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

Zebedee R.J. Nicholls[1,1], Malte Alexander Meinshausen[1,1], Jared Lewis[2,2], Maisa Rojas Corradi[3,3], Kalyn Dorheim[4,4], Thomas Gasser[5,5], Robert Gieseke[6,6], Austin Patrick Hope[7,7], Nicholas Leach[8,8], Laura Anne McBride[7,7], Yann Quilcaille[9,9], Joeri Rogelj[10,10], Ross J. Salawitch[7,7], Bjørn Hallvard Samset[11,11], Marit Sandstad[12,12], Alexey N Shiklomanov[4,4], Ragnhild Bieltvedt Skeie[13,13], Christopher J Smith[14,14], Steven Smith[15,15], Xuanming SU[16,16], Junichi Tsutsui[17,17], Benjamin Aaron Vega-Westhoff[18,18], and Dawn Woodard[4,4]

[1]University of Melbourne
[2]Climate & Energy College, School of Earth Sciences, University of Melbourne
[3]University of Chile
[4]Pacific Northwest National Laboratory
[5]International Institute for Applied Systems Analysis (IIASA)
[6]Independent Researcher
[7]University of Maryland, College Park
[8]University of Oxford
[9]International Institute for Applied Systems Analysis
[10]Imperial College London
[11]CICERO Center for International Climate and Environmental Research - Oslo
[12]CICERO
[13]Center for International Climate and Environmental Research - Oslo (CICERO)
[14]University of Leeds
[15]Pacific Northwest National Laboratory (DOE)
[16]National Institute for Environmental Studies
[17]Central Research Institute of Electric Power Industry
[18]University of Illinois at Urbana-Champaign

November 30, 2022

Abstract

Over the last decades, climate science has evolved rapidly across multiple expert domains. Our best tools to capture state-of-the-art knowledge in an internally self-consistent modelling framework 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 ESMs alone. The tools of choice are instead 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 ESM experiments. Here we present Phase 2 of the Reduced Complexity Model Intercomparison Project (RCMIP Phase 2), the first comprehensive intercomparison of RCMs that are probabilistically calibrated with key benchmark ranges from specialised research communities. Unsurprisingly, but crucially, we find that models which have been constrained

to reflect the key benchmarks better reflect the key benchmarks. Under the low-emissions SSP1-1.9 scenario, across the RCMs, median peak warming projections range from 1.3 to 1.7{degree sign}C (relative to 1850-1900, using an observationally-based historical warming estimate of 0.8{degree sign}C between 1850-1900 and 1995-2014). Further developing methodologies to constrain these projection uncertainties seems paramount given the international community's goal to contain warming to below 1.5{degree sign}C above pre-industrial in the long-term. Our findings suggest that users of RCMs should carefully evaluate their RCM, specifically its skill against key benchmarks and consider the need to include projections benchmarks either from ESM results or other assessments to reduce divergence in future projections.

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. L. Woodard[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:

- Probabilistic global-mean temperature projections often use reduced complexity climate models (RCMs) because of their low computational cost
- We evaluate how well RCMs' probabilistic setups can synthesise knowledge from multiple research domains for policy relevant projections
- No RCM is able to capture all forcing, warming, heat uptake and carbon cycle metrics we evaluate, however some come close across a range

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

Abstract

Over the last decades, climate science has evolved rapidly across multiple expert domains. Our best tools to capture state-of-the-art knowledge in an internally self-consistent modelling framework 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 ESMs alone. The tools of choice are instead 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 ESM experiments. Here we present Phase 2 of the Reduced Complexity Model Intercomparison Project (RCMIP Phase 2), the first comprehensive intercomparison of RCMs that are probabilistically calibrated with key benchmark ranges from specialised research communities. Unsurprisingly, but crucially, we find that models which have been constrained to reflect the key benchmarks better reflect the key benchmarks. 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). Further developing methodologies to constrain these projection uncertainties seems paramount given the international community's goal to contain warming to below 1.5°C above pre-industrial in the long-term. Our findings suggest that users of RCMs should carefully evaluate their RCM, specifically its skill against key benchmarks and consider the need to include projections benchmarks either from ESM results or other assessments to reduce divergence in future projections.

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. 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 knowledge from specialised research communities. We find that the agreement between each RCM and the benchmarks varies, although the best performing models show good agreement across the majority of benchmarks. Under a very-low emissions scenario median peak warming projections range from 1.3 to 1.7°C (relative to 1850-1900, assuming historical warming of 0.8°C between 1850-1900 and 1995-2014). Investigating new ways to reduce these projection uncertainties seems paramount given the international community's goal to limit warming to below 1.5°C above pre-industrial in the long-term.

1 Introduction

Coupled Earth System Models (ESMs) have evolved for decades as primary climate research tools (Kawamiya et al., 2020). 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, large scenario 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., Murphy et al., 2014) that aim to explore a 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 the ability to completely explore 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. Earth System Models of Intermediate Complexity (EMICs) may initially appear to be another option. However, due to the process-based representations used by EMICs, their computational complexity and data requirements are still orders of magnitude greater than RCMs. As a result, even EMICs are not a feasible choice for the large-scale, probabilistic assessment discussed here.

Typically, RCMs achieve 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. In general, RCMs don't include representations of interannual variability, although the EMGC model (Table 1) is a clear exception to this rule. 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`).

Running WG3-type scenarios requires at least some representation of greenhouse gas cycles, atmospheric chemistry and dynamic vegetation modules. While some of the world's most comprehensive ESMs have the required components, they could not be used to sample scenario and parametric uncertainty for reasons of computational cost. The most comprehensive RCMs include parameterised representations of the required components (including feedbacks of climate on permafrost and wetland methane emissions), enabling the exploration of interacting uncertainties from multiple parts of the climate system in an internally consistent setup.

While RCMs do not include the detail of ESMs across the emissions-climate change cause-effect chain, 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. the degree to which RCMs can synthesise multiple lines of evidence across the emissions-climate change cause-effect chain.

Typically, RCMs attempt to perform this synthesis using probabilistic parameter ensembles (see also Section 3), which are distinct from the emulator mode in which RCMs can also be run (see Z. R. J. Nicholls et al. (2020) for a discussion of emulation with RCMs). These probabilistic parameter ensembles are derived based on knowledge of specific Earth system quantities drawn from multiple, often independent, research communities, 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 can then be used in a variety of applications, e.g. to assess the likelihood that different warming levels are reached under a specific emissions scenario (e.g. 50% and 66%) based on the combined available evidence. As a result of their probabilistic nature, the ensembles resulting from RCMs are conceptually different from an ensemble of multiple model outputs that has not been constructed with any relative probabilities in mind (such as those from CMIP6) taken without constraining or any other sort of post-processing.

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 goal of this knowledge transfer is consistency between WG3's scenario classification and 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 subsection 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.

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 independent radiative forcing, warming, heat uptake and concentration assessments simultaneously is of vital importance. Given that these assessments are independent, a single, internally consistent, model may not be able to capture them all. If RCMs are inherently biased in some way or they are unable to simultaneously capture the independent assessments, 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. While there have been studies with single models which choose parameter sets that match various assessments of ECS and TCR (e.g. Meinshausen et al., 2009; Rogelj et al., 2012) and Smith, Forster, et al. (2018) compared two models' probabilistic outputs, no previous study into RCM probabilistic distributions is of the same breadth.

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 of RCMIP (Z. R. J. 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). Models were invited to participate via an open invitation made available at `rcmip.org` and circulated via relevant researcher networks. All interested modelling teams were included. 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, the models and techniques cover (to the best of our knowledge) the full range of techniques seen 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, also referred to as a probabilistic distribution. As discussed in the introduction, a probabilistic setup means that each RCM is run with an ensemble of parameter configurations. Specifically, for a given experiment, each RCM is run multiple times, each time in a different configuration i.e. with 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). In each probabilistic run, the parameter ensemble is fixed i.e. the same set of parameter configurations will be used in each experiment. This choice ensures that the model outputs are deterministic, rather than including a random element due to e.g. sampling parameter values from a range or probability distribution for each run. 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. In this study, two models (MAGICC7 and MCE-v1-2) have used a common set of target assessed ranges, i.e. benchmarks, to derive their probabilistic distributions. For these models, we are able to rule out the choice of data as the cause of difference between these models. Accordingly, we can more clearly identify the importance of model structure and constraining technique for future projections.

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 (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. Notably, in this study, the proxy assessment does not include any future projections. While we examine future projections coming from the models, we do not explicitly compare them against future projections coming from another line of evidence because there is no obvious choice for

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	591 members sub-sampled from a posterior of 30 040 members to form a set that match the proxy assessment ocean heat content distribution while excluding parameter sets with unrealistic aerosol ERF or unrealistic 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); Z. R. J. Nicholls et al. (2020); Skeie et al. (2021)
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 likelihood of 2010-2019 level and rate of attributable warming, calculated using the Global Warming Index methodology (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)
MAGICCv7.5.1	7 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 v2.1	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	0.46		0.54		0.61
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	

such a line of evidence - apart from the 'assessed ranges' of SSP scenarios that will be communicated in the forthcoming IPCC report (but are not available for this study). As discussed in more detail in Section 4.3, the inclusion of future projections in the proxy assessment would narrow the range of model projections but any such narrowing should be carefully considered because - depending on the types of constraints - it may lead to underestimates of uncertainty.

In order to keep the study's scope manageable, our proxy assessment focuses on climate response parameters, with the carbon cycle examined only via the TCRE. We aim to perform a detailed analysis on carbon cycle response in the next phase of RCMIP.

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 only via the transient climate response to emissions (TCRE). TCRE is defined as the ratio of surface air temperature change 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 ECS assessment comes from a study which uses multiple lines of evidence, the TCR assessment is based on a constrained set of CMIP6 models and the TCRE assessment is based on unconstrained CMIP6 Earth System Models. We discuss the importance of this inconsistency and its consequences in Section 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 methane ERF. For aerosol and CO_2 ERF, we use the more recent work of Smith et al. (2020). For methane 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 (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) were available at the time of the study's inception.

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 previously discussed, this allowed some modelling teams (although crucially not all) to use the same 'constraining benchmarks' (with a number of different techniques being employed to consider the constraining benchmarks, see Table 1). We use these consistently constrained models to 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 inclusion of results from models which were not constrained using the same benchmarks allows us to quantify the importance of constraining when using reduced complexity climate models as integrators of knowledge.

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. OSCARv3.1), 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 Z. R. J. Nicholls et al. (2020), where required, models may use prescribed effective radiative forcing if 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 focuses on scenarios defined in terms of emissions alone. We use the data sources described in Z. R. J. 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 archived with Zenodo (Z. Nicholls & Lewis, 2021) - 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, Figure 1, Supplementary Table S1 and Supplementary Figures S1 - S9). In general, the RCMs capture the central assessed values better than the likely and very likely ranges. Historical warming, TCR and the TCRE are notable exceptions to this. For the TCR, the upper likely and very likely upper assessed values are captured by the RCMs about as well as the central value. For TCRE and historical warming, 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 ocean heat content and effective radiative forcing metrics is better captured by the RCMs than the other metrics. For the ocean heat content and 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 effective radiative forcing metrics, with median multi-model differences being up to 19% (aerosol effective radiative forcing).

For the other metrics (historical warming, ECS, TCR and TCRE), 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

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). Bold cells indicate that this model is within 20% of the proxy assessment at all likelihood levels for this metric. 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. 5^{th} percentile), 'll' (likely lower, 17^{th} percentile), 'c' (central, 50^{th} percentile), 'lu' (likely upper, 83^{th} percentile) and 'vlu' (very-likely upper, 95^{th} 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	7%		11%		25%
Equilibrium Climate Sensitivity	**16%**	**15%**	**12%**	**14%**	**20%**
Transient Climate Response	38%	18%	7%	4%	7%
Transient Climate Response to Emissions	**9%**	**11%**	**20%**	**19%**	**20%**
2014 CO_2 Effective Radiative Forcing		**5%**	**5%**	**1%**	
2014 Aerosol Effective Radiative Forcing		**14%**	**10%**	**19%**	
2018 Ocean Heat Content rel. to 1971		**1%**	**1%**	**16%**	
2011 CH_4 Effective Radiative Forcing		**4%**	**6%**	**18%**	
2011 N_2O Effective Radiative Forcing		**11%**	**2%**	**10%**	
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	**6%**		**12%**		**17%**	-30%		-11%		35%	14%		23%		37%
Equilibrium Climate Sensitivity	9%	4%	-2%	-12%	-22%	-43%	-42%	-38%	-28%	-12%	-16%	-11%	-3%	11%	32%
Transient Climate Response	33%	14%	1%	-6%	-12%						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		28%	37%	64%			**16%**	**16%**	**16%**			**2%**	**0%**	**-4%**	
2018 Ocean Heat Content rel. to 1971		**0%**	**-0%**	**-0%**			-9%	2%	26%			-0%	12%	24%	
2011 CH_4 Effective Radiative Forcing		12%	-12%	-27%			23%	-3%	-20%			**3%**	**-8%**	**-15%**	
2011 N_2O 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					MAGICCv7.5.1				
Assessed range	vll	ll	c	lu	vlu	vll	ll	c	lu	vlu	vll	ll	c	lu	vlu
2000-2019 GMST rel. to 1961-1990	11%		22%		38%	7%		16%		25%	**1%**		**1%**		**2%**
Equilibrium Climate Sensitivity	-22%	-15%	-2%	15%	31%	**-20%**	**-17%**	**-8%**	**0%**	**16%**	**-13%**	**-15%**	**-14%**	**-16%**	**-17%**
Transient Climate Response	26%	12%	4%	3%	4%	45%	25%	11%	3%	0%	32%	16%	7%	2%	0%
Transient Climate Response to Emissions	**-1%**	**-16%**	**-20%**	**-19%**	**-20%**						**9%**	**-2%**	**-1%**	**1%**	**-2%**
2014 CO_2 Effective Radiative Forcing		**3%**	**8%**	**14%**								**-2%**	**-1%**	**-1%**	
2014 Aerosol Effective Radiative Forcing		-12%	-13%	-27%			51%	44%	29%			**-1%**	**-7%**	**-19%**	
2018 Ocean Heat Content rel. to 1971												**1%**	**1%**	**1%**	
2011 CH_4 Effective Radiative Forcing		**5%**	**-1%**	**-5%**								**-1%**	**-3%**	**-3%**	
2011 N_2O Effective Radiative Forcing		**4%**	**-2%**	**-7%**								**1%**	**2%**	**2%**	
2011 F-Gases Effective Radiative Forcing		**3%**	**5%**	**7%**								**1%**	**1%**	**1%**	

