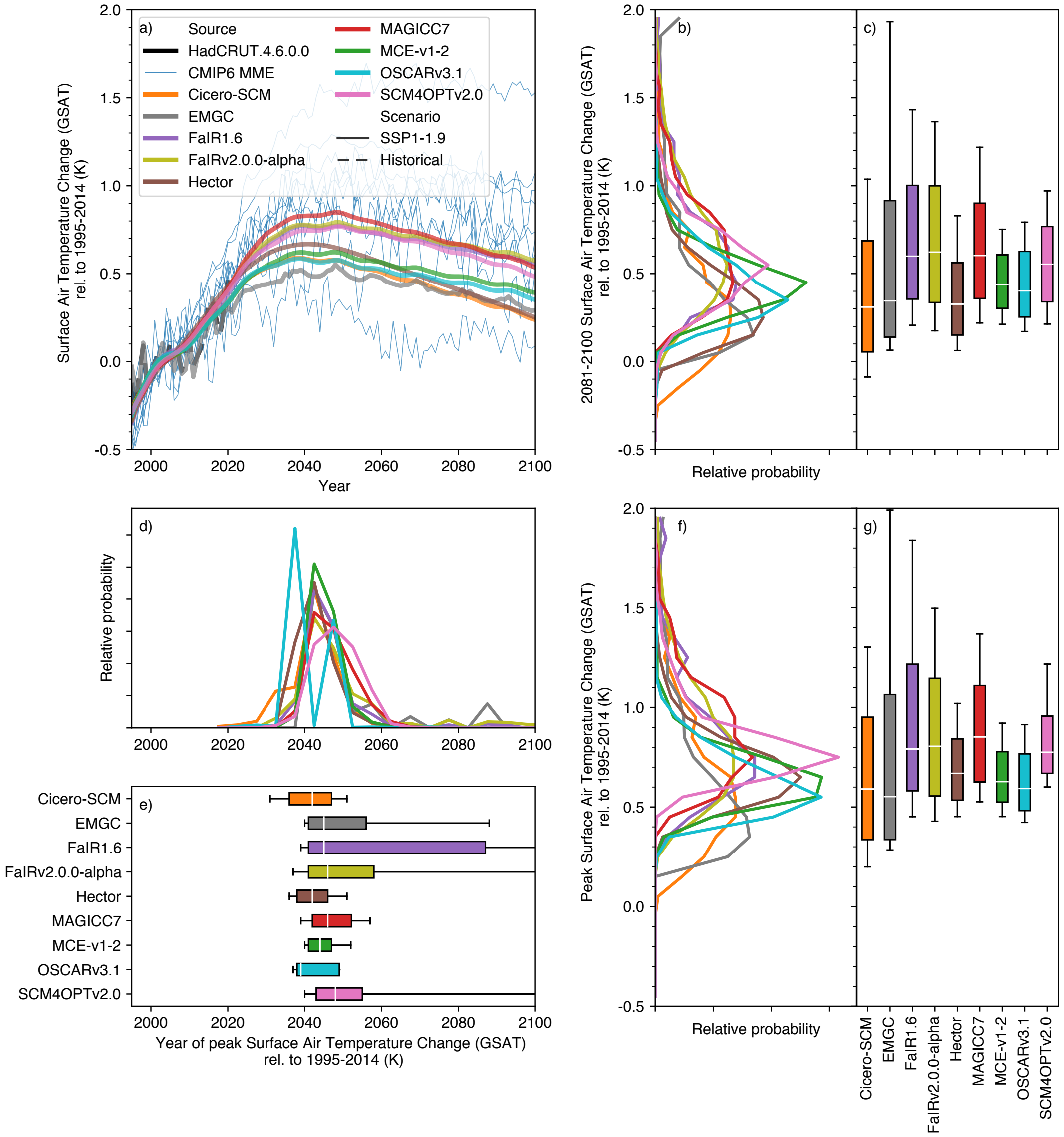

Figure 1. Surface air temperature (also referred to as global-mean surface air temperature, GSAT) change under the very low-emissions SSP1-1.9 scenario. a) GSAT projections from 1995 to 2100. We show the median RCM projections (coloured lines), GMST observations from HadCRUT4.6.0.0 (Morice et al., 2012) up to 2019 (dashed black line) and CMIP6 model projections (thin blue lines, we show the average of all available ensemble members for each CMIP6 model); b) distribution of 2081-2100 mean GSAT from each RCM; c) very likely (whiskers), likely (box) and central (white line) 2081-2100 mean GSAT estimate from each RCM; d) as in b) except for the year in which GSAT peaks; e) as in c) except for the year in which GSAT peaks; f) as in b) except for the peak GSAT; g) as in c) except for the peak GSAT. All results are shown relative to the 1995-2014 reference period.

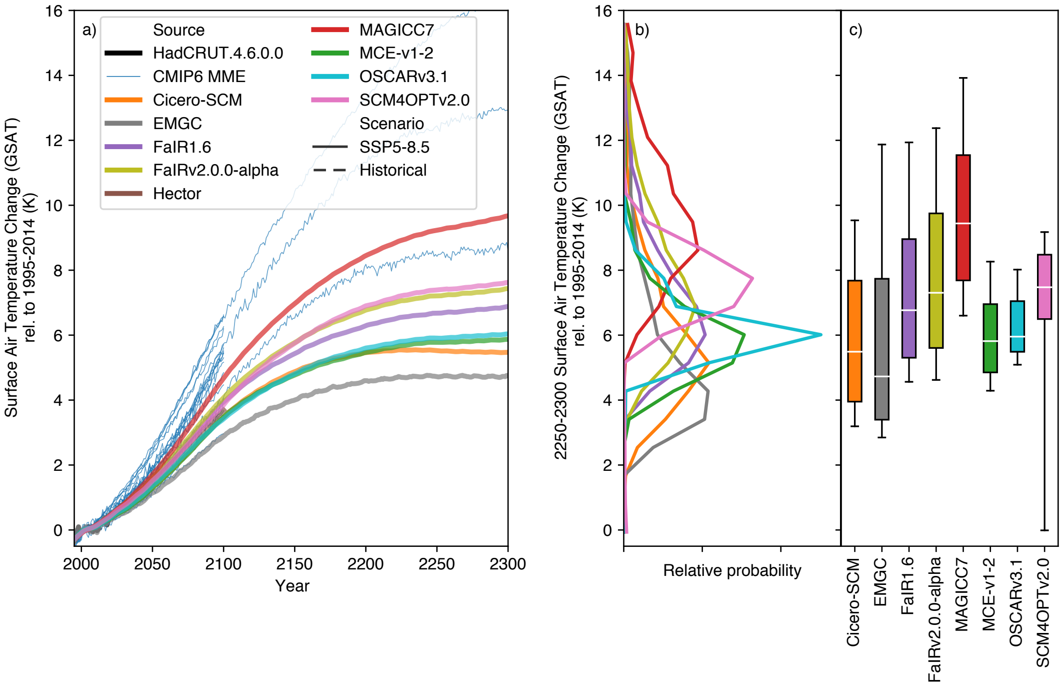

Figure 2. Long-term surface air temperature (also referred to as global-mean surface air temperature, GSAT) change under the high-emissions SSP5-8.5 scenario. a) GSAT projections from 1995 to 2300. We show the median RCM projections (coloured lines), GMST observations from (Morice et al., 2012) up to 2019 (dashed black line) and available CMIP6 model projections (thin blue lines, we show the average of all available ensemble members for each CMIP6 model); b) distribution of 2250-2300 mean GSAT from each RCM; c) very likely (whiskers), likely (box) and central (white line) 2250-2300 mean GSAT estimate from each RCM. All results are shown relative to the 1995-2014 reference period.

reach up to 17°C of global-mean warming. Across all the RCMs, only Cicero-SCM shows any sign of temperatures peaking by 2300 under such a high-emissions scenario.

4.2.2 Effective Radiative Forcing

Compared to temperatures, there is relatively less variance in end of century total effective radiative forcing projections (Figure 3, Supplementary Figure S16 and Supplementary Figure S17), with SCM4OPTv2.0 being a clear outlier. This finding reinforces the understanding that the parameterisation of the climate response to effective radiative forcing is a key driver of climate projection uncertainty.