Climate model	MCE-v1-2					OSCARv3.1					SCM4OPTv2.1				
Assessed range	vll	ll	c	lu	vlu	vll	ll	c	lu	vlu	vll	ll	c	lu	vlu
2000-2019 GMST rel. to 1961-1990	**-1%**		**3%**		**3%**	**3%**		**2%**		**0%**	-8%		10%		28%
Equilibrium Climate Sensitivity	-24%	-22%	-22%	-23%	-26%	**3%**	**-9%**	**-15%**	**-14%**	**-20%**	**13%**	**4%**	**12%**	**6%**	**-3%**
Transient Climate Response	23%	4%	-6%	-12%	-15%	44%	21%	-1%	-10%	-15%	61%	35%	9%	0%	-6%
Transient Climate Response to Emissions	0%	-18%	-23%	-23%	-27%	11%	-11%	-21%	-24%	-28%					
2014 CO_2 Effective Radiative Forcing		**-1%**	**-0%**	**1%**			**13%**	**6%**	**0%**			**6%**	**-0%**	**-6%**	
2014 Aerosol Effective Radiative Forcing		**10%**	**6%**	**-9%**			-21%	-10%	-1%			-14%	-10%	-23%	
2018 Ocean Heat Content rel. to 1971		**-1%**	**-0%**	**1%**			-20%	5%	28%			-21%	0%	16%	
2011 CH_4 Effective Radiative Forcing		**-1%**	**-3%**	**-4%**			5%	-17%	-32%			2%	-19%	-34%	
2011 N_2O Effective Radiative Forcing		**-1%**	**-0%**	**-1%**			26%	5%	-11%			21%	0%	-14%	
2011 F-Gases Effective Radiative Forcing		**0%**	**-0%**	**0%**			**15%**	**4%**	**-6%**			27%	14%	4%	

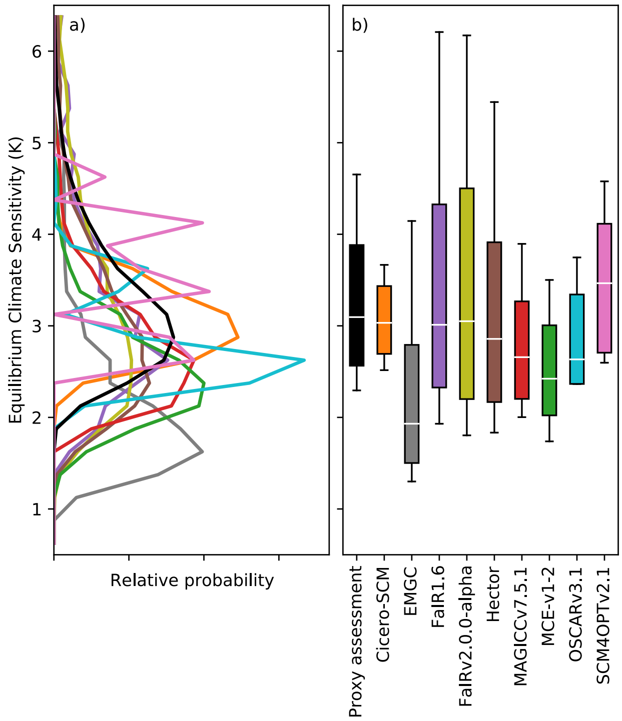

Figure 1. Distribution of Equilibrium Climate Sensitivity (ECS) from each RCM (coloured lines) and the proxy assessed range (solid black line). a) Distribution of ECS; b) Very likely (whiskers), likely (box) and central (white solid line) from the proxy assessment and each RCM.

median matches the very likely lower historical warming (rows labelled '2000-2019 GMST rel. to 1961-1990' in Table 3) to within 7%. However, the multi-model median differs from the central and very likely upper historical warming by 11% and 25%, indicating that the models are having greater difficulty capturing the upper-end warming estimates.

There is also significant spread in performance across the models. MAGICCv7.5.1 performs 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 ECS while MCE-v1-2 performs better than the multi-model median across all metrics and assessed ranges except for three metrics (ECS, TCR and TCRE). However, all RCMs had at least one metric where they matched the proxy assessment at all likelihood levels to within 20% (bolld cells in Table 3). For many applications, agreement to within 20% will be sufficient given the uncertainty associated with assessed ranges. However, for some applications, using an RCM's probabilistic distribution which has differences greater than 5-10% (for certain metrics) may be problematic as such differences could bias projections to an unacceptably large degree. For example, the WG3 classification of scenarios in terms of their peak warming levels should ideally be consistent with the range of evidence assessed in IPCC WG1. To have confidence that such an application is reflecting the WG1 assessment, the RCMs should be within 5-10% of the assessed results (particularly for any future warming assessment).

When interpreting these results it is vital to keep in mind that, for some models, the same benchmarks are used to both constrain and evaluate the models. The reason for this choice is that we are evaluating the ability of the models to act as integrators of knowledge i.e. to simultaneously capture all the independent assessments (see also discussion in Section 1). We are not attempting to do a calibration followed by an out-of-sample evaluation, instead we are looking at how well the RCMs can act as integrators of knowledge.

As a result, it is not so surprising that the models which calibrated to the benchmarks, specifically MAGICC7 and MCE-v1-2, better reflect the benchmarks during the evaluation phase. However, the results presented here highlight just how important it is to calibrate if the model is to be used as an integrator of knowledge. If the goal is an integrator of knowledge which reflects key benchmarks, our results suggest that models which are calibrated will perform better.

4.2 Projection results

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.

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 0.4°C across the models (from Hector with 0.3°C of warming to SCM4OPTv2.1 with 0.7°C, Figure 2 a)-c)). Variations in 5^{th} percentile warming show a similar range, from 0.0°C to 0.4°C. In contrast, upper-end, 95^{th} percentile warming shows far greater variation, from 0.8°C to 1.9°C. 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} percentiles. We discuss reasons for this difference in Section 4.3.

A slightly smaller spread is seen in peak temperature (Figure 2 f)-g)). Across the RCM ensemble, SSP1-1.9 median peak warming ranges from 0.55°C to 0.8°C while the 5^{th} and 95^{th} percentiles range from 0.3°C to 0.7°C and 0.9°C to 2.0°C, respectively. The year of peak warming shows much more variation, particularly at the upper end (Figure 2 d)-e)). While the median peak year is fairly consistent across the RCMs' ensembles, around 2045 (although SCM4OPTv2.1's 2055 peak is a clear outlier), and the 5^{th} percentile peak year varies from 2030 to 2040, the 95^{th} percentile varies from 2050 to beyond the end of this century. In the EMGC, FaIR1.6 and FaIRv2.0.0-alpha probabilistic distributions, there is a significant area of parameter space which results in ongoing warming even after CO_2 emissions have reached net zero. These models also drive the spread in end of century temperature projections, particularly in the 95^{th} percentile (Figure 2 b)-c)).

In the SSP1-2.6 scenario (Supplementary Figure S10), median peak warming ranges from 0.65-1.1°C (0.1-0.3°C higher than in SSP1-1.9). Median end of century warming (relative to 1995-2014) ranges from 0.6°C to 1.0°C. End of century 5^{th} percentile warming ranges from 0.2°C to 0.8°C and 95^{th} percentile warming ranges from 1.2°C to 2.0°C. As in SSP1-1.9, a number of CMIP6 model projections lie above the upper end of the constrained RCMs.

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 5^{th} percentile and median are fairly consistent (ranging from 0.3°C to 0.9°C and 0.65°C to 1.1°C respectively). However, 95^{th} percentile projections vary from 1.2°C to 2.8°C. The divergence in upper-end warming between SSP1-2.6 and SSP1-1.9 is driven by FaIR1.6, and appears to be the result of persistent warming after CO_2 emissions reach net zero given that its 83^{rd} 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

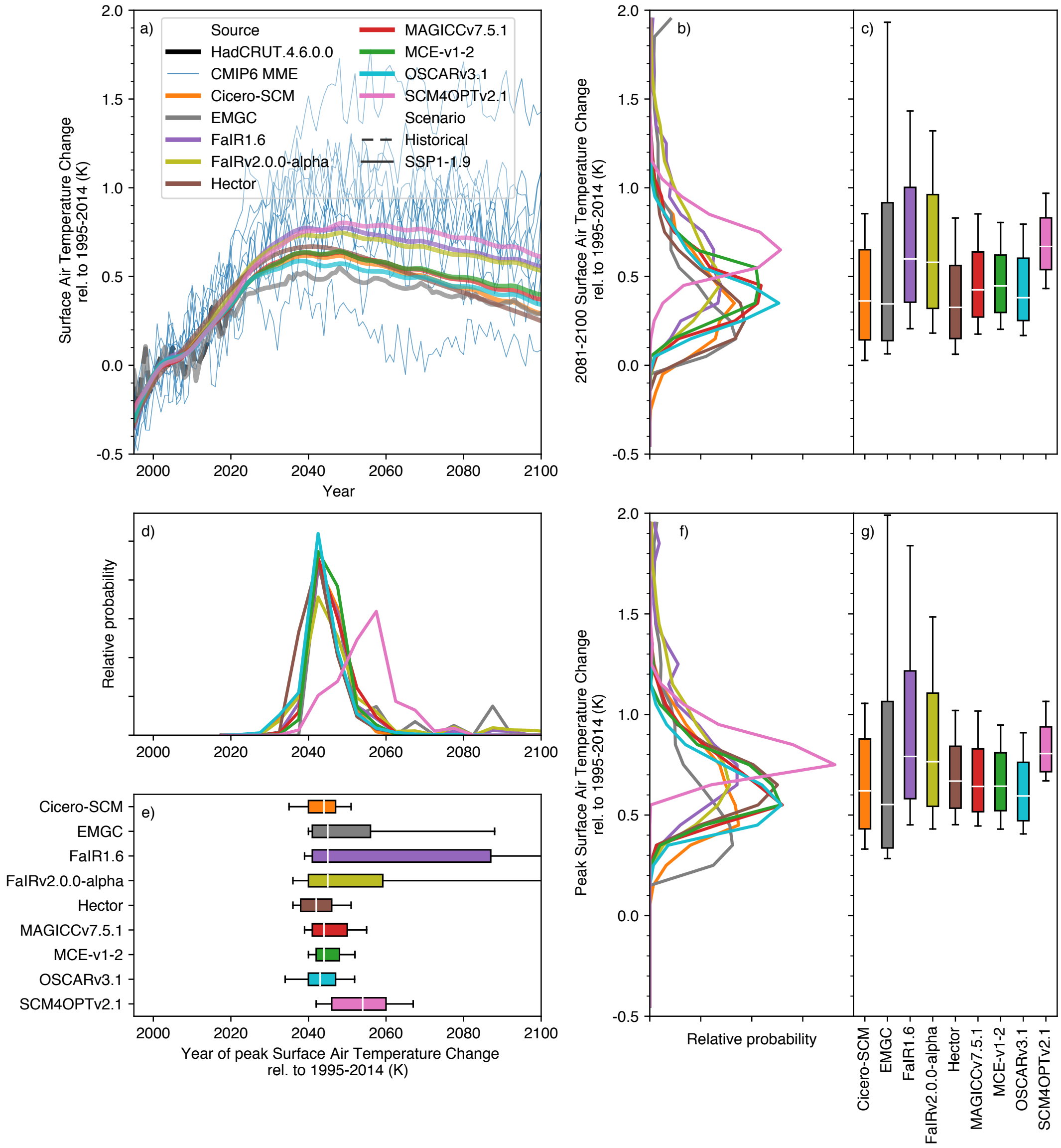

Figure 2. Surface air temperature (also referred to as global-mean surface air temperature, GSAT) change under the very low-emissions SSP1-1.9 scenario. a) GSAT projections from 1995 to 2100. We show the median RCM projections (coloured lines), GMST observations from HadCRUT4.6.0.0 (Morice et al., 2012) up to 2019 (dashed black line) and CMIP6 model projections (thin blue lines, we show a single ensemble member for each CMIP6 model to preserve the CMIP6 models' natural variability signal); 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.

median (ranging from 2065 to 2075). Once again, the 5^{th} percentile (ranging from 2050 to 2060) shows a much smaller spread across the models than the 95^{th} percentile (ranging from 2075 to beyond the end of the 21^{st} Century).

The warmest RCMs in mitigation scenarios are also the warmest under the high-emissions, SSP5-8.5, scenario (Supplementary Figure S11). The exceptions to this are MAGICC7, which is one of the warmest models in SSP5-8.5 even though it was around the median in mitigation scenarios, and SCM4OPTv2.1, which was the warmest model in mitigation scenarios but is slightly cooler than the warmest models in SSP5-8.5. Under SSP5-8.5, median end of century warming ranges from 2.5°C to 3.6°C across the RCMs. Unlike the mitigation scenarios, there is a similar level of disagreement in 5^{th} and 95^{th} percentile warming, with the 5^{th} percentile ranging from 1.7°C to 3.1°C and the 95^{th} percentile ranging from 3.8°C to 5.4°C. The RCMs all make future warming projections in the lower-half of the CMIP6 multi-model ensemble. Such a difference is largely explained by the constraints applied to the RCMs (see discussion in Section 4.3).

If we consider long-term (2250-2300) warming under the SSP5-8.5 scenario (Figure 3, see Supplementary Figure S12 and Supplementary Figure S13 for long-term warming under SSP1-1.9 and SSP1-2.6 respectively), the difference between RCMs and CMIP6 is even clearer (although the few CMIP6 models which have run the SSP5-8.5 extension are all at or above the median of the CMIP6 multi-model ensemble in 2100). On these timescales, MAGICC7 is clearly the warmest model, despite having slightly lower long-term effective radiative forcing than FaIR1.6, FaIR-v2.0.0-alpha and MCE-v1-2 (Supplementary Figure S14). There is a significant spread in long-term projections across the RCMs, with the median ranging from 4.5°C to 8.0°C, 5^{th} percentile from 3°C (ignoring SCM4OPTv2.1 as an outlier) to 5.8°C and 95^{th} from 7.8°C to 12.3°C. Even these upper end projections are well below the highest CMIP6 projections, which reach over 16°C of global-mean warming (again, likely due to constraining, see discussion in Section 4.3). 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 less variance in end of century total effective radiative forcing projections (Figure 4, Supplementary Figure S15 and Supplementary Figure S16). 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. The 5^{th} percentile ranges from 1.8 W / m^2 to 2.1 W / m^2 across the models (excluding CICERO-SCM which has an extremely narrow range). The spread is larger for the 95^{th} percentile, which ranges from 2.4 W / m^2 to 3.2 W / m^2. This pattern, of uncertainty being higher for upper percentiles than lower percentiles, is seen across other key scenarios and highlights that the high-end effective radiative forcing projections are much more uncertain than the best case and low-end effective radiative forcing projections.

In SSP1-2.6 (Supplementary Figure S15, once again excluding 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 5^{th} percentile only ranges from 2.4 W / m^2 to 2.7 W / m^2 and the 95^{th} percentile has a much wider range of 3.1 W / m^2 to 4.1 W / m^2. Under SSP5-8.5 (Supplementary Figure S16, excluding EMGC and CICERO-SCM as outliers), median 2081-2100 total effective radiative forcing ranges from 8.0 W / m^2 to 9.3 W / m^2 while the 5^{th} percentile only ranges from 7.4 W / m^2 to 7.8 W / m^2 and the 95^{th} percentile has a much wider range of 8.4 W / m^2 to 11.0 W / m^2.

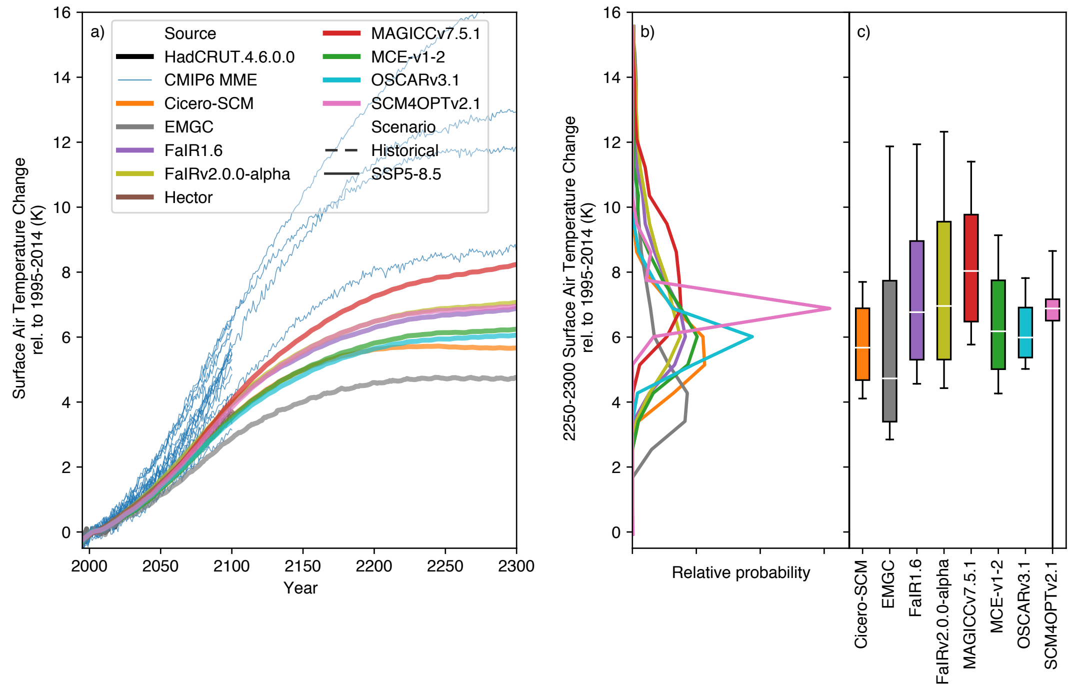

Figure 3. Long-term surface air temperature (also referred to as global-mean surface air temperature, GSAT) change under the high-emissions SSP5-8.5 scenario. a) GSAT projections from 1995 to 2300. We show the median RCM projections (coloured lines), GMST observations from (Morice et al., 2012) up to 2019 (dashed black line) and available CMIP6 model projections (thin blue lines, we show a single ensemble member for each CMIP6 model to preserve the CMIP6 models' natural variability signal); 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.

The approximate 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 5 and Supplementary Figures S17 - S29, which also show ERF output up to the year 2300). The key exceptions to this are SCM4OPTv2.1 and OSCARv3.1's aerosol effective radiative forcing. This negative aerosol forcing is driven by SCM4OPTv2.1 and OSCARv3.1's inclusion of a climate feedback on aerosol effective radiative forcing. The climate feedback makes their median end of century aerosol effective radiative forcing 0.3 - 0.8 W / m^2 more negative than other RCMs across the scenarios, although the effect is stronger in OSCARv3.1 than in SCM4OPTv2.1. The strong aerosol forcing is somewhat compensated by other forcing agents although both these models have long-term ERF which is at the low end of the RCM ensemble under SSP5-8.5 (Supplementary Figure S14). The different aerosol ERF parameterisations warrant further attention, particularly because models without this aerosol ERF - climate feedback may be underestimating the spread in future temperature projections.

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. Clearly, the analysis presented here covers only a limited subset of the full range of carbon cycle behaviour and metrics. The analysis is intended to highlight variance in carbon cycle behaviour across the RCMs, providing the motivation for a more detailed future analysis. 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 (Supplementary Figure S30, Supplementary Figure S31 and Figure 6). In esm-SSP1-1.9 (Supplementary Figure S30, excluding CICERO-SCM because of its narrow range), the spread in median peak atmospheric CO_2 concentrations (430 ppm to 450 ppm) is similar to the spread in 2081-2100 median concentrations (385 ppm to 410 ppm). Similarly, in esm-SSP1-2.6 (Supplementary Figure S31, again excluding CICERO-SCM), the spread in median peak atmospheric CO_2 concentrations (450 ppm to 480 ppm) shows a spread similar to the spread in 2081-2100 median concentrations (430 ppm to 460 ppm). Under both scenarios, there are wide variances in percentile ranges across the models, with MAGICC7 showing the largest uncertainty in 2081-2100 atmospheric CO_2 concentrations and SCM4OPTv2.1 showing the least (arguably, this model's range is overly confident). The considerable spread in projections from the models highlights the importance of carbon cycle uncertainty for emissions-driven projections. The spread reinforces the need for a detailed study into available techniques for evaluating and potentially constraining carbon cycle behaviour. Such a study would provide information about whether any of these projections can be ruled out based on other lines of evidence.

Next, we consider esm-SSP5-8.5, the only scenario with available CMIP6 Earth System Model results (Figure 6). Median 2081-2100 atmospheric CO_2 concentrations range from 920 ppm to 1 000 ppm while 5th percentile and 95th percentile concentrations range from 800 ppm to 930 ppm and 910 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 widen to a greater extent in higher-warming scenarios in FaIR1.6. FaIRv2.0.0-alpha, MCE-v1-2, OSCARv3.1 and SCM4OPTv2.1 compared to MAGICC7.

Median atmospheric CO_2 projections from all of the RCMs lie within the plume of available CMIP6 results (Figure 6). FaIR1.6 lies at the top end of the CMIP6 plume, and its 5-95th range does not include low end CMIP6 results. In contrast, SCM4OPTv2.1

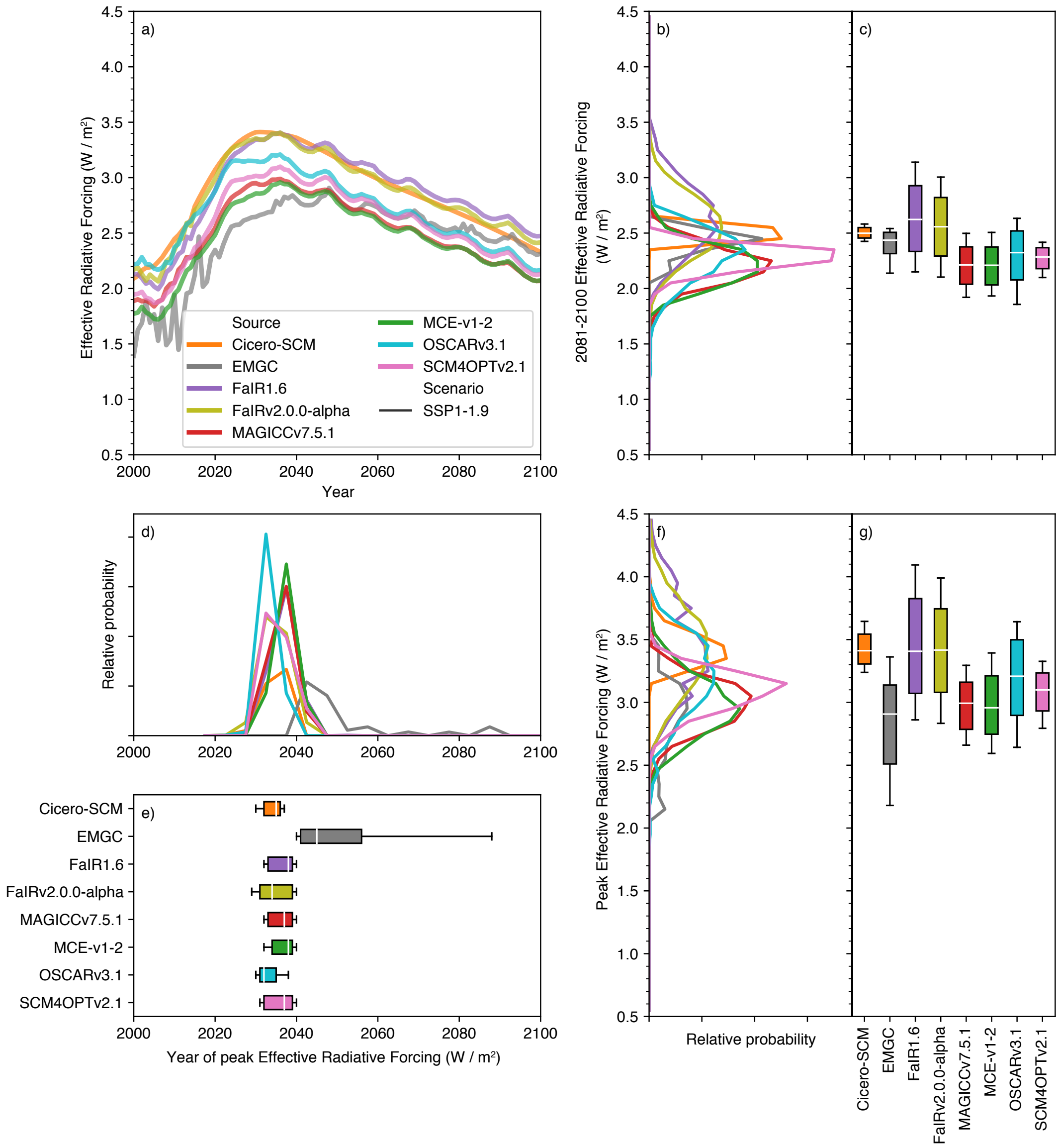

Figure 4. Effective radiative forcing under the very low-emissions SSP1-1.9 scenario. a) Median effective radiative forcing projections from 1995 to 2100 for each RCM; b) distribution of 2081-2100 mean effective radiative forcing from each RCM; c) very likely (whiskers), likely (box) and central (white 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.

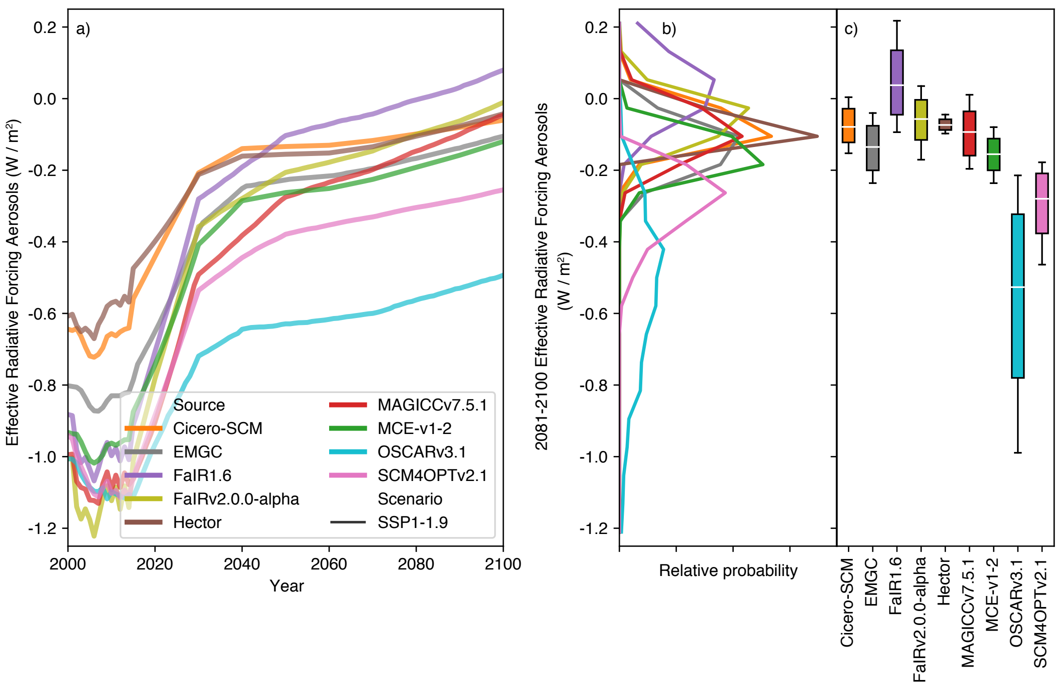

Figure 5. As in panels a), b) and c) of Figure 4, except for effective radiative forcing due to aerosols.