In SSP1-1.9, 2081-2100 mean total effective radiative forcing varies from 2.2 W / m^2 to 2.6 W / m^2, with SCM4OPTv2.0 being a an outlier with only 1.7 W / m^2. The spread is larger for the upper, 95^{th}, percentile and lower for the lower, 5^{th} percentile. The 95^{th} percentile ranges from 2.5 W / m^2 to 3.2 W / m^2 while the 5^{th} percentile ranges from 1.9 W / m^2 to 2.1 W / m^2 across the models (excluding SCM4OPTv2.0 and Cicero-SCM which has an extremely narrow range). This trend, of uncertainty being higher for upper percentiles than lower percentiles, is seen across other key scenarios and highlights that the high effective radiative forcing risks are much more uncertain than the best case low effective radiative forcing projections.

In SSP1-2.6 (Supplementary Figure S16, once again excluding SCM4OPTv2.0 as an outlier and Cicero-SCM because of its narrow range) median 2081-2100 total effective radiative forcing ranges from 2.9 W / m^2 to 3.4 W / m^2 while the 5th percentile only ranges from 2.5 W / m^2 to 2.7 W / m^2 and the 95th percentile has a much wider range of 3.2 W / m^2 to 4.1 W / m^2. Under SSP5-8.5 (Supplementary Figure S17, excluding EMGC and Cicero-SCM as outliers), median 2081-2100 total effective radiative forcing ranges from 7.9 W / m^2 to 9.0 W / m^2 while the 5th percentile only ranges from 7.4 W / m^2 to 7.7 W / m^2 and the 95th percentile has a much wider range of 9.0 W / m^2 to 10.8 W / m^2.

The general agreement in total effective radiative forcing is reflected in the agreement of each of the key contributors to this total, namely CO_2 and aerosol effective radiative forcing (Figure 4 and Supplementary Figures S18 - S22). The key exceptions to this relate to aerosol effective radiative forcing, particularly in SCM4OPTv2.0 and OSCARv3.1. SCM4OPTv2.0's low effective radiative forcing is driven by its strong negative aerosol effective radiative forcing. This negative aerosol forcing is driven by SCM4OPTv2.0's inclusion of a climate feedback on aerosol effective radiative forcing, which makes their end of century aerosol effective radiative forcing 0.3 - 0.4 W / m^2 more negative the across the scenarios. This effect is absent in all other models except OSCARv3.1. However, the strong aerosol forcing is somewhat cancelled out by other factors in OSCARv3.1, for example its relatively large tropospheric ozone forcing (Supplementary Figure S23). As a result, OSCARv3.1's total effective radiative forcing is more in line with the other models.

4.2.3 Carbon Cycle

Moving beyond effective radiative forcing and its temperature response, we consider the behaviour of the carbon cycle in the different RCMs. For these comparisons, we use the emissions-driven ESM-SSPX-Y.Y set of scenarios, in which emissions of CO_2 are prescribed and atmospheric CO_2 concentrations are allowed to freely evolve (in contrast to the SSP experiments in which CO_2 concentrations are prescribed). There are considerable variations between the RCMs which submitted relevant results. However, these variations mainly occur in the width of the projections (i.e. the upper and lower percentiles) and the medians are surprisingly consistent across the RCMs which submitted data (Supplementary Figure S24, Supplementary Figure S25 and Figure 5).

In esm-SSP1-1.9 (Supplementary Figure S24, excluding Cicero-SCM because of its narrow range), the spread in median peak atmospheric CO_2 concentrations (430 ppm to 445 ppm) is smaller than the spread in 2081-2100 median concentrations (385 ppm to 405 ppm). In contrast, in esm-SSP1-2.6 (Supplementary Figure S25, again excluding Cicero-SCM), the spread in median peak atmospheric CO_2 concentrations (455 ppm to 480 ppm) is the same width as the spread in 2081-2100 median concentrations (25ppm, 430 ppm to 455 ppm). Under both scenarios, there are wide variances in percentile ranges across the models, with MAGICC7 showing largest uncertainty in 2081-2100 atmospheric CO_2 concentrations and FaIRv1.6 showing the least.

Next, we consider esm-SSP5-8.5, the only scenario with available CMIP6 Earth System Model results (Figure 5). Median atmospheric CO_2 concentrations range from 920 ppm to 1 000 ppm while 5th percentile and 95th percentile concentrations range from 800 ppm to 920 ppm and 1 020 ppm to 1 130 ppm respectively. MAGICC7 once again shows the largest uncertainties, but is more similar to the other RCMs than in the other scenarios. These comparisons highlight differences in the dynamics of the carbon cycle (and its feedbacks) in the various RCMs: uncertainties scale more quickly with temperature in MCE, FaIR1.6 and OSCARv3.1 than they do in MAGICC7.

Median atmospheric CO_2 projections from all of the RCMs lie within the plume of available CMIP6 results (Figure 5). FaIR1.6 lies at the top end of the CMIP6 plume,

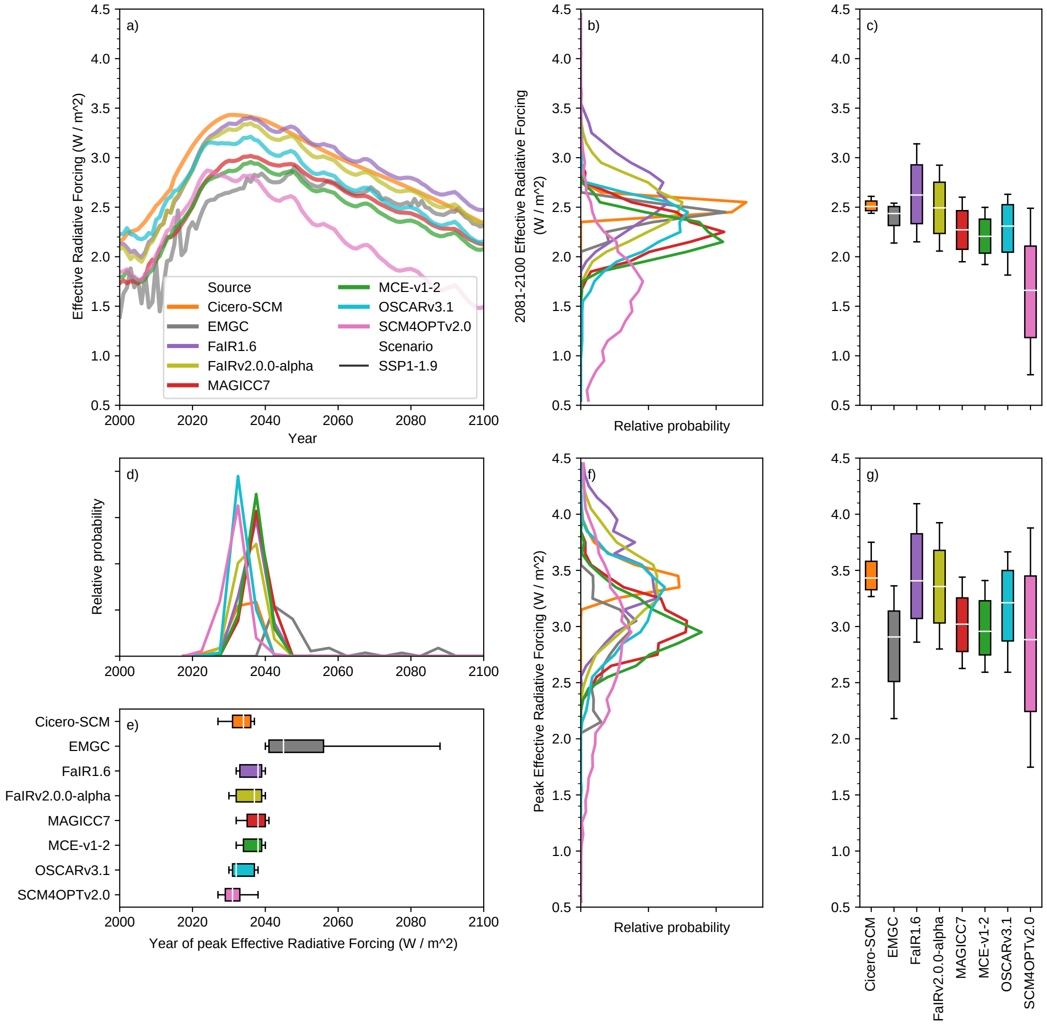

Figure 3. Effective radiative forcing under the very low-emissions SSP1-1.9 scenario. a) Median effective radiative forcing projections from 1995 to 2100 for each RCM; b) distribution of 2081-2100 mean effective radiative forcing from each RCM; c) very likely (whiskers), likely (box) and central (white line) 2081-2100 mean effective radiative forcing estimate from each RCM; d) as in b) except for the year in which effective radiative forcing peaks; e) as in c) except for the year in which effective radiative forcing peaks; f) as in b) except for the peak effective radiative forcing; g) as in c) except for the peak effective radiative forcing.