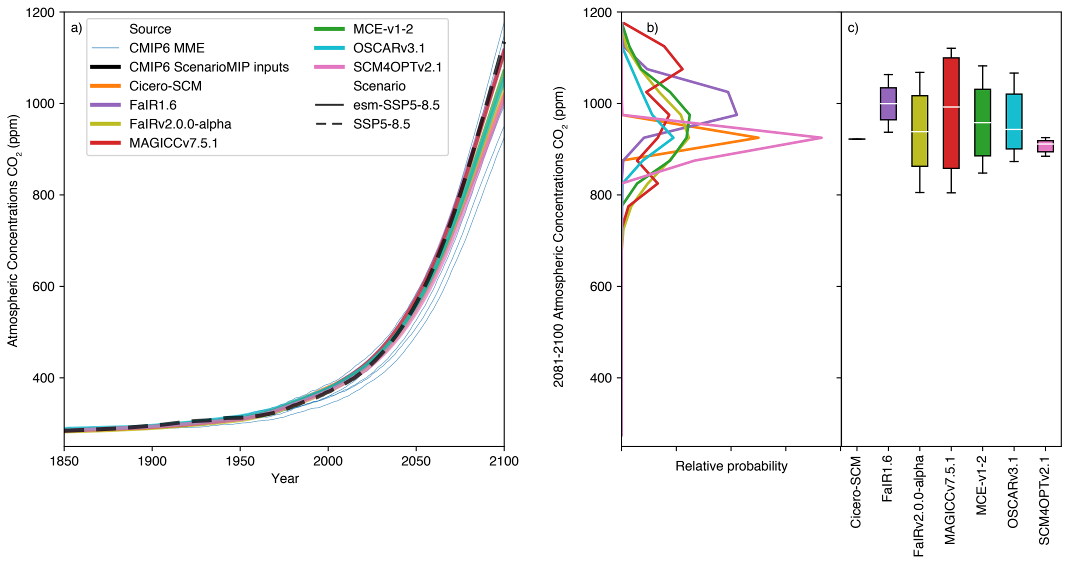

Figure 6. Atmospheric CO_2 concentration projections in the esm-SSP5-8.5 experiment. a) Atmospheric CO_2 concentration projections from 1995 to 2100. We show the median RCM projections (coloured lines), prescribed CMIP6 ScenarioMIP input concentrations from the SSP5-8.5 concentration-driven experiment (dashed black line) and available CMIP6 model projections (thin blue lines, we show a single ensemble member for each CMIP6 model to preserve the CMIP6 models' natural variability signal); 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.

lies at the bottom end of the CMIP6 plume. FaIR-v2.0.0-alpha, MAGICC7, MCE-v1-2 and OSCARv3.1 approximately span the CMIP6 range, with FaIR-v2.0.0-alpha's and MCE-v1-2's ranges being almost exactly in line with the CMIP6 range whilst MAGICC7's projections are slightly wider than the CMIP6 range and OSCARv3.1's projections are slightly narrower 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.

Despite the limits of our carbon cycle evaluation, it is notable that the CMIP6 ScenarioMIP input concentrations are generally higher than the RCMs' medians in emissions-driven runs across all considered scenarios. Emissions-driven scenario data from CMIP6 ESMs is almost exclusively related to the esm-SSP5-8.5 experiment. Hence, while the pattern 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. 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 emissions-driven runs (Friedlingstein et al., 2014). The CMIP6 concentrations were derived using an alpha version of MAGICC7, calibrated to approximately the median of the CMIP5 ESM carbon cycle responses with the inclusion of permafrost CO_2 and methane feedbacks (Meinshausen et al., 2020). 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 6). 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.

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 ESMs for a large number of scenarios and parameter configurations in such a setup is computationally prohibitive (and few ESMs include key feedbacks such as methane permafrost and wetland emissions), 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 efforts into how to validate RCMs in this fully-coupled, all greenhouse gas emissions driven setup.

Five models (CICERO-SCM, FaIR1.6, FaIRv2.0.0-alpha, MAGICC7 and SCM4OPTv2.1) have submitted results for the all greenhouse gas emissions-driven scenarios. The results suggest that the all greenhouse gas emissions-driven runs are cooler and peak earlier than the concentration-driven runs (Figure 7, Supplementary Figure S32 and Supplementary Figure S33). However, the magnitude of the difference varies across the models. For median projections, MAGICC7 suggests the smallest difference between concentration-driven and all greenhouse gas emissions-driven runs while CICERO-SCM and SCM4OPTv2.1 imply differences of up to 0.3°C for peak and 2081-2100 warming and a peak in warming up to ten years earlier. The range of projections in the all greenhouse gas emissions-driven runs are generally about the same or slightly wider than in the concentration-driven runs, with MAGICC7 showing the largest increase in projection ranges.

The lower-warming and wider projection ranges seen in all greenhouse gas emissions-driven runs are consistent with two other bits of knowledge. The first is that median CO_2 concentrations are lower in all greenhouse gas emissions-driven runs than in concentration-driven runs (Section 4.2.3). The second is that carbon cycle and other greenhouse gas cycle uncertainties are included in temperature projections in all greenhouse gas emissions-driven runs, whilst these uncertainties are missing in concentration-driven runs. The difference between the all greenhouse gas emissions-driven runs and concentration-driven runs reinforces the need for further consideration of RCM behaviour beyond the climate response to ERF.

4.3 Further Discussion

Our results 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 can reveal biases in a probabilistic distribution. The quantification makes it possible for the users of these distributions to identify where the biases might impact 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 probabilistic distributions. 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 is likely to be slightly inconsistent. Unfortunately, inconsistency between metric values is an inevitable risk of using independent lines of evidence. The potential inconsistency could in part explain our finding that the RCMs' TCR ranges are generally too high, while their ECS and TCRE ranges are generally too low. To explain the inconsistency in more detail, firstly consider the ratio between TCR and ECS i.e. the realised warming fraction. The realised warming fraction implied by our TCR and ECS distributions is around 0.5. This is at the low end of the assessment by Millar et al. (2015). Hence, it can be argued that greater consistency within the proxy assessment would be achieved if either our proxy assessed TCR values were larger, or our proxy assessed ECS values were smaller. Similarly, the airborne fraction implied by our TCR and TCRE assessment is around 0.65. This is at the high-end of the CMIP5 and CMIP6 range quantified by Arora et al. (2020). Once again, it can be argued that greater consistency within the proxy assessment would be achieved if either our proxy assessed TCR values were larger, or our proxy assessed TCRE values were smaller. 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 reference periods 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 were too warm compared to observations. However, the level of agreement is clearly reference period dependent (Figures 8a) and 8b)). In Figure 8a), which uses a 1961-1990

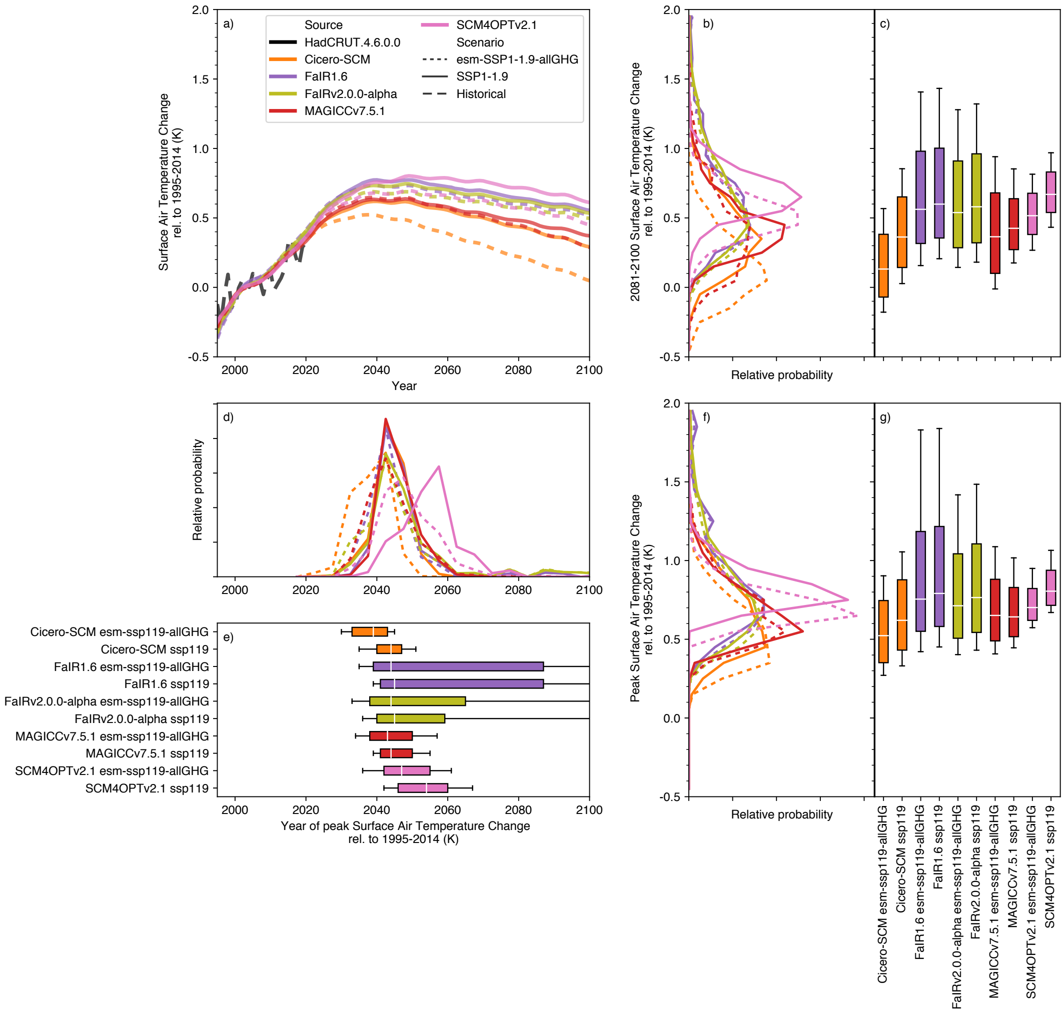

Figure 7. Surface air temperature (also referred to as global-mean surface air temperature, GSAT) change in the concentration-driven SSP1-1.9 experiment and the all greenhouse gas emissions driven esm-SSP1-1.9-allGHG experiment. a) GSAT projections from 1995 to 2100. We show the median RCM projections (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.

reference period, MAGICC7, MCE-v1-2 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 (Figure 8b)), 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, we find that higher historical warming, ECS and TCR values generally lead to higher warming projections (an intuitive result). Hector provides an exception to this pattern, with relatively low temperature projections, especially in SSP1-1.9, despite its relatively high historical warming and TCR.

In the strong mitigation scenarios (SSP1-1.9 and SSP1-2.6), there is agreement to within ~ 0.1°C in future projections (both best-estimate and range) between the models which best reflect historical warming (MAGICC7, MCE-v1-2 and OSCARv3.1). This agreement suggests that constraining greatly increases confidence in future projections. However, a limited set of models also provided probabilistic distributions that are constrained to match HadCRUT.5.0.1.0 (Morice et al., 2021), which is significantly warmer than the HadCRUT.4.6.0.0 based constrained used in the rest of the study. The future projections from these HadCRUT.5.0.1.0-constrained distributions are noticeably warmer (Supplementary Figures S34 - S36) than projections from HadCRUT.4.6.0.0-constrained distributions, which demonstrates that projections are sensitive to the choice of constraint.

Given the sensitivity of conclusions to the constraint, the use of constraints must be carefully considered as it could lead to overconfidence (Sanderson et al., 2017). Even though considerable care is taken both here and elsewhere to identify and use relevant, physically justifiable, constraints, it is still possible that future research may show that the constraints are leading to overconfident future projections. Having said this, Herger et al. (2019) suggest that using multiple constraints, as is done by many RCMs here, reduces the likelihood of overconfidence.

Studies which constrain the raw CMIP6 model ensemble help explain the difference between the RCM-based results presented here and the raw CMIP6 model ensemble. Brunner et al. (2020), Liang et al. (2020) and Tokarska et al. (2020) all find significant reductions in both the best-estimate and 5-95% range GSAT projections after applying observed-warming constraints to the CMIP6 model ensemble. For the SSP1-2.6 and SS5-8.5 scenarios respectively, these studies find 5-95% GSAT (relative to 1995-2014) ranges of: Tokarska et al. (2020): 0.41-1.46°C and 2.26-4.60°C; Liang et al. (2020) 0.52-1.66°C and 2.72-4.77°C and Brunner et al. (2020) 0.61-1.85°C and 2.72-4.86°C. These estimates, particularly for the SSP1-2.6 scenario, are slightly wider than our results based on RCMs. However, the constrained CMIP6 estimates are much closer to our RCM-based estimates than the raw CMIP6 model ensemble, in particular for the 95$^{\text{th}}$ percentile. This suggests that the majority of the difference between our RCM-based results and the raw CMIP6 model ensemble is explained by the constraining applied to the RCMs, rather than structural differences between RCMs and CMIP6 models (although structural differences may explain the disagreement between constrained CMIP6 output and our results). Further studies are needed to explore the validity of the constraining approaches for both ESMs and RCMs - as investigated here - but this study lays the foundation for systematically investigating probabilistic RCM ensembles in more detail.

Given the proxy assessment and results, we make one final observation: to extrapolate assessed warming ranges from one set of scenarios (e.g. the RCP or SSP-based sce-

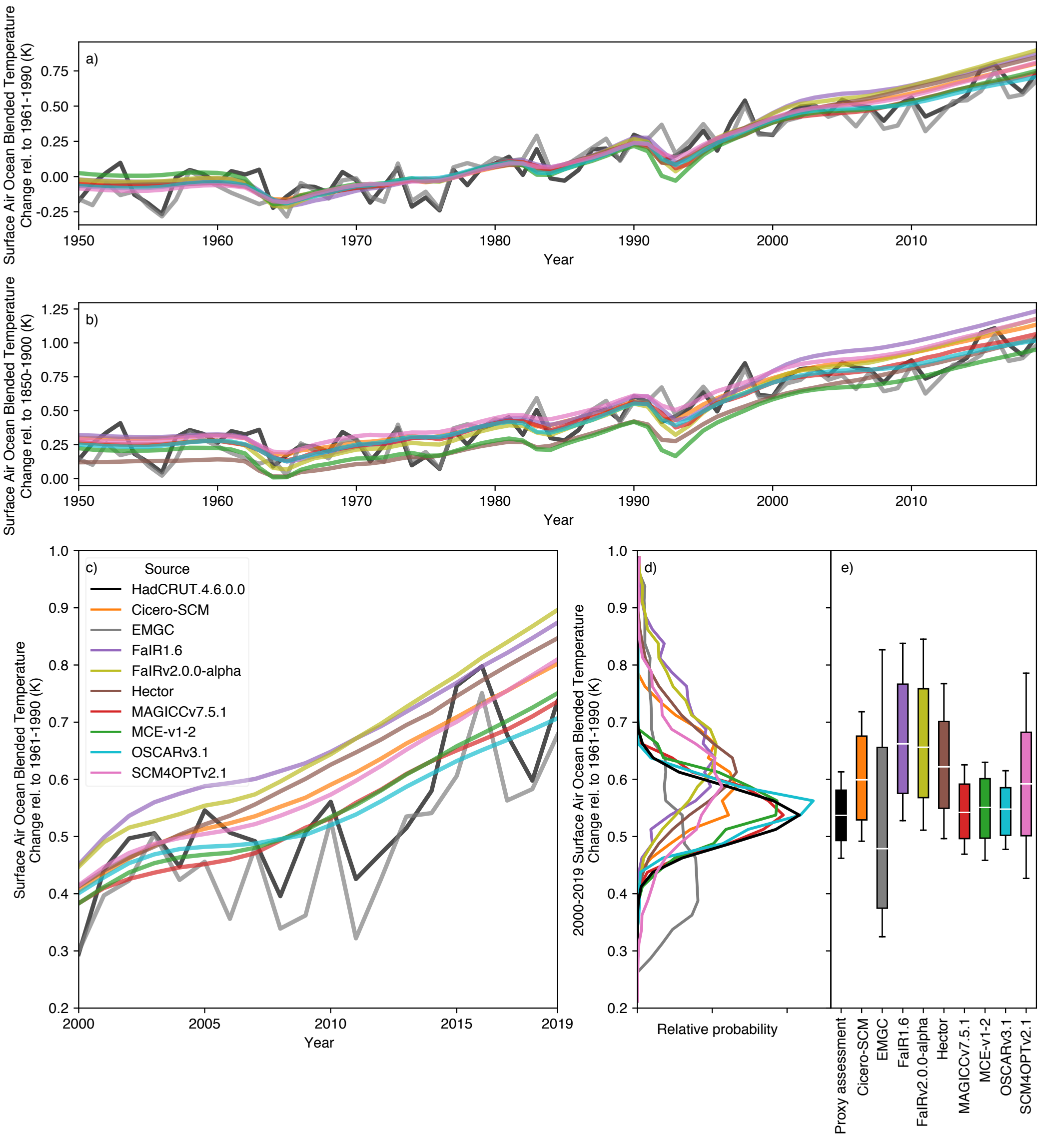

Figure 8. Historical surface air ocean blended temperature change (also referred to as global-mean surface temperature, GMST) from each RCM. We compare observations from HadCRUT4.6.0.0 (Morice et al., 2012) (solid black line) to the distribution from each RCM (coloured lines). All panels use 1961-1990 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.

narios) to a wider set of scenarios, it may be beneficial to include a benchmark of assessed future warming under the benchmark scenarios. This benchmark could be taken from other studies, e.g. those that constrain CMIP projections (for the limited number of scenarios run by CMIP) based on historical observations (e.g. Brunner et al., 2020; Liang et al., 2020; Tokarska et al., 2020). 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 other lines of evidence. Including scenarios with similar end of century total ERF but different transient evolutions (like the SSP4-3.4 and SSP5-3.4-overshoot scenario pair) would provide an even stronger check of the models' transient response. Such quantifications could be key when assessing future projections under large sets of scenarios, like the WG3 scenario database climate assessment. Of course, the risk of adding such benchmarks is an artificial narrowing of uncertainty in projected warming. Hence, future projections should only be included where there is a clear need and justification for consistency between the RCMs' projections and the projections from other lines of evidence.

5 Future work

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. One obvious area for examination would be the aerosol ERF, particularly the inclusion of a climate feedback in aerosol ERF parameterisations. Such an exercise could also provide greater insights into differences between the constrained RCMs' probabilistic distributions, the raw CMIP6 multi-model ensemble and constrained CMIP6 output (building on the discussion in Section 4.3).

A clear limitation of this study is the relative lack of examination of carbon cycle behaviour and carbon cycle related metrics. Given the importance of the carbon cycle for emissions-driven projections, this is another clear area for future work. In the limited examination we have performed, we chose to focus on emissions-driven simulations. This choice provides the cleanest comparison between RCMs and CMIP6 models, given that many RCMs do not separate the land and ocean carbon pools, although it 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). An increase in the number of emissions-driven CMIP6 ESM model output, particularly for mitigation scenarios, would greatly aid such evaluations. Using the concentration-driven simulations in future work will also provide a greater set of comparison data and will facilitate evaluation of RCMs' land and ocean carbon cycles under more varied scenarios.

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. Such an evaluation would provide insights into the drivers of differences between future projections based on the concentration-driven experiments typical of CMIP and results based on the all greenhouse gas emissions-driven experiments required by WG3. 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 matched the proxy assessment well for at least one metric.

Our evaluation is the first multi-model comparison of probabilistic projections from RCMs. This exercise provides a unique insight into RCMs probabilistic parameter ensembles, specifically how they compare with a set of proxy assessed ranges, which reflect wider scientific understanding of key climate metrics, and the implications of differences in probabilistic distributions for climate projections across a range of climate variables and scenarios.

Notably, although unsurprisingly, we found that models whose probabilistic distribution were constrained to the proxy assessed ranges were better able to reflect the proxy assessed ranges. This point is notable because it makes clear that if RCMs are to be used as integrators of knowledge, conveying multiple lines of evidence from one domain to another (e.g. IPCC WG1 to IPCC WG3), then RCMs whose probabilistic distributions have been constrained to the intended lines of evidence are likely to be the best tool.

Even amongst models which had similar levels of agreement with the proxy assessment, some divergence in future projections was observed. Given the various model structures 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, although such extra constraints should be carefully considered given that they risk artificially narrowing projection uncertainty.

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. It is possible to make judgements about what is more reasonable based on the evaluation performed here, and to rule out clearly incorrect projections, yet it must be recognised that 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. In addition, 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 warming seen in CMIP6 models. However, studies which constrain CMIP6 models based on observational constraints also exclude such high warming which suggests that the lack of high warming is due to the constraining applied to the RCMs, rather than structural differences between RCMs and CMIP6 models. 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. Although, further investigations into carbon cycle behaviour are required to provide a clearer picture of the influence of carbon cycle uncertainties on emissions-driven projections. 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

Brunner, L., Pendergrass, A. G., Lehner, F., Merrifield, A. L., Lorenz, R., & Knutti, R. (2020). Reduced global warming from cmip6 projections when weighting models by performance and independence. *Earth System Dynamics Discussions*, *2020*, 1–23. Retrieved from `https://esd.copernicus.org/preprints/esd-2020-23/` doi: 10.5194/esd-2020-23

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

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

Herger, N., Abramowitz, G., Sherwood, S., Knutti, R., Angélil, O., & Sisson, S. A. (2019). Ensemble optimisation, multiple constraints and overconfidence: a case study with future Australian precipitation change. *Climate Dynamics*, *53*(3), 1581–1596. Retrieved from https://doi.org/10.1007/s00382-019-04690-8 doi: 10.1007/s00382-019-04690-8

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

Kawamiya, M., Hajima, T., Tachiiri, K., Watanabe, S., & Yokohata, T. (2020). Two decades of Earth system modeling with an emphasis on Model for Interdisciplinary Research on Climate (MIROC). *Progress in Earth and Planetary Science*, *7*(1), 64. Retrieved from https://doi.org/10.1186/s40645-020-00369-5 doi: 10.1186/s40645-020-00369-5

Leach, N. J., Jenkins, S., Nicholls, Z., Smith, C. J., Lynch, J., Cain, M., … Allen, M. R. (2020). Fairv2.0.0: a generalised impulse-response model for climate uncertainty and future scenario exploration. *Geoscientific Model Development Discussions*, *2020*, 1–48. Retrieved from https://gmd.copernicus.org/preprints/gmd-2020-390/ doi: 10.5194/gmd-2020-390

Liang, Y., Gillett, N. P., & Monahan, A. H. (2020). Climate model projections of 21st century global warming constrained using the observed warming trend. *Geophysical Research Letters*, *47*(12), e2019GL086757. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019GL086757 (e2019GL086757 2019GL086757) doi: https://doi.org/10.1029/2019GL086757

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

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., ... Simpson, I. R. (2021). An updated assessment of near-surface temperature change from 1850: The hadcrut5 data set. *Journal of Geophysical Research: Atmospheres*, *126*(3), e2019JD032361. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD032361` (e2019JD032361 2019JD032361) doi: https://doi.org/10.1029/2019JD032361

Murphy, J. M., Booth, B. B. B., Boulton, C. A., Clark, R. T., Harris, G. R., Lowe, J. A., & Sexton, D. M. H. (2014). Transient climate changes in a perturbed parameter ensemble of emissions-driven earth system model simulations. *Climate Dynamics*, *43*(9), 2855–2885. Retrieved from `https://doi.org/10.1007/s00382-014-2097-5` doi: 10.1007/s00382-014-2097-5

Nicholls, Z., & Lewis, J. (2021, March). *Reduced Complexity Model Intercomparison Project (RCMIP) protocol.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.4589756` doi: 10.5281/zenodo.4589756

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/

Sanderson, B. M., Wehner, M., & Knutti, R. (2017). Skill and independence weighting for multi-model assessments. *Geoscientific Model Development*, *10*(6), 2379–2395. Retrieved from https://gmd.copernicus.org/articles/10/2379/2017/ doi: 10.5194/gmd-10-2379-2017

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

Skeie, R. B., Peters, G. P., Fuglestvedt, J., & Andrew, R. (2021). A future perspective of historical contributions to climate change. *Climatic Change*, *164*(1), 24. Retrieved from https://doi.org/10.1007/s10584-021-02982-9 doi: 10.1007/s10584-021-02982-9

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

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

Supporting Information for "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. L. Woodard[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

March 17, 2021, 8:57pm

[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

Contents of this file

10. Figure S16: SSP5-8.5 ERF projections until 2100

11. Figure S17: SSP1-1.9 ERF projections until 2300

12. Figure S18: SSP1-2.6 ERF projections until 2300

13. Figure S19: SSP1-1.9 aerosol ERF projections until 2300

14. Figure S20: SSP1-1.9 CO_2 ERF projections until 2100

15. Figure S21: SSP1-1.9 CO_2 ERF projections until 2300

16. Figure S22: SSP1-2.6 aerosol ERF projections until 2100

17. Figure S23: SSP1-2.6 aerosol ERF projections until 2300

18. Figure S24: SSP1-2.6 CO_2 ERF projections until 2100

19. Figure S25: SSP1-2.6 CO_2 ERF projections until 2300

20. Figure S26: SSP5-8.5 aerosol ERF projections until 2100

21. Figure S27: SSP5-8.5 aerosol ERF projections until 2300

22. Figure S28: SSP5-8.5 CO_2 ERF projections until 2100

23. Figure S29: SSP5-8.5 CO_2 ERF projections until 2300

24. Figure S30: ESM-SSP1-1.9 atmospheric CO_2 concentration projections until 2100

25. Figure S31: ESM-SSP1-2.6 atmospheric CO_2 concentration projections until 2100

26. Figure S32: SSP1-2.6 and ESM-SSP1-2.6-allGHG GSAT projections until 2100

27. Figure S33: SSP5-8.5 and ESM-SSP5-8.5-allGHG GSAT projections until 2100

28. Figure S34: SSP1-1.9 HadCRUT.4.6.0.0 and HadCRUT.5.0.1.0 constrained GST projections until 2300

29. Figure S35: SSP1-2.6 HadCRUT.4.6.0.0 and HadCRUT.5.0.1.0 constrained GST projections until 2300

30. Figure S36: SSP5-8.5 HadCRUT.4.6.0.0 and HadCRUT.5.0.1.0 constrained GST projections until 2300

Text S1.

CICERO-SCM

The CICERO Simple climate model (CICERO-SCM) consists of a carbon cycle model (Joos et al., 1996), simplified expressions relating emissions of components to forcing, either directly or via concentrations (Etminan et al., 2016; Skeie et al., 2017) and an energy balance/upwelling diffusion model (Schlesinger et al., 1992). A detailed description of the CICERO-SCM is presented in Skeie et al. (2017) with recent updates in Nicholls et al. (2020). The energy balance/upwelling diffusion model calculates warming separately for the two hemispheres and includes 40 vertical layers in the ocean. The parameters that govern the mixing of heat in the ocean as well at the climate sensitivity and radiative forcing is estimated in a Bayesian approach using observational based time series of global mean surface temperature change and ocean heat content and prior estimates of radiative forcing times series (Skeie et al., 2018). The posteriori distribution of the parameters in the energy balance/upwelling diffusion model including the climate sensitivity and aerosol forcing are used in the probabilistic run. Uncertainties in other climate drivers are ignored. The climate sensitivity estimated in Skeie et al. (2018) is the inferred effective climate sensitivity (ECS_{inf}), known to be lower than the equilibrium climate sensitivity (ECS). This is due to the use of global-mean surface temperature instead of global-mean surface air temperature as well as feedback on long time scales that have not come into play when inferring the climate sensitivity from the historical record. The assessed range for the equilibrium climate sensitivity is therefore shifted by 0.9°C (Skeie et al., 2018) to lower values when selecting the parameter set to be used in the probabilistic simulations. Another selection criterion is the assessed range of surface air temperature change in 1985-2014 relative to the baseline 1850-1900. The 30 040 original parameter set from Skeie et al. (2018) is subsetted using the RCMIP defined ECS distribution as the primary constraint. By binning the data, a subset following this same distribution is built. Simultaneously, the surface air temperature constraint is used as a hard cutoff to choose between parameter sets in this distribution, producing a final subset of between 550 and 600 ensemble members.