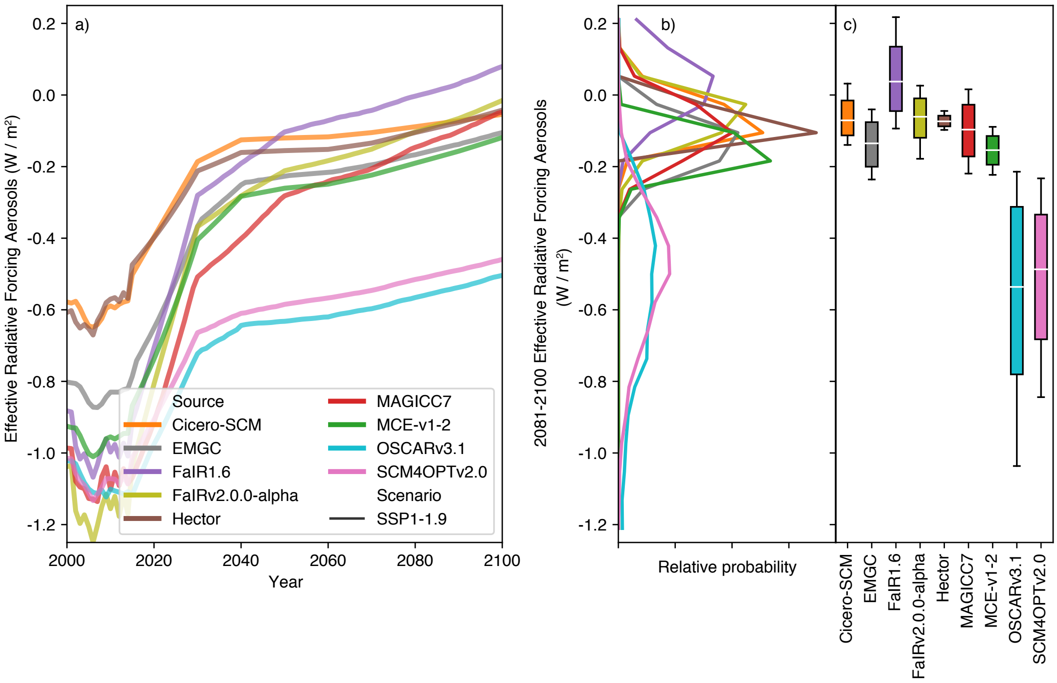

Figure 4. As in panels a), b) and c) of Figure 3, except for effective radiative forcing due to aerosols.

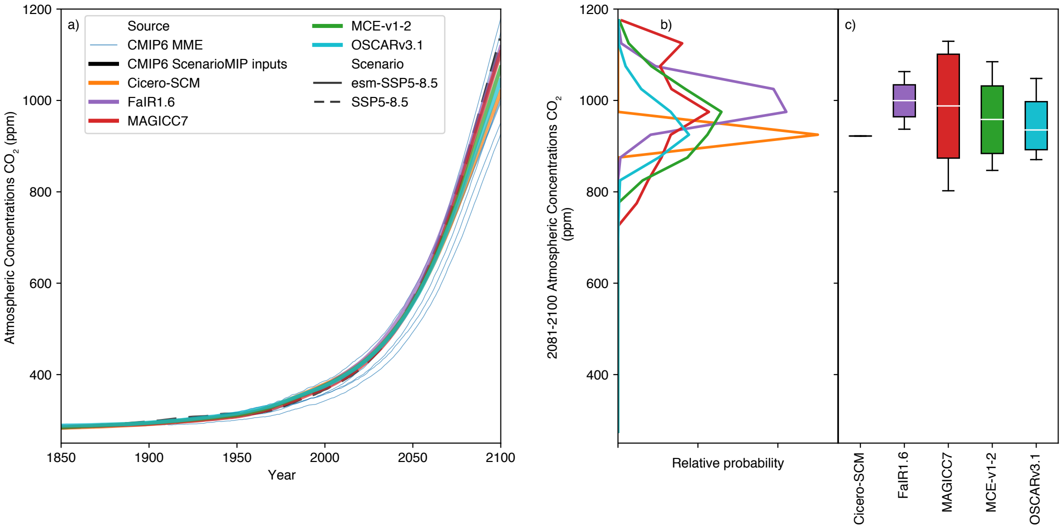

Figure 5. Atmospheric CO_2 concentration projections in the esm-SSP5-8.5 experiment. a) Atmospheric CO_2 concentration projections from 1995 to 2100. We show the median RCM projections (coloured lines), prescribed CMIP6 ScenarioMIP input concentrations from the SSP5-8.5 concentration-driven experiment (dashed black line) and available CMIP6 model projections (thin blue lines, we show the average of all available ensemble members for each CMIP6 model); b) distribution of 2081-2100 mean atmospheric CO_2 concentration projections from each RCM; c) very likely (whiskers), likely (box) and central (white line) 2081-2100 mean atmospheric CO_2 concentration projections estimate from each RCM. Note that FaIR1.6 data is taken from the esm-SSP5-8.5-allGHG simulations because esm-SSP5-8.5 simulations are not available. [TODO fix panel labels]

and its 5-95th range does not include low end CMIP6 results. In contrast, OSCARv3.1 lies at the bottom end of the CMIP6 plume, and its 5-95th range does not include high end CMIP6 results. MAGICC7 and MCE span the CMIP6 range, with MCE's range being almost exactly in line with the CMIP6 range whilst MAGICC7's projections are slightly wider than the CMIP6 range. Cicero-SCM does not include uncertainty in the carbon cycle, nor temperature feedbacks on the carbon cycle, hence produces only a single best-estimate projection.

4.2.4 All greenhouse gas emissions-driven runs

The final set of experiments we present are the experiments which are most relevant to WG3: all greenhouse gas emissions-driven runs. As discussed in Section 1, WG3 describes scenarios in terms of their emissions hence needs models which can run in a fully-emissions driven setup. The cost of running Earth System Models in such a setup is computationally prohibitive, hence there is a paucity of data against which to evaluate the projections of RCMs in such experiments. Nonetheless, here we present the results of such experiments in the hope that they will inspire further thinking into how to validate RCMs in this fully-coupled, all greenhouse gas emissions driven setup.

Only three models (Cicero-SCM, MAGICC7 and FaIR1.6) have submitted results for the all greenhouse gas emissions-driven scenarios. The MAGICC7 and FaIR1.6 models suggest that there is little difference between the concentration-driven and all greenhouse gas emissions-driven runs (Figure 6, Supplementary Figure S26 and Supplementary Figure S27). For both these models, the emissions-driven results have slightly lower temperature projections (both long-term and peak) and slightly earlier warming peaks, with slightly wider uncertainties. These differences are consistent with: a) their slightly lower median CO_2 concentrations in emissions-driven runs and b) the fact that emissions-driven runs introduce carbon cycle uncertainties into temperature projections, an uncertainty which is missing in concentration-driven runs. The Cicero-SCM results suggest a larger discrepancy between all greenhouse gas emissions-driven runs and the concentration-driven runs. In general, Cicero-SCM's warming projections are notably lower in emissions-driven runs, with the same uncertainty (Cicero-SCM does not include carbon cycle uncertainties or temperature feedbacks), reflecting their lower CO_2 concentration projections in emissions-driven runs.

4.3 Further Discussion

The results presented previously prompt consideration of a number of further points. Firstly, the assessment performed here provides a way to easily identify differences between an RCM's behaviour and the assessed range of a particular metric. Such differences are important to quantify, as they often point to a bias in the model's behaviour or setup. The quantification makes it possible for the users of these models to consider the impact of these biases on their own conclusions.