EMGC

The University of Maryland Empirical Model of Global Climate (EMGC) is a multiple linear regression energy balance representation of various factors (both natural and anthropogenic) that control global-mean surface temperature (GMST) (Canty et al., 2013; Hope et al., 2017). Values of climate feedback and ocean heat uptake efficiency are found in a regression framework constrained by observed time series of GMST, the radiative forcing due to tropospheric aerosols (AER RF), and ocean heat content. Several natural and anthropogenic components that affect GMST are also considered. Recently, we have added an interactive ocean module that represents the warming of the ocean profile in response to rising GMST (Hope et al., 2020; McBride et al., 2020). As a result, transport of heat from the atmosphere to the world's oceans evolves over time in a more realistic fashion compared to earlier versions of our model.

The EMGC forecasts of GMST all use values of radiative forcing due to future greenhouse gas abundances and aerosols prescribed by RCMIP. Values of climate feedback and ocean heat uptake for each simulation are found based upon regression analysis of the data record for GMST from HadCRUT4 (Morice et al., 2012) that spans 1850 until the end of 1999 (an alternate parameter set is also submitted based on using the HadCRUT.5.0.1.0 dataset instead) and an ocean heat content record that is the average of data from five groups that spans 1955 to 1999 (Hope et al., 2020; McBride et al., 2020). The time series of radiative forcing due to tropospheric aerosols is scaled such that the values in 2011 (AER RF_{2011}) statistically sample the uncertainty in the value of AER RF_{2011} given by Chapter 8 of AR5 (Myhre et al., 2013). Projections of GMST are based upon analyses of the subset of the 160,000 possible combinations of climate feedback and AER RF_{2011} that provide a good fit (reduced-chi-squared metric less than or equal to 2) to the observed variation in GMST and OHC, as described in Hope et al. (2020) and McBride et al. (2020). The largest factor driving spread in our future projections of GMST from the ensemble of model runs that satisfy the reduced-chi-squared metric is imprecise knowledge of the radiative forcing of climate by tropospheric aerosols over the historical time period. Ensemble members with largest future warming are characterized by values of AER RF_{2011} towards the low end of the distribution (i.e., -1.5 W m^{-2} to -1.9 W m^{-2}) and the ensemble members with smallest future warming are characterized by values of AER RF_{2011} towards the high end of the distribution (i.e., -0.1 W m^{-2} to -0.4 W m^{-2}).

FaIRv1.6.1

The Finite-amplitude Impulse Response (FaIR) model is an emissions-driven simple climate model written in Python. Since the v1.3 model description paper (Smith et al., 2018), a number of features have been added. FaIR v1.6 separately reports greenhouse gas forcings from 28 different fluorinated species (which were all aggregated in FaIR v1.3), and breaks aerosol forcing down into five direct species and indirect aerosol forcing.

Functionality has been added for running experiments in concentration-driven mode and for deriving CO_2 emissions from prescribed concentrations, enabling a greater set of the RCMIP experiments to be run. Additionally, it is now possible to run FaIR using the parameters of the two-layer model defined by Geoffroy, Saint-Martin, Bellon, et al. (2013), given that this model is simply a mathematical transformation of FaIR's impulse response setup (as shown by Geoffroy, Saint-Martin, Olivié, et al. (2013)). Finally, the carbon cycle has been optimised following FaIR 2.0.0, speeding up runtime (Leach et al., 2020).

An initial ensemble of 3000 members were drawn for RCMIP. Forcing uncertainties for CH4, N2O, other GHGs, tropospheric ozone, stratospheric ozone, contrails, black carbon on snow, land use change, solar and volcanic are taken from AR5 uncertainty ranges (Myhre et al., 2013). Two-layer model parameters for ocean heat exchange coefficient, climate feedback parameter, efficacy of deep ocean heat uptake and heat capacity of the mixed layer and deep ocean are sampled with distributions informed by 44 CMIP6 models built from joint kernel density distributions that take correlations of terms into account (Smith et al., submitted). The ERF from 4xCO2 is also taken from the ensemble based on abrupt-4xCO2 experiments from CMIP6 models and used to inform the uncertainty range for CO2 forcing. For ESM runs, the carbon cycle parameters (pre-industrial airborne fraction, and sensitivity to temperature and atmospheric CO2 burden) are sampled from normal distributions as in Smith et al. (2018). Direct aerosol forcing from SO2, BC and OC is sampled from CMIP6 models participating in RFMIP and AerChemMIP (Smith et al., submitted). Nitrate and secondary organic aerosol are not included. Indirect aerosol forcing is sampled by scaling the 1850-2010 aerosol forcing to a Gaussian distribution centred on -0.85 W / m^2 with standard deviation of 0.91 W / m^2. The 3000-member prior ensemble is reduced to a final ensemble of 501 members, where this ensemble was selected from the members with the smallest RMSE for their GMST from historical (1850-2014) integrations compared to the Cowtan and Way (2014) dataset (v2.0.0) for 1850-2014. FaIR does not report GMST, but the simple assumption that GSAT anomalies are 4% greater than GMST is used based on CMIP6 models and reanalysis datasets.

FaIRv2.0.0-alpha

FaIRv2.0.0-alpha (Leach et al., 2020) is an update to the FaIR model (version 1.6 is described above). This update reduces the model's structural complexity as comprehensively as possible. The result is a set of six equations - the five equations that made up the impulse-response model used for GHG metric calculations in the IPCC 5^{th} Assessment Report (Myhre et al., 2013), plus one additional equation that introduces a state-dependence to the carbon and methane cycles (Millar et al., 2017).

A 1 million-member ensemble is generated by perturbing parameters relating to the modelled carbon-cycle, ERF and thermal response. Prior carbon-cycle and thermal response distributions are inferred from parameter samples obtained by tuning the model to idealised experiments in the CMIP6 ensemble. The carbon-cycle was tuned to 11 models from C4MIP (Arora et al., 2020); the thermal response cycle was tuned to 28 models using a maximum likelihood method (Cummins et al., 2020). Prior ERF parameter uncertainties were taken from AR5 uncertainty ranges (Myhre et al., 2013) for all forcing classes except for aerosol-radiation and -cloud interaction. These were sampled from distributions informed by tuning the aerosol ERF parameterisations to 10 CMIP6 models (Smith et al., submitted) and then quantile mapped to match the process-based assessment in Bellouin et al. (2020).

This large prior ensemble is constrained by setting the selection probability of an individual member equal to the likelihood of its corresponding present-day level and rate of anthropogenic warming calculated using the Global Warming Index methodology (Haustein et al., 2017) (with the HadCRUT.4.6.0.0 timeseries for the main results but also the HadCRUT.5.0.1.0 timeseries for illustration). A 5 000-member subset of this constrained ensemble (total size 250 651) is used in RCMIP phase 2.

Hector v2.5.0

Hector is an open-source globally resolved, process-based carbon-climate model that calculates the annual energy fluxes between the ocean, atmosphere, and terrestrial biosphere (Hartin et al., 2015). As of Hector v2.0 (Vega-Westhoff et al., 2019), the model uses an implementation of the 1-D ocean heat diffusion model, DOECLIM (Kriegler, 2005; Urban et al., 2014). Recent model updates to v2.5.0 include: reorganizing the code as an R package, constraining pre-industrial atmospheric CO_2 to a prescribed value during model spin-up, and updating the OH lifetime.

For each scenario, Hector was run 10 000 times with parameters (equilibrium climate sensitivity, ocean heat diffusivity, and aerosol forcing) randomly sampled from the joint posterior distribution from the Vega-Westhoff et al. (2019) MCMC calibration against historical global surface temperature observations and ocean heat content. Using the parametrization from the posterior distribution we produced probabilistic Hector output for global mean air temperature, air-ocean blended temperature, and aerosol radiative forcing.

MAGICC7

MAGICC's climate core is based on a 50-layer, hemispherically resolved upwelling-diffusion-entrainment ocean model coupled to a four-box (hemispheric land/ocean) spatial resolution for effective radiative forcing. MAGICC and runs on monthly timesteps, which improves its representation of the response to volcanic eruptions compared to an annual timestep. The version of MAGICC used here (v7.4.1) is an update of MAGICC6 (Meinshausen et al., 2011) and the setup used to generate the GHG concentration projections (Meinshausen et al., 2020) for the historical and SSP-based CMIP6 experiments (Eyring et al., 2016; O'Neill et al., 2016). The key updates are the inclusion of a state-dependent climate feedback factor (previously it was only forcing-dependent) which has been calibrated to CMIP6 models (Nicholls et al., 2020), accounting for the effect of large historical anthropogenic biomass burning aerosol precursor emissions on aerosol effective radiative forcing, a nitrate aerosol forcing scheme which accounts for the sulfate competition for ammonia based on Hauglustaine, Balkanski, and Schulz (2014) and the inclusion of a non-ocean heat uptake parameterisation which represents land surface and cryosphere heat uptake in each hemisphere. In addition, it includes an updated effective radiative forcing parameterisations for CO_2, CH_4 and N_2O that capture results by Etminan et al. (2016), while allowing for a wider range of input concentrations (see Meinshausen et al. (2020)] for details).

We derive a posterior parameter distribution using the methodology of Meinshausen et al. (2009), updated to use observations of global-mean temperature up to 2019 based on HadCRUT4.6.0.0(Morice et al., 2012) (an alternate set which uses HadCRUT.5.0.0.0 is also included for sensitivity analysis) and ocean heat content up to 2018 based on von Schuckmann et al. (2020) as well as the proxy effective radiative forcing assessment used in this study. We run a Monte Carlo Markov Chain with 20 million steps, from which we draw every 200$^{\mathrm{th}}$ member, resulting in a 100 000 member posterior distribution. The probabilistic distribution used here is the result of sub-sampling the posterior distribution to draw a set of 600 parameter sets which best match the proxy assessed ranges and also maintain the covariance of MAGICC's parameters as derived from the posterior distribution.

MCE

MCE consists of a thermal response module and a carbon cycle module. These are represented by impulse response functions (Hooss et al., 2001; Joos et al., 1996), responding to anthropogenic carbon input which then alters the ERF of the atmospheric CO_2 and natural processes in the ocean and terrestrial carbon cycle. The carbon cycle incorporates temperature feedbacks via dissociation constants in the chemical equilibrium of the carbonic acid system in seawater and the respiration of organic materials in the terrestrial biosphere. After being used in RCMIP Phase 1, the CO_2 forcing scheme was slightly changed, and schemes for non-CO_2 well-mixed GHGs were newly incorporated instead of using prescribed scenario data. The CO_2 scheme has two control parameters: one for scaling in terms of the logarithm of CO_2 concentrations, and the other for amplifying deviations from the logarithmic increase (Tsutsui, 2017). The latter is activated when the concentration exceeds a two-times level with a quadratic term, but was modified here to be linear when the concentration further exceeds a four-times level. The non-CO_2 schemes use those by Etminan et al. (2016) for CH_4 and N_2O, and a simple linear formula for halocarbon gases with their lifetimes and radiative efficiencies assessed in AR5.

The probabilistic runs were conducted with 600-member parameter sets, varied for (1) CO_2 forcing and thermal response, (2) non-CO_2 forcing scaling, and (3) ocean and land CO_2 uptake. The first sets were generated from a multivariate normal distribution built on principal components of individual parameters adjusted to CMIP5 and CMIP6 models (Tsutsui, 2017, 2020). Cross-correlation between the parameters of this group reflects the variation of the CMIP models, such that the ratio of TCR-to-ECS tends to decrease with increase in ECS, and that CO_2 forcing is weakly correlated with response properties. The second sets were implemented as scaling factors of non-CO_2 forcing, and individually generated from a probability distribution modeled for each of the prescribed likely ranges, the third sets were implemented as perturbations on the amplitudes of the impulse response function for the ocean CO_2 uptake, and on two land-CO_2 parameters for the fertilization effect and the temperature dependency of respiration. These perturbations were individually generated from a uniform distribution so that resulting carbon budgets encompass the range of those from CMIP5 and CMIP6 Earth system models presented in Arora et al. (2020). A Bayesian updating was applied to constrain the parameter sets with a Metropolis-Hastings sampling algorithm sequentially as to land CO_2 uptake, the ERF of CO_2, TCR, and the two metrics for the surface blended temperature and the ocean heat content. The land CO_2 constraint was targeted for the excess carbon at doubling along a CO_2 concentration pathway under an idealized 1%-per-year increase scenario from the CMIP Earth system models while the other constraints follow the assessed ranges. It is supposed that the second posterior conforms to the CMIP ensemble and the assessed forcing ranges, and that the last (fourth) posterior, from which the 600 members were sampled, is a compromised distribution reflecting all the metrics together with the CMIP ensemble. For constraining the temperature and heat content metrics, a bivariate normal distribution was built with factors of 1.04 and 1.08 for conversion from the blended temperature to the air temperature, and from the ocean heat content to the total heat content, respectively.

OSCARv3.1

OSCAR v3.1 is an open-source reduced-form Earth system model, whose modules mimic models of higher complexity in a probabilistic setup (Gasser, Ciais, et al., 2017). The response of the global surface temperature to radiative forcing is the two-layer model (Geoffroy, Saint-Martin, Bellon, et al., 2013). OSCAR calculates the effective radiative forcing caused by greenhouse gases (CO_2, CH_4, N_2O, 37 halogenated compounds), short-lived climate forcers (tropospheric and stratospheric ozone, stratospheric water vapor, nitrates, sulfates, black carbon, primary and secondary organic aerosols) and changes in surface albedo. The ocean carbon cycle is based on the mixed-layer response function of Joos et al. (1996), albeit with an added stratification of the upper ocean derived from CMIP5 (Arora et al., 2013) and with an updated carbonate chemistry. The land carbon cycle is divided into five biomes and five regions, and each of the 25 biome/region combinations follows a three-box model (soil, litter and vegetation). Land cover change, wood harvest and shifting cultivation are also accounted for, thanks to a dedicated book-keeping module that allows OSCAR to estimate its own CO_2 emissions from land-use change (Gasser, Peters, et al., 2017; Gasser et al., 2020). Permafrost thaw and the consequent emission of CO_2 and CH_4 is also modeled (Gasser et al., 2018). In addition, biomass burning emissions are calculated endogenously following the book-keeping module and the wildfire feedback. These emissions were therefore subtracted from the input RCMIP data used to drive OSCAR to avoid double counting.