There are, however, cases where the issue lies in the combination of the proxy assessed ranges taken together, rather than in the models. In this study, we used a combination of ECS from the literature (based on multiple lines of evidence), TCR from constrained CMIP6 models and TCRE from unconstrained CMIP6 Earth System Models. This combination has likely resulted in slight inconsistencies between these metrics as the metrics are sourced from various lines of evidence, yet are strongly interdependent. This potential inconsistency could in part explain our finding that the RCMs' TCR ranges are generally too high, while their TCRE ranges are generally too low. The inconsistency is further demonstrated by the fact that a) the realised warming fraction implied by our TCR and ECS distributions, i.e. the ratio between TCR and ECS, is around 0.5, at the low end of the assessment by Millar et al. (2015) and b) the airborne fraction implied by our TCR and TCRE assessment is around 0.65, at the high-end of the CMIP5 and

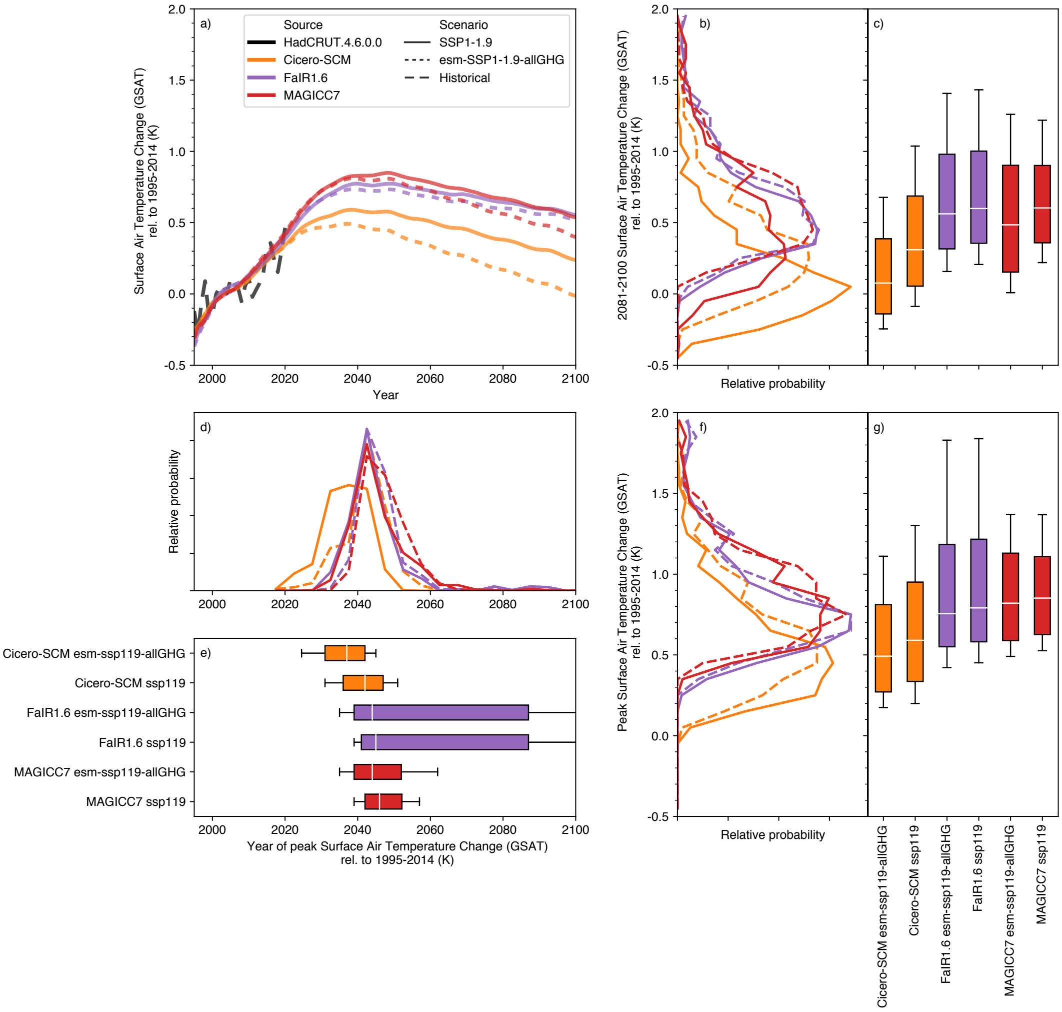

Figure 6. Surface air temperature (also referred to as global-mean surface air temperature, GSAT) change in the concentration-driven SSP1-1.9 experiment and the all greenhouse gas emissions driven esm-SSP1-1.9-allGHG experiment. a) GSAT projections from 1995 to 2100. We show the median RCM projections (coloured lines) for the concentration-driven experiment (solid) and all greenhouse gas emissions driven experiment (dashed) as well as observations up to 2019 (dashed black line); b) distribution of 2081-2100 mean GSAT for each scenario from each RCM; c) very likely (whiskers), likely (box) and central (white line) 2081-2100 mean GSAT estimate for each scenario from each RCM; d) as in b) except for the year in which GSAT peaks; e) as in c) except for the year in which GSAT peaks; f) as in b) except for the peak GSAT; g) as in c) except for the peak GSAT. All results are shown relative to the 1995-2014 reference period.

CMIP6 range quantified by Arora et al. (2020). Identifying such inconsistencies is a useful secondary benefit of exercises such as the one performed here.

Next, while they are a useful way of quickly visualising a model's agreement with the (here proxy) assessed ranges, summary tables of the form of Table 3 hide the full story. Specifically, for timeseries based variables, assessed ranges can only consider the trend or change between specific timepoints and don't consider the entire timeseries as a whole.

Not considering the entire timeseries can lead to problematic interpretations of the agreement between a model and the assessment. A clear example here is historical surface air ocean blended temperature change. In our proxy assessment, we focussed on 2000-2019 warming relative to the 1961-1990 reference period. On this measure, many of the RCMs showed poor agreement with the observations. However, the level of agreement is clearly reference period dependent (Figures 7a) and 7b)). In Figure 7a), which uses a 1961-1990 reference period, MAGICC7, MCE and OSCARv3.1 show the best agreement with observations (as also seen in Table 3). However, if we use a different reference period, e.g. 1850-1900 (Figures 7b)), that impression changes with Hector, MAGICC7, and OSCARv3.1 being the closest to observations in the recent period.

Considering the entire timeseries provides a more robust check on model behaviour. Fitting only to one evaluation and reference period can be achieved by slightly adjusting different model behaviour e.g. aerosol effective radiative forcing. However, if the entire timeseries are considered with multiple reference periods, such tuning quickly becomes impossible and the check provides detail into how well a model's dynamics are consistent with observations.

Moving away from evaluating the models, it is clear that historical performance alone does not determine a model's projections. For example, MAGICC7 and MCE have very similar fits to historical temperatures and historical effective radiative forcing yet have vastly different ECS and TCR distributions and make notably different projections about the magnitude, peak and timing of future warming. Investigating the extent to which the difference in ECS and TCR distributions could be rectified, without degrading the historical temperature simulations, is an area for future work. More generally, we find that higher ECS and TCR values lead to higher warming projections. Hector provides an exception to this trend, with relatively low temperature projections, especially in SSP1-1.9, despite its relatively high TCR. There is clear uncertainty in RCM projections, and it does not disappear even if the reduced complexity modelling groups all start with the same target ranges. In the strong mitigation scenarios (SSP1-1.9 and SSP1-2.6), the range in median warming across the RCMs is around 0.3°C and is much higher for the upper-end (95th percentile) of the range, being at least 1.0°C. In the context of international climate policy, even the relatively small deviations in median temperature projections presented here are not trivial. For a 1.5°C target, deviations of 0.3°C are roughly 60% of our remaining warming.

While historical performance alone does not determine a model's projections, the constraining process does have an impact on projections. This is most obvious when comparing the constrained RCM-based projections with the CMIP6-based projections (Figure 2, Supplementary Figure S13 and Supplementary Figure S14). Clearly, constraining the RCMs to match our proxy assessment across a range of metrics causes projections to be lower than the CMIP6 multi-model ensemble, perhaps because the high ECS seen in many CMIP6 models (Zelinka et al., 2020) is hard to reconcile with historical observations without a compensating strong negative aerosol forcing. This study lays the foundation for examining why this is the case in detail, using the comprehensive set of experiments (and possibly more) and data handling infrastructure implemented here.

There is another corollary from our finding that future warming diverges, even among a set of RCMs that share the same historical benchmarks: to extrapolate assessed warm-

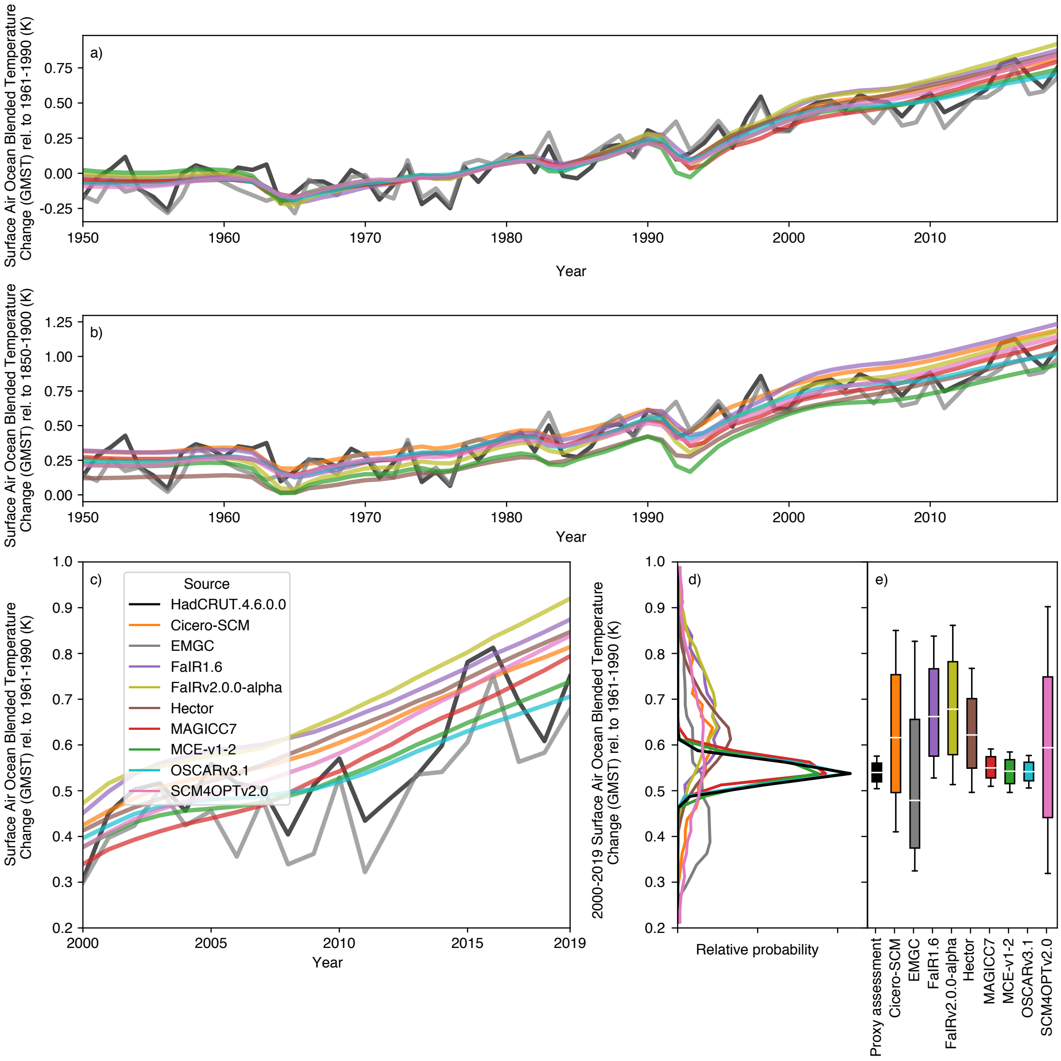

Figure 7. Historical surface air ocean blended temperature change (also referred to as global-mean surface temperature, GMST) from each RCM. We compare observations from HadCRUT4.6.0.0 (Morice et al., 2012) (solid black line) to the distribution from each RCM (coloured lines). All panels use 1961-1990 as the reference period, the same reference period as is used in our proxy assessed ranges, except b) which uses 1850-1900. a), b) median GMST from 1950 to 2019; c) median GMST from 2000 to 2019 (the proxy assessment period); d) distribution of 2000-2019 mean GMST from each RCM and the proxy assessed range; e) Very likely (whiskers), likely (box) and central (white line) estimate of 2000-2019 mean GMST from each RCM and the proxy assessed range. The historical simulation has been extended with SSP2-4.5 for the period 2015-2019.

ing ranges from one set of scenarios (e.g. the RCP or SSP-based scenarios) to a wider set of scenarios, it may be beneficial to include a benchmark of assessed future warming under the benchmark scenarios. Adding such a benchmark (to the historical observations, present-day assessments and idealised metrics used in this study) would highlight where future warming significantly diverges from wider understanding. Such quantifications could be key when assessing future projections under large sets of scenarios, like the WG3 scenario database climate assessment.

Nonetheless, deciding which projections are most sensible will never be an exact science. It is possible to make judgements about what is more reasonable based on the evaluation performed here, and to rule out clearly incorrect projections, but a definitive answer is impossible: we will not know which projections are correct until we get there, by which time it is too late for climate policy. Hence while it is important to continue to evaluate and improve our models to remove as many sources of error as possible, it is also important that research into decision making under uncertainty (e.g. Weaver et al., 2013; Dittrich et al., 2016) continues to develop and be used because the uncertainty in projections will not disappear anytime soon, never in fact.

Beyond the implications for policy, there are other scientific outcomes to consider. The first is the difference between these RCMs and the more comprehensive CMIP6 models. Here, the most obvious difference is the behaviour in high-warming scenarios. In such scenarios, MAGICC7 is a clear outlier from the rest of the RCMs, yet it appears to be the most 'CMIP6-like', showing sustained warming out to 2300 in SSP5-8.5 (Figure 2). This 'more CMIP6-like' impression is reinforced by the similarity between MAGICC7 and the CMIP6 models' relatively strong recovery in SSP1-2.6, something which is not as prominent in the other RCMs except for Cicero-SCM. In SSP1-2.6, MAGICC7 shows a similar peak median warming to FaIR1.6 and FaiRv2.0.0-alpha before exhibiting a stronger cooling trend than the other RCMs (with the exception of Cicero-SCM, Supplementary Figure S14). A likely explanation for the MAGICC7 'outlier' behaviour, particularly for the sustained warming seen in SSP5-8.5, is MAGICC7's state-dependent climate sensitivity, which arises from its calibration to CMIP6 models (Nicholls et al., 2020) and reflects the finding that CMIP models have state-dependent climate sensitivities (Rugenstein et al., 2020). It appears that the state-dependent climate sensitivity is a feature of MAGICC7 which is either missing or less prominent in the other RCMs.

We have limited our evaluation of the carbon-cycle behaviour to emissions-driven simulations. While this decision limits us to a relatively small set of CMIP6-comparison data (given that only few emissions-driven simulations (Jones et al., 2016) have been run by CMIP6 models), it provides the cleanest comparison between RCMs and CMIP6 models, given that many RCMs do not separate the land and ocean carbon pools. Using the concentration-driven simulations would allow us to evaluate the RCMs' land and ocean carbon cycles (for those RCMs which include such a distinction) under more varied scenarios. We reserve such evaluation for future work.

It is notable that the CMIP6 ScenarioMIP input concentrations are generally higher than the RCMs' medians in emissions-driven runs across all considered scenarios (Figure 5, Supplementary Figure S24 and Supplementary Figure S25). Emissions-driven scenario data from CMIP6 ESMs is almost exclusively related to the esm-SSP5-8.5 experiment. Hence while the trend appears to be that the prescribed SSP5-8.5 CMIP6 concentrations are at the high-end of the range compared to the esm-SSP5-8.5 CMIP6 ESM results, there is little data with which to determine whether the prescribed CO_2 concentrations in the low-emissions scenarios would be within the projected concentration change by emission-driven ESM models. In hindsight, the input atmospheric CO_2 concentrations used in the concentration-driven runs may turn out to be at the high-end of CMIP6 ESM results across a range of scenarios. Given that only one set of input concentrations can be used in CMIP6, it is not surprising that the CO_2 concentrations prescribed for CMIP6 experiments do not sit exactly in the middle of later emissions-driven runs (see further discussion in Section 4.3). The opposite was observed in CMIP5: the input CO_2 concentrations (derived with MAGICC6) were found to be in the lower-half of the CMIP5 emissions-driven runs that later emerged from the CMIP5 emission driven runs (Friedlingstein et al., 2014). Choosing a carbon cycle parameterisation more in line with the median of CMIP5 models appears to have lead to CO_2 concentrations which are now in the upper-half of CMIP6 ESM projections (Figure 5). Whenever a single estimate of the relationship between CO_2 emissions and concentrations is used, there is always the risk that it will not be the central estimate of the next generation of ESMs as our understanding of the carbon cycle improves and the ensembles of participating ESMs changes in each intercomparison phase. While this does not invalidate the design of concentration-driven experiments which are developed in this way, it must be kept in mind when relating emissions scenarios and the output of concentration-driven CMIP experiments.