In RCMIP phase 2, the same 10 000 elements of the Monte-Carlo ensemble used in RCMIP phase 1 are used. Each simulation is run using all these configurations. The parameters of OSCARv3.1 are not tuned to reflect the assessed ranges required, but instead, each configuration is weighted. The weights are determined by comparing the performances over the emissions-driven historical experiment to the assessed ranges for the cumulative net land to atmosphere and ocean to atmosphere fluxes to constrain long-term dynamics, and the rate of increase in atmospheric CO_2 for short-term dynamics. More details about this weighting approach can be found in Gasser et al. (2020). We choose not to use the historical surface air-ocean blended temperature as an additional constraint, as it causes the final range of the equilibrium climate sensitivity of OSCARv3.1 to be drastically reduced. All final outputs are provided as the resulting quantiles.

SCM4OPT v2.1

The Simple Climate Model for Optimization version 2.1 (SCM4OPT v2.1) (Su et al., 2020) is a simple climate model which can simulate the radiative forcing and global temperature change resulting from a full suite of greenhouse gases, pollutants and aerosols, as well as land-use albedo. The SCM4OPT v2.1 is designed to be lightweight and capable of being used in an integrated assessment model (IAM) with a large-scale optimization process. Compared to the older version (Su et al., 2017, 2018), we updated the ocean carbon cycle following Hector v1.0 (Hartin et al., 2015) and used the Diffusion Ocean Energy balance CLIMate (DOECLIM) model (Kriegler, 2005; Tanaka et al., 2007) to calculate global-mean temperature change. We fitted the CO2 concentration and temperature change of the SCM4OPT v2.1 to the associated outputs of four RCP experiments (RCP2.6, RCP4.5, RCP6.0 and RCP8.5) of 26 coupled atmosphere ocean general circulation models (AOGCMs) in CMIP5. In addition, the method used to estimate aerosol forcing was also renewed based on OSCAR v2.2 (Gasser, Ciais, et al., 2017). However, we removed a few parameter sets which could generate unrealistic outliers, and re-tuned the forcing efficiencies and other related parameters against the aerosol forcings presented in IPCC AR5 (IPCC, 2013). An ensemble of 2000 members was adopted for RCMIP to represent the uncertainties caused by the carbon cycle, aerosol forcings and temperature change by using randomized parameter sets as described above.

References

Arora, V. K., Boer, G. J., Friedlingstein, P., Eby, M., Jones, C. D., Christian, J. R., … Wu, T. (2013, 07). Carbon–Concentration and Carbon–Climate Feedbacks in CMIP5 Earth System Models. *Journal of Climate*, *26*(15), 5289-5314. Retrieved from https://doi.org/10.1175/JCLI-D-12-00494.1 doi: 10.1175/JCLI-D-12-00494.1

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

Bellouin, N., Davies, W., Shine, K. P., Quaas, J., Mülmenstädt, J., Forster, P. M., … Myhre, G. (2020). Radiative forcing of climate change from the copernicus reanalysis of atmospheric composition. *Earth System Science Data*, *12*(3), 1649–1677. Retrieved from https://essd.copernicus.org/articles/12/1649/2020/

doi: 10.5194/essd-12-1649-2020

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

Cowtan, K., & Way, R. G. (2014). Coverage bias in the hadcrut4 temperature series and its impact on recent temperature trends. *Quarterly Journal of the Royal Meteorological Society*, *140*(683), 1935–1944. doi: 10.1002/qj.2297

Cummins, D. P., Stephenson, D. B., & Stott, P. A. (2020). Optimal estimation of stochastic energy balance model parameters. *Journal of Climate*, *33*(18), 7909–7926.

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).

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

Gasser, T., Peters, G. P., Fuglestvedt, J. S., Collins, W. J., Shindell, D. T., & Ciais, P. (2017). Accounting for the climate–carbon feedback in emission metrics. *Earth System Dynamics*, *8*(2), 235–253. Retrieved from https://esd.copernicus.org/articles/8/235/2017/ doi: 10.5194/esd-8-235-2017

Geoffroy, O., Saint-Martin, D., Bellon, G., Voldoire, A., Olivié, D. J. L., & Tytéca, S. (2013, 03). Transient Climate Response in a Two-Layer Energy-Balance Model. Part II: Representation of the Efficacy of Deep-Ocean Heat Uptake and Validation for CMIP5 AOGCMs. *Journal of Climate*, *26*(6), 1859-1876. Retrieved from https://doi.org/10.1175/JCLI-D-12-00196.1 doi: 10.1175/JCLI-D-12-00196.1

Geoffroy, O., Saint-Martin, D., Olivié, D. J. L., Voldoire, A., Bellon, G., & Tytéca, S. (2013, mar). Transient climate response in a two-layer energy-balance model. part i: Analytical solution and parameter calibration using CMIP5 AOGCM experiments. *Journal of Climate*, *26*(6), 1841–1857. doi: 10.1175/jcli-d-12-00195.1

Hartin, C. A., Patel, P., Schwarber, A., Link, R. P., & Bond-Lamberty, B. P. (2015, apr). A simple object-oriented and open-source model for scientific and policy analyses of the global climate system – hector v1.0. *Geoscientific Model Development*, *8*(4), 939–955. doi: 10.5194/gmd-8-939-2015

Hauglustaine, D. A., Balkanski, Y., & Schulz, M. (2014). A global model simulation of present and future nitrate aerosols and their direct radiative forcing of climate. *Atmospheric Chemistry and Physics*, *14*(20), 11031–11063. Retrieved from https://acp.copernicus.org/articles/14/11031/2014/ doi: 10.5194/acp-14-11031-2014

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

IPCC. (2013). Summary for policymakers [Book Section]. In T. Stocker et al. (Eds.), *Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate* *change* (p. 1–30). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. Retrieved from http://www.climatechange2013.org doi: 10.1017/CBO9781107415324.004

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

Kriegler, E. (2005). *Imprecise probability analysis for integrated assessment of climate change* (Doctoral thesis, Universität Potsdam). Retrieved from https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/index/index/docId/497

Leach, N. J., Jenkins, S., Nicholls, Z., Smith, C. J., Lynch, J., Cain, M., ... Allen, M. R. (2020). Fairv2.0.0: a generalised impulse-response model for climate uncertainty and future scenario exploration. *Geoscientific Model Development Discussions*, *2020*, 1–48. Retrieved from https://gmd.copernicus.org/preprints/gmd-2020-390/ doi: 10.5194/gmd-2020-390

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

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

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., ... Simpson, I. R. (2021). An updated assessment of near-surface temperature change from 1850: The hadcrut5 data set. *Journal of Geophysical Research: Atmospheres*, *126*(3), e2019JD032361. Retrieved from https://agupubs.onlinelibrary

.wiley.com/doi/abs/10.1029/2019JD032361 (e2019JD032361 2019JD032361) doi: https://doi.org/10.1029/2019JD032361

Myhre, G., Shindell, D., Bréon, F.-M., Collins, W., Fuglestvedt, J., Huang, J., ... Zhang, H. (2013). Anthropogenic and natural radiative forcing [Book Section]. In T. Stocker et al. (Eds.), *Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change* (p. 659–740). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. Retrieved from http://www.climatechange2013.org doi: 10.1017/CBO9781107415324.018

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

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).

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., et al. (submitted).

JGR: Atmospheres.

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.

Tanaka, K., Kriegler, E., Bruckner, T., Hooss, G., Knorr, W., Raddatz, T., & Tol, R. (2007). *Aggregated carbon cycle, atmospheric chemistry and climate model (acc2): description of forward and inverse mode.* Retrieved from https://pure.mpg.de/rest/items/item_994422/component/file_994421/content

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

Urban, N. M., Holden, P. B., Edwards, N. R., Sriver, R. L., & Keller, K. (2014). Historical and future learning about climate sensitivity. *Geophysical Research Letters*, *41*(7), 2543-2552. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014GL059484 doi: 10.1002/2014GL059484

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

Table S1. Comparison of each model's probabilistic distribution with the proxy assessment. 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).

	Source	Assessed ranges					CICERO-SCM				
	Assessed range	vll	ll	c	lu	vlu	vll	ll	c	lu	vlu
Metric	Unit										
2000-2019 GMST rel. to 1961-1990	K	0.46		0.54		0.61	0.49		0.60		0.72
Equilibrium Climate Sensitivity	K	2.30	2.60	3.10	3.90	4.70	2.52	2.69	3.03	3.43	3.67
Transient Climate Response	K	0.98	1.26	1.64	2.02	2.29	1.31	1.44	1.66	1.91	2.03
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			1.94	1.94	1.94	
2014 Aerosol Effective Radiative Forcing	W / m^2		-1.37	-1.01	-0.63			-0.99	-0.64	-0.23	
2018 Ocean Heat Content rel. to 1971	ZJ		303	320	337			303	320	337	
2011 CH_4 Effective Radiative Forcing	W / m^2		0.47	0.60	0.73			0.53	0.53	0.53	
2011 N_2O Effective Radiative Forcing	W / m^2		0.14	0.17	0.20			0.16	0.16	0.16	
2011 F-Gases Effective Radiative Forcing	W / m^2		0.03	0.03	0.03						

Table S1 (cont.). Comparison of each model's probabilistic distribution with the proxy assessment. 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).

	Source	Assessed ranges					EMGC				
	Assessed range	vll	ll	c	lu	vlu	vll	ll	c	lu	vlu
Metric	Unit										
2000-2019 GMST rel. to 1961-1990	K	0.46		0.54		0.61	0.32		0.48		0.83
Equilibrium Climate Sensitivity	K	2.30	2.60	3.10	3.90	4.70	1.30	1.50	1.93	2.79	4.14
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			1.90	1.90	1.90	
2014 Aerosol Effective Radiative Forcing	W / m^2		-1.37	-1.01	-0.63			-1.16	-0.84	-0.53	
2018 Ocean Heat Content rel. to 1971	ZJ		303	320	337			277	327	426	
2011 CH_4 Effective Radiative Forcing	W / m^2		0.47	0.60	0.73			0.58	0.58	0.58	
2011 N_2O Effective Radiative Forcing	W / m^2		0.14	0.17	0.20			0.18	0.18	0.18	
2011 F-Gases Effective Radiative Forcing	W / m^2		0.03	0.03	0.03						

Table S1 (cont.). Comparison of each model's probabilistic distribution with the proxy assessment. The assessed ranges are labelled as 'vll' (very-likely lower i.e. 5^{th} percentile), 'll' (likely lower, 17^{th} percentile), 'c' (central, 50^{th} percentile), 'lu' (likely upper, 83^{th} percentile) and 'vlu' (very-likely upper, 95^{th} percentile).

	Source	Assessed ranges					FaIR1.6				
	Assessed range	vll	ll	c	lu	vlu	vll	ll	c	lu	vlu
Metric	Unit										
2000-2019 GMST rel. to 1961-1990	K	0.46		0.54		0.61	0.53		0.66		0.84
Equilibrium Climate Sensitivity	K	2.30	2.60	3.10	3.90	4.70	1.93	2.33	3.01	4.33	6.21
Transient Climate Response	K	0.98	1.26	1.64	2.02	2.29	1.40	1.55	1.80	2.12	2.48
Transient Climate Response to Emissions	K / TtC	1.03	1.40	1.77	2.14	2.51	1.21	1.36	1.59	1.91	2.22
2014 CO_2 Effective Radiative Forcing	W / m^2		1.69	1.80	1.91			1.65	1.89	2.17	
2014 Aerosol Effective Radiative Forcing	W / m^2		-1.37	-1.01	-0.63			-1.34	-1.01	-0.65	
2018 Ocean Heat Content rel. to 1971	ZJ		303	320	337			303	359	417	
2011 CH_4 Effective Radiative Forcing	W / m^2		0.47	0.60	0.73			0.49	0.55	0.62	
2011 N_2O Effective Radiative Forcing	W / m^2		0.14	0.17	0.20			0.15	0.17	0.18	
2011 F-Gases Effective Radiative Forcing	W / m^2		0.03	0.03	0.03			0.03	0.03	0.03	

Table S1 (cont.). Comparison of each model's probabilistic distribution with the proxy assessment. The assessed ranges are labelled as 'vll' (very-likely lower i.e. 5^{th} percentile), 'll' (likely lower, 17^{th} percentile), 'c' (central, 50^{th} percentile), 'lu' (likely upper, 83^{th} percentile) and 'vlu' (very-likely upper, 95^{th} percentile).

	Source	Assessed ranges					FaIRv2.0.0-alpha				
	Assessed range	vll	ll	c	lu	vlu	vll	ll	c	lu	vlu
Metric	Unit										
2000-2019 GMST rel. to 1961-1990	K	0.46		0.54		0.61	0.51		0.66		0.85
Equilibrium Climate Sensitivity	K	2.30	2.60	3.10	3.90	4.70	1.80	2.20	3.05	4.50	6.17
Transient Climate Response	K	0.98	1.26	1.64	2.02	2.29	1.23	1.41	1.71	2.08	2.38
Transient Climate Response to Emissions	K / TtC	1.03	1.40	1.77	2.14	2.51	1.02	1.17	1.41	1.74	2.01
2014 CO_2 Effective Radiative Forcing	W / m^2		1.69	1.80	1.91			1.73	1.95	2.18	
2014 Aerosol Effective Radiative Forcing	W / m^2		-1.37	-1.01	-0.63			-1.53	-1.14	-0.80	
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			0.50	0.59	0.69	
2011 N_2O Effective Radiative Forcing	W / m^2		0.14	0.17	0.20			0.15	0.17	0.19	
2011 F-Gases Effective Radiative Forcing	W / m^2		0.03	0.03	0.03			0.03	0.03	0.04	

Table S1 (cont.). Comparison of each model's probabilistic distribution with the proxy assessment. The assessed ranges are labelled as 'vll' (very-likely lower i.e. 5^{th} percentile), 'll' (likely lower, 17^{th} percentile), 'c' (central, 50^{th} percentile), 'lu' (likely upper, 83^{th} percentile) and 'vlu' (very-likely upper, 95^{th} percentile).