Finally, we find that there is relatively little difference in climate projections between the concentration-driven experiments typical of CMIP and the emissions-driven experiments required by WG3. This finding provides confidence that validating RCMs using concentration-driven experiments covers the most important earth system uncertainties and features of the RCMs climate projections. However, this confidence is tempered by the sparsity of available emissions-driven CMIP6 ESM model output, particularly for mitigation scenarios. Given that all greenhouse gas emissions driven experiments should also include uncertainties from each non-CO_2 greenhouse gas cycle, it is somewhat surprising that the uncertainties in RCMs all greenhouse gas emissions driven experiment temperature projections are not wider. We suggest this could be explained in two ways: a) the uncertainties in non-CO_2 greenhouse gas cycles are relatively small hence don't add much to the uncertainty from the carbon cycle and temperature response to effective radiative forcing and/or b) the RCMs are underestimating the uncertainty in the relationship between non-CO_2 greenhouse gas emissions and changes in atmospheric concentrations.

5 Extensions

This exercise is a first step towards more comprehensive, routine evaluation of RCMs' probabilistic parameter ensembles and their corresponding projections. However, there is still much room for future work to improve on this study and the first phase of RCMIP. As a first suggestion, repeating this exercise with the assessed ranges from Working Group 1 of the Intergovernmental Panel on Climate Change's Sixth Assessment Report (due in mid 2021) would provide an evaluation of the extent to which RCMs can capture the latest international assessment of the scientific literature.

This future work could go beyond evaluation and also diagnose the root causes of differences between the models. Such an exercise could also provide insights into why the constrained RCMs' probabilistic distributions systematically lead to lower temperature projections than the CMIP6 multi-model ensemble (as discussed in Section 4.3).

Finally, given how RCMs are typically used by WG3, it appears that a truly thorough evaluation would need to consider a larger set of individual steps in the emissions-climate change cause-effect chain. While it is not completely clear to us which components would need to be considered (and which could be ignored), a first suggestion of important components is: the carbon cycle, other earth system feedbacks e.g. representation of permafrost, representation of aerosols, non-CO_2 greenhouse gas cycles, translation between changes in greenhouse gas concentrations and effective radiative forcing, ozone representation, land-use change albedo representation, temperature response to effective radiative forcing and all the feedbacks and interactions. To see the full picture, a broad range of literature would need to be considered as a validation source and a wide range of experiments, spanning historical, scenario-based and idealised experiments, would need to be performed. In performing a more thorough evaluation, an updated evaluation technique may be required. Specifically, using percentage differences from the assessed range will lead to problems when the assessed range is close to or spans zero. Hence, more sophisticated ways of evaluating the agreement between model results and assessed ranges may be required. For reasons of scope, we haven't achieved such a thorough evaluation here, but we hope that this work provides a basis upon which future work can aim for the lofty goal of more complete evaluation of all of the relevant parts of the climate system.

6 Conclusions

We have found that the best performing RCMs can match our proxy assessment across a range of climate metrics. However, no RCM matched the proxy assessment across all metrics. At the same time, all RCMs had at least one strength where they matched the proxy assessment well.

Our evaluation of probabilistic projections from RCMs provides a comparison where, for the first time, each reduced complexity modelling team knew the target distributions before developing and submitting their results. This exercise provides a unique insight into RCMs probabilistic parameter ensembles, specifically how they compare with the target distributions and their implications for climate projections across a range of climate variables and scenarios.

Notably, we found that agreement with the proxy assessment, i.e. past performance, did not determine future performance (i.e. projections) from the RCMs. Given the various model structure that the reduced complexity models employ, ranging from linearised impulse response functions to 50-layer ocean models, it is not surprising that models may diverge in scenarios that go significantly beyond the domain of the validation data. Adding constraints on future performance i.e. extending the domain of validation data (for example based on an independent assessment of warming in a limited subset of scenarios) would likely reduce the divergence. Deciding which projections are most likely to be correct will never be an exact science. While exercises such as the one performed here can provide helpful information about where the biases may lie, they cannot provide definitive answers about what the future holds. Those who use RCMs for climate projections should carefully consider how they're going to use the RCMs and how they're going to validate them before making conclusions about the implications of their projections.

In addition, we found that many of the RCMs did not reproduce the high, long-term warming seen in CMIP6 models under high-emissions scenarios. Given that the exception was MAGICC7, it appears that its state-dependent climate sensitivity is a key feature for replicating CMIP6-style high-warming responses. Beyond the question of temperature projections, we found that the prescribed CO_2 concentrations used in the CMIP6 SSP-based experiments are at the high-end of projections made with historically constrained carbon cycles. Finally, we observed that a change in reference period significantly altered how well some models agreed with observations, reinforcing the need to consider more than one reference period when evaluating models.

With sufficient validations, RCMs provide a unique synthesis tool to integrate the latest scientific understanding, including its uncertainties, along the complex cause-effect chain from emissions to global-mean temperatures. Integrating this understanding in an internally consistent RCM framework, with all the implicit cross-correlations, is our best method to inform decision-making and other scientific domains, for example the likelihood of exceeding a given global-mean temperature threshold under a specific emissions scenario. Further developing these tools opens vast opportunities to go beyond global-mean variables and temperature changes, and to robustly represent the complex science beneath.

Acknowledgments

Data and code to produce all the figures and tables can be obtained from `https://doi.org/10.5281/zenodo.4269711`.

ZN benefited from support provided by the ARC Centre of Excellence for Climate Extremes (CE170100023).

KD's, DW's, AS's, SJS's and BVW's participation was supported by the U.S. Department of Energy, Office of Science, as part of research in MultiSector Dynamics, Earth and Environmental System Modeling Program.

CJS was supported by a NERC/IIASA Collaborative Research Fellowship (NE/T009381/1).

References

Arora, V. K., Katavouta, A., Williams, R. G., Jones, C. D., Brovkin, V., Friedlingstein, P., … Ziehn, T. (2020). Carbon–concentration and carbon–climate feedbacks in cmip6 models and their comparison to cmip5 models. *Biogeosciences*, *17*(16), 4173–4222. Retrieved from `https://bg.copernicus.org/articles/17/4173/2020/` doi: 10.5194/bg-17-4173-2020

Canty, T., Mascioli, N. R., Smarte, M. D., & Salawitch, R. J. (2013). An empirical model of global climate – part 1: A critical evaluation of volcanic cooling. *Atmospheric Chemistry and Physics*, *13*(8), 3997–4031. Retrieved from `https://www.atmos-chem-phys.net/13/3997/2013/` doi: 10.5194/acp-13-3997-2013

Clarke, L., Jiang, K., Akimoto, K., Babiker, M., Blanford, G., Fisher-Vanden, K., … et al. (2014). Assessing transformation pathways. In O. Edenhofer et al. (Eds.), *Climate change 2014: Mitigation of climate change. contribution of working group iii to the fifth assessment report of the intergovernmental panel on climate change* (p. 413–510). Cambridge University Press.

Dittrich, R., Wreford, A., & Moran, D. (2016). A survey of decision-making approaches for climate change adaptation: Are robust methods the way forward? *Ecological Economics*, *122*, 79 - 89. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0921800915004887` doi: https://doi.org/10.1016/j.ecolecon.2015.12.006

Edwards, P. N. (2000). A brief history of atmospheric general circulation modeling. *International Geophysics Series*, *70*, 67–90.

Etminan, M., Myhre, G., Highwood, E. J., & Shine, K. P. (2016, dec). Radiative forcing of carbon dioxide, methane, and nitrous oxide: A significant revision of the methane radiative forcing. *Geophysical Research Letters*, *43*(24), 12,614–12,623. doi: 10.1002/2016gl071930

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development (Online)*, *9*(LLNL-JRNL-736881).

Forster, P. M., Richardson, T., Maycock, A. C., Smith, C. J., Samset, B. H., Myhre, G., … Schulz, M. (2016). Recommendations for diagnosing effective radiative forcing from climate models for cmip6. *Journal of Geophysical Research: Atmospheres*, *121*(20), 12,460-12,475. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016JD025320` doi: 10.1002/2016JD025320

Friedlingstein, P., Meinshausen, M., Arora, V. K., Jones, C. D., Anav, A., Liddicoat, S. K., & Knutti, R. (2014, 01). Uncertainties in CMIP5 Climate Projections due to Carbon Cycle Feedbacks. *Journal of Climate*, *27*(2), 511-526. Retrieved from `https://doi.org/10.1175/JCLI-D-12-00579.1` doi:

10.1175/JCLI-D-12-00579.1

Gasser, T., Ciais, P., Boucher, O., Quilcaille, Y., Tortora, M., Bopp, L., & Hauglustaine, D. (2017). The compact earth system model oscar v2.2: description and first results. *Geoscientific Model Development*, *10*(1), 271–319. Retrieved from `https://gmd.copernicus.org/articles/10/271/2017/` doi: 10.5194/gmd-10-271-2017

Gasser, T., Crepin, L., Quilcaille, Y., Houghton, R. A., Ciais, P., & Obersteiner, M. (2020). Historical co2 emissions from land use and land cover change and their uncertainty. *Biogeosciences*, *17*(15), 4075–4101. Retrieved from `https://bg.copernicus.org/articles/17/4075/2020/` doi: 10.5194/bg-17-4075-2020

Gasser, T., Kechiar, M., Ciais, P., Burke, E. J., Kleinen, T., Zhu, D., … Obersteiner, M. (2018, Nov 01). Path-dependent reductions in co2 emission budgets caused by permafrost carbon release. *Nature Geoscience*, *11*(11), 830-835. Retrieved from `https://doi.org/10.1038/s41561-018-0227-0` doi: 10.1038/s41561-018-0227-0

Harmsen, M. J. H. M., van Vuuren, D. P., van den Berg, M., Hof, A. F., Hope, C., Krey, V., … Schaeffer, M. (2015, sep). How well do integrated assessment models represent non-CO2 radiative forcing? *Climatic Change*, *133*(4), 565–582. doi: 10.1007/s10584-015-1485-0

Haustein, K., Allen, M. R., Forster, P. M., Otto, F. E. L., Mitchell, D. M., Matthews, H. D., & Frame, D. J. (2017, Nov 13). A real-time global warming index. *Scientific Reports*, *7*(1), 15417. Retrieved from `https://doi.org/10.1038/s41598-017-14828-5` doi: 10.1038/s41598-017-14828-5

Hooss, G., Voss, R., Hasselmann, K., Maier-Reimer, E., & Joos, F. (2001, dec). A nonlinear impulse response model of the coupled carbon cycle-climate system (NICCS). *Climate Dynamics*, *18*(3-4), 189–202. doi: 10.1007/s003820100170

Hope, A. P., Canty, T. P., Salawitch, R. J., Tribett, W. R., & Bennett, B. F. (2017). Forecasting global warming [Book Section]. In *Paris climate agreement: Beacon of hope* (p. 51-114). Springer Climate.

Hope, A. P., McBride, L. A., Canty, T. P., Bennett, B. F., Tribett, W. R., & Salawitch, R. J. (2020). Examining the human influence on global climate using an empirical model. *Earth and Space Science Open Archive*, 79. Retrieved from `https://doi.org/10.1002/essoar.10504179.1` doi: 10.1002/essoar.10504179.1

Huppmann, D., Rogelj, J., Kriegler, E., Krey, V., & Riahi, K. (2018). A new scenario resource for integrated 1.5 °c research. *Nature Climate Change*, *8*(12), 1027–1030. doi: 10.1038/s41558-018-0317-4

Jones, C. D., Arora, V., Friedlingstein, P., Bopp, L., Brovkin, V., Dunne, J., … Zaehle, S. (2016). C4mip – the coupled climate–carbon cycle model intercomparison project: experimental protocol for cmip6. *Geoscientific Model Development*, *9*(8), 2853–2880. Retrieved from `https://gmd.copernicus.org/articles/9/2853/2016/` doi: 10.5194/gmd-9-2853-2016

Joos, F., Bruno, M., Fink, R., Siegenthaler, U., Stocker, T. F., Quéré, C. L., & Sarmiento, J. L. (1996). An efficient and accurate representation of complex oceanic and biospheric models of anthropogenic carbon uptake. *Tellus B: Chemical and Physical Meteorology*, *48*(3), 394-417. Retrieved from `https://doi.org/10.3402/tellusb.v48i3.15921` doi: 10.3402/tellusb.v48i3.15921

Leach, N. J., Nicholls, Z., Jenkins, S., Smith, C. J., Lynch, J., Cain, M., … Allen, M. R. (2020). Gir v1.0.0: a generalised impulse-response model for climate uncertainty and future scenario exploration. *Geoscientific Model Development Discussions*, *2020*, 1–29. Retrieved from `https://www.geosci-model-dev-discuss.net/gmd-2019-379/` doi: 10.5194/gmd-2019-379

McBride, L. A., Hope, A. P., Canty, T. P., Bennett, B. F., Tribett, W. R., & Salawitch, R. J. (2020). Comparison of cmip6 historical climate simulations and future projected warming to an empirical model of global climate. *Earth System Dynamics Discussions*, *2020*, 1–59. Retrieved from `https://esd.copernicus.org/preprints/esd-2020-67/` doi: 10.5194/esd-2020-67

Meinshausen, M., Meinshausen, N., Hare, W., Raper, S. C. B., Frieler, K., Knutti, R., … Allen, M. R. (2009, Apr 01). Greenhouse-gas emission targets for limiting global warming to 2 °c. *Nature*, *458*(7242), 1158-1162. Retrieved from `https://doi.org/10.1038/nature08017` doi: 10.1038/nature08017

Meinshausen, M., Nicholls, Z. R. J., Lewis, J., Gidden, M. J., Vogel, E., Freund, M., … Wang, R. H. J. (2020). The shared socio-economic pathway (ssp) greenhouse gas concentrations and their extensions to 2500. *Geoscientific Model Development*, *13*(8), 3571–3605. Retrieved from `https://gmd.copernicus.org/articles/13/3571/2020/` doi: 10.5194/gmd-13-3571-2020

Meinshausen, M., Raper, S. C. B., & Wigley, T. M. L. (2011). Emulating coupled atmosphere-ocean and carbon cycle models with a simpler model, MAGICC6 – part 1: Model description and calibration. *Atmospheric Chemistry and Physics*, *11*(4), 1417–1456. doi: 10.5194/acp-11-1417-2011

Millar, R. J., Nicholls, Z. R., Friedlingstein, P., & Allen, M. R. (2017, jun). A modified impulse-response representation of the global near-surface air temperature and atmospheric concentration response to carbon dioxide emissions. *Atmospheric Chemistry and Physics*, *17*(11), 7213–7228. doi: 10.5194/acp-17-7213-2017

Millar, R. J., Otto, A., Forster, P. M., Lowe, J. A., Ingram, W. J., & Allen, M. R. (2015, Jul 01). Model structure in observational constraints on transient climate response. *Climatic Change*, *131*(2), 199-211. Retrieved from `https://doi.org/10.1007/s10584-015-1384-4` doi: 10.1007/s10584-015-1384-4

Morice, C. P., Kennedy, J. J., Rayner, N. A., & Jones, P. D. (2012, apr). Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research: Atmospheres*, *117*(D8), n/a–n/a. doi: 10.1029/2011jd017187

Nicholls, Z. R. J., Meinshausen, M., Lewis, J., Gieseke, R., Dommenget, D., Dorheim, K., … Xie, Z. (2020). Reduced complexity model intercomparison project phase 1: introduction and evaluation of global-mean temperature response. *Geoscientific Model Development*, *13*(11), 5175–5190. Retrieved from `https://gmd.copernicus.org/articles/13/5175/2020/` doi: 10.5194/gmd-13-5175-2020

O'Neill, B. C., Tebaldi, C., van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., … Sanderson, B. M. (2016). The scenario model intercomparison project (scenariomip) for cmip6. *Geoscientific Model Development*, *9*(9), 3461–3482. Retrieved from `https://gmd.copernicus.org/articles/9/3461/2016/` doi: 10.5194/gmd-9-3461-2016

Rogelj, J., Meinshausen, M., & Knutti, R. (2012, Apr 01). Global warming under old and new scenarios using ipcc climate sensitivity range estimates. *Nature Climate Change*, *2*(4), 248-253. Retrieved from `https://doi.org/10.1038/nclimate1385` doi: 10.1038/nclimate1385

Rogelj, J., Shindell, D., Jiang, K., Fifita, S., Forster, P., Ginzburg, V., … et al. (2018). Mitigation pathways compatible with 1.5°c in the context of sustainable development. In G. Flato, J. Fuglestvedt, R. Mrabet, & R. Schaeffer (Eds.), *Global warming of 1.5 °c: an ipcc special report on the impacts of global warming of 1.5 °c above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty* (p. 93–174). IPCC/WMO. Retrieved from `http://www.ipcc.ch/report/sr15/`

Rugenstein, M., Bloch-Johnson, J., Gregory, J., Andrews, T., Mauritsen, T., Li, C.,

… Knutti, R. (2020). Equilibrium climate sensitivity estimated by equilibrating climate models. *Geophysical Research Letters*, *47*(4), e2019GL083898. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019GL083898` (e2019GL083898 10.1029/2019GL083898) doi: 10.1029/2019GL083898

Schlesinger, M. E., Jiang, X., & Charlson, R. J. (1992). Implication of anthropogenic atmospheric sulphate for the sensitivity of the climate system. In *Climate change and energy policy: Proceedings of the international conference on global climate change: Its mitigation through improved production and use of energy [rosen, l. and r. glasser (eds.)]. amer. inst. phys., new york, ny, usa* (pp. 75–108).

Schwarber, A. K., Smith, S. J., Hartin, C. A., Vega-Westhoff, B. A., & Sriver, R. (2019, nov). Evaluating climate emulation: fundamental impulse testing of simple climate models. *Earth System Dynamics*, *10*(4), 729–739. doi: 10.5194/esd-10-729-2019

Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., … Zelinka, M. D. (2020). An assessment of earth's climate sensitivity using multiple lines of evidence. *Reviews of Geophysics*, *58*(4), e2019RG000678. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019RG000678` (e2019RG000678 2019RG000678) doi: 10.1029/2019RG000678

Simmons, A. J., Berrisford, P., Dee, D. P., Hersbach, H., Hirahara, S., & Thépaut, J.-N. (2017). A reassessment of temperature variations and trends from global reanalyses and monthly surface climatological datasets. *Quarterly Journal of the Royal Meteorological Society*, *143*(702), 101-119. Retrieved from `https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2949` doi: 10.1002/qj.2949

Skeie, R. B., Berntsen, T., Aldrin, M., Holden, M., & Myhre, G. (2018, jun). Climate sensitivity estimates – sensitivity to radiative forcing time series and observational data. *Earth System Dynamics*, *9*(2), 879–894. doi: 10.5194/esd-9-879-2018

Skeie, R. B., Fuglestvedt, J., Berntsen, T., Peters, G. P., Andrew, R., Allen, M., & Kallbekken, S. (2017, feb). Perspective has a strong effect on the calculation of historical contributions to global warming. *Environmental Research Letters*, *12*(2), 024022. doi: 10.1088/1748-9326/aa5b0a

Smith, C. J., Forster, P. M., Allen, M., Leach, N., Millar, R. J., Passerello, G. A., & Regayre, L. A. (2018, jun). FAIR v1.3: a simple emissions-based impulse response and carbon cycle model. *Geoscientific Model Development*, *11*(6), 2273–2297. doi: 10.5194/gmd-11-2273-2018

Smith, C. J., Kramer, R. J., Myhre, G., Alterskjær, K., Collins, W., Sima, A., … Forster, P. M. (2020). Effective radiative forcing and adjustments in cmip6 models. *Atmospheric Chemistry and Physics*, *20*(16), 9591–9618. Retrieved from `https://acp.copernicus.org/articles/20/9591/2020/` doi: 10.5194/acp-20-9591-2020

Smith, C. J., Kramer, R. J., Myhre, G., Forster, P. M., Soden, B. J., Andrews, T., … Watson-Parris, D. (2018). Understanding rapid adjustments to diverse forcing agents. *Geophysical Research Letters*, *45*(21), 12,023-12,031. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018GL079826` doi: 10.1029/2018GL079826

Stainforth, D. A., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D. J., … Allen, M. R. (2005, Jan 01). Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, *433*(7024), 403-406. Retrieved from `https://doi.org/10.1038/nature03301` doi: 10.1038/nature03301

Su, X., Shiogama, H., Tanaka, K., Fujimori, S., Hasegawa, T., Hijioka, Y., … Liu, J. (2018). How do climate-related uncertainties influence 2 and 1.5° c pathways? *Sustainability science*, *13*(2), 291–299.

Su, X., Tachiiri, K., Tanaka, K., Watanabe, M., & Kawamiya, M. (2020). Source attributions of radiative forcing by regions, sectors, and climate forcers. *arXiv preprint arXiv:2009.07472*.

Su, X., Takahashi, K., Fujimori, S., Hasegawa, T., Tanaka, K., Kato, E., ... Emori, S. (2017). Emission pathways to achieve 2.0 c and 1.5 c climate targets. *Earth's Future*, *5*(6), 592–604.

Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., Lehner, F., & Knutti, R. (2020). Past warming trend constrains future warming in cmip6 models. *Science Advances*, *6*(12). Retrieved from `https://advances.sciencemag.org/content/6/12/eaaz9549` doi: 10.1126/sciadv.aaz9549

Tsutsui, J. (2017, jan). Quantification of temperature response to CO2 forcing in atmosphere–ocean general circulation models. *Climatic Change*, *140*(2), 287–305. doi: 10.1007/s10584-016-1832-9

Tsutsui, J. (2020, apr). Diagnosing transient response to CO2 forcing in coupled atmosphere-ocean model experiments using a climate model emulator. *Geophysical Research Letters*, *47*(7). doi: 10.1029/2019gl085844

Uhe, P., Otto, F. E., Rashid, M. M., & Wallom, D. C. (2016). Utilising amazon web services to provide an on demand urgent computing facility for climateprediction. net. In *2016 ieee 12th international conference on e-science (e-science)* (pp. 407–413).

van Vuuren, D. P., Lowe, J., Stehfest, E., Gohar, L., Hof, A. F., Hope, C., ... Plattner, G.-K. (2011, jan). How well do integrated assessment models simulate climate change? *Climatic Change*, *104*(2), 255–285. doi: 10.1007/s10584-009-9764-2

Vega-Westhoff, B., Sriver, R. L., Hartin, C. A., Wong, T. E., & Keller, K. (2019, jun). Impacts of observational constraints related to sea level on estimates of climate sensitivity. *Earth's Future*, *7*(6), 677–690. doi: 10.1029/2018ef001082

von Schuckmann, K., Cheng, L., Palmer, M. D., Hansen, J., Tassone, C., Aich, V., ... Wijffels, S. E. (2020). Heat stored in the earth system: where does the energy go? *Earth System Science Data*, *12*(3), 2013–2041. Retrieved from `https://essd.copernicus.org/articles/12/2013/2020/` doi: 10.5194/essd-12-2013-2020

Weaver, C. P., Lempert, R. J., Brown, C., Hall, J. A., Revell, D., & Sarewitz, D. (2013). Improving the contribution of climate model information to decision making: the value and demands of robust decision frameworks. *WIREs Climate Change*, *4*(1), 39-60. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/wcc.202` doi: 10.1002/wcc.202

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., ... Taylor, K. E. (2020). Causes of higher climate sensitivity in cmip6 models. *Geophysical Research Letters*, *47*(1), e2019GL085782. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019GL085782` (e2019GL085782 10.1029/2019GL085782) doi: 10.1029/2019GL085782