	Source	Assessed ranges					Hector				
	Assessed range	vll	ll	c	lu	vlu	vll	ll	c	lu	vlu
Metric	Unit										
2000-2019 GMST rel. to 1961-1990	K	0.46		0.54		0.61	0.50		0.62		0.77
Equilibrium Climate Sensitivity	K	2.30	2.60	3.10	3.90	4.70	1.83	2.17	2.86	3.91	5.44
Transient Climate Response	K	0.98	1.26	1.64	2.02	2.29	1.42	1.58	1.81	2.07	2.30
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			-0.68	-0.57	-0.45	
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						

Table S1 (cont.). Comparison of each model's probabilistic distribution with the proxy assessment. 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).

	Source	Assessed ranges					MAGICCv7.5.1				
	Assessed range	vll	ll	c	lu	vlu	vll	ll	c	lu	vlu
Metric	Unit										
2000-2019 GMST rel. to 1961-1990	K	0.46		0.54		0.61	0.47		0.54		0.63
Equilibrium Climate Sensitivity	K	2.30	2.60	3.10	3.90	4.70	2.00	2.20	2.66	3.27	3.90
Transient Climate Response	K	0.98	1.26	1.64	2.02	2.29	1.30	1.46	1.76	2.07	2.30
Transient Climate Response to Emissions	K / TtC	1.03	1.40	1.77	2.14	2.51	1.13	1.38	1.75	2.17	2.47
2014 CO_2 Effective Radiative Forcing	W / m^2		1.69	1.80	1.91			1.66	1.78	1.89	
2014 Aerosol Effective Radiative Forcing	W / m^2		-1.37	-1.01	-0.63			-1.39	-1.08	-0.75	
2018 Ocean Heat Content rel. to 1971	ZJ		303	320	337			305	323	340	
2011 CH_4 Effective Radiative Forcing	W / m^2		0.47	0.60	0.73			0.47	0.58	0.71	
2011 N_2O Effective Radiative Forcing	W / m^2		0.14	0.17	0.20			0.14	0.17	0.20	
2011 F-Gases Effective Radiative Forcing	W / m^2		0.03	0.03	0.03			0.03	0.03	0.03	

Table S1 (cont.). Comparison of each model's probabilistic distribution with the proxy assessment. 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).

	Source	Assessed ranges					MCE-v1-2				
	Assessed range	vll	ll	c	lu	vlu	vll	ll	c	lu	vlu
Metric	Unit										
2000-2019 GMST rel. to 1961-1990	K	0.46		0.54		0.61	0.46		0.55		0.63
Equilibrium Climate Sensitivity	K	2.30	2.60	3.10	3.90	4.70	1.74	2.02	2.42	3.01	3.50
Transient Climate Response	K	0.98	1.26	1.64	2.02	2.29	1.20	1.31	1.54	1.77	1.95
Transient Climate Response to Emissions	K / TtC	1.03	1.40	1.77	2.14	2.51	1.03	1.15	1.36	1.64	1.84
2014 CO_2 Effective Radiative Forcing	W / m^2		1.69	1.80	1.91			1.67	1.79	1.93	
2014 Aerosol Effective Radiative Forcing	W / m^2		-1.37	-1.01	-0.63			-1.23	-0.95	-0.69	
2018 Ocean Heat Content rel. to 1971	ZJ		303	320	337			299	319	340	
2011 CH_4 Effective Radiative Forcing	W / m^2		0.47	0.60	0.73			0.47	0.58	0.70	
2011 N_2O Effective Radiative Forcing	W / m^2		0.14	0.17	0.20			0.14	0.17	0.20	
2011 F-Gases Effective Radiative Forcing	W / m^2		0.03	0.03	0.03			0.03	0.03	0.03	

Table S1 (cont.). Comparison of each model's probabilistic distribution with the proxy assessment. The assessed ranges are labelled as 'vll' (very-likely lower i.e. 5^{th} percentile), 'll' (likely lower, 17^{th} percentile), 'c' (central, 50^{th} percentile), 'lu' (likely upper, 83^{th} percentile) and 'vlu' (very-likely upper, 95^{th} percentile).

	Source	Assessed ranges					OSCARv3.1				
	Assessed range	vll	ll	c	lu	vlu	vll	ll	c	lu	vlu
Metric	Unit										
2000-2019 GMST rel. to 1961-1990	K	0.46		0.54		0.61	0.48		0.55		0.62
Equilibrium Climate Sensitivity	K	2.30	2.60	3.10	3.90	4.70	2.36	2.37	2.63	3.34	3.75
Transient Climate Response	K	0.98	1.26	1.64	2.02	2.29	1.41	1.52	1.62	1.81	1.96
Transient Climate Response to Emissions	K / TtC	1.03	1.40	1.77	2.14	2.51	1.15	1.25	1.41	1.62	1.82
2014 CO_2 Effective Radiative Forcing	W / m^2		1.69	1.80	1.91			1.91	1.91	1.91	
2014 Aerosol Effective Radiative Forcing	W / m^2		-1.37	-1.01	-0.63			-1.65	-1.11	-0.64	
2018 Ocean Heat Content rel. to 1971	ZJ		303	320	337			241	335	432	
2011 CH_4 Effective Radiative Forcing	W / m^2		0.47	0.60	0.73			0.50	0.50	0.50	
2011 N_2O Effective Radiative Forcing	W / m^2		0.14	0.17	0.20			0.18	0.18	0.18	
2011 F-Gases Effective Radiative Forcing	W / m^2		0.03	0.03	0.03			0.03	0.03	0.03	

Table S1 (cont.). Comparison of each model's probabilistic distribution with the proxy assessment. The assessed ranges are labelled as 'vll' (very-likely lower i.e. 5^{th} percentile), 'll' (likely lower, 17^{th} percentile), 'c' (central, 50^{th} percentile), 'lu' (likely upper, 83^{th} percentile) and 'vlu' (very-likely upper, 95^{th} percentile).

	Source	Assessed ranges					SCM4OPTv2.1				
	Assessed range	vll	ll	c	lu	vlu	vll	ll	c	lu	vlu
Metric	Unit										
2000-2019 GMST rel. to 1961-1990	K	0.46		0.54		0.61	0.43		0.59		0.79
Equilibrium Climate Sensitivity	K	2.30	2.60	3.10	3.90	4.70	2.60	2.71	3.47	4.11	4.58
Transient Climate Response	K	0.98	1.26	1.64	2.02	2.29	1.57	1.69	1.78	2.03	2.16
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			1.80	1.80	1.80	
2014 Aerosol Effective Radiative Forcing	W / m^2		-1.37	-1.01	-0.63			-1.56	-1.11	-0.78	
2018 Ocean Heat Content rel. to 1971	ZJ		303	320	337			241	321	392	
2011 CH_4 Effective Radiative Forcing	W / m^2		0.47	0.60	0.73			0.48	0.48	0.48	
2011 N_2O Effective Radiative Forcing	W / m^2		0.14	0.17	0.20			0.17	0.17	0.17	
2011 F-Gases Effective Radiative Forcing	W / m^2		0.03	0.03	0.03			0.03	0.03	0.03	

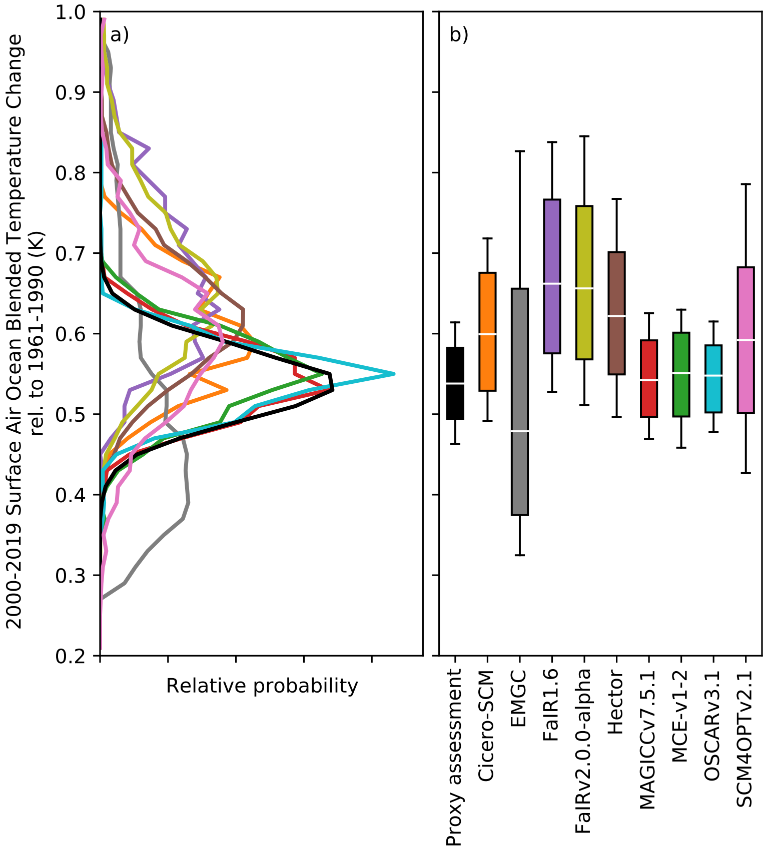

Figure S1. As in Figure 1, except for 2000-2019 mean global-mean surface temperature (GMST) change relative to 1961-1990.

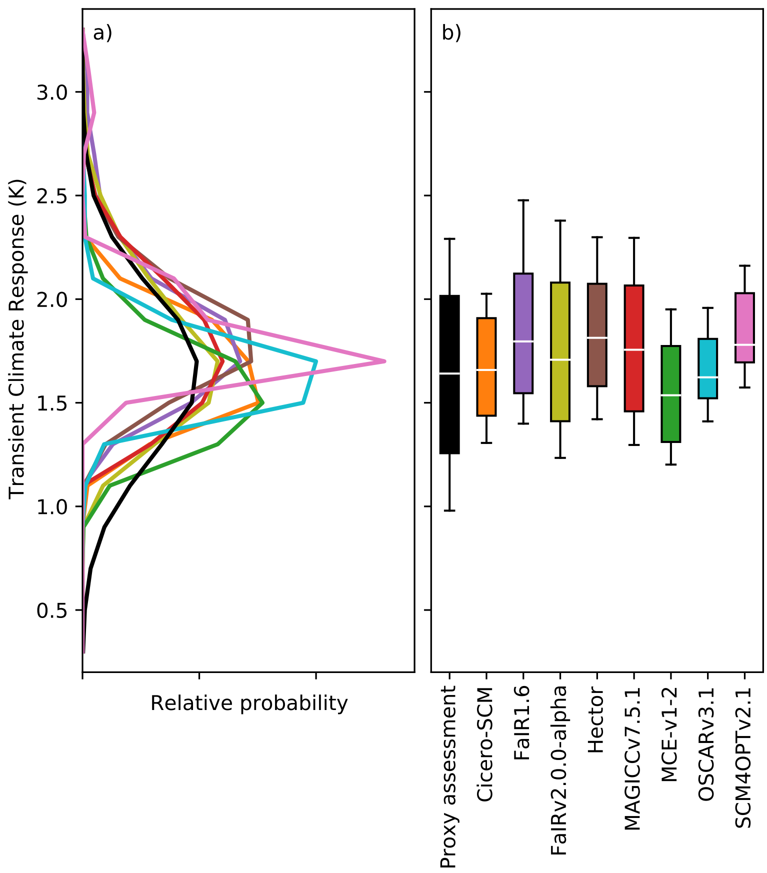

Figure S2. As in Figure 1, except for TCR.

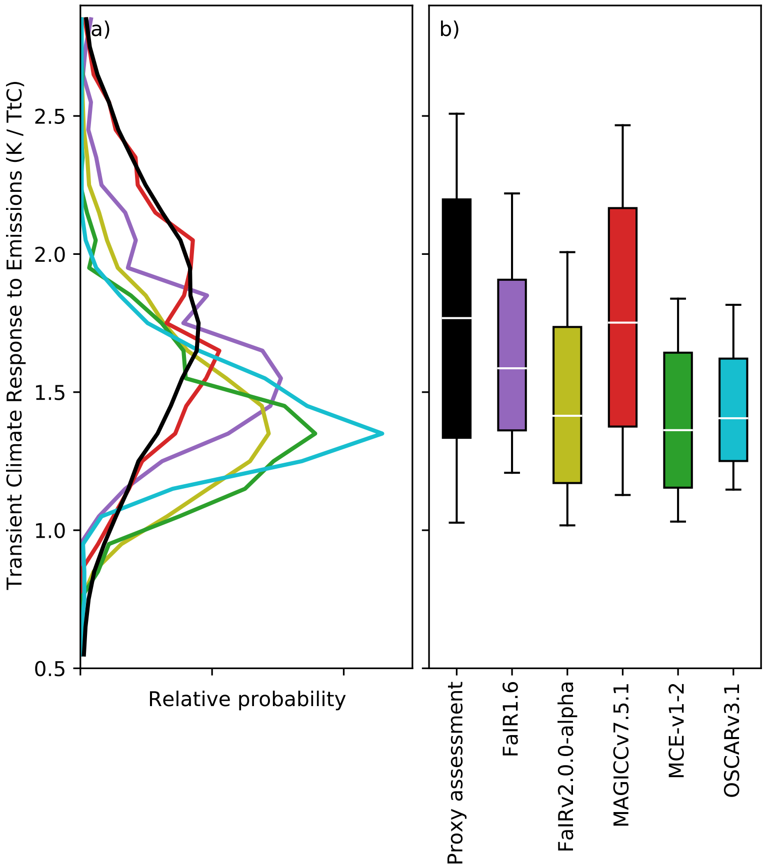

Figure S3. As in Figure 1, except for TCRE.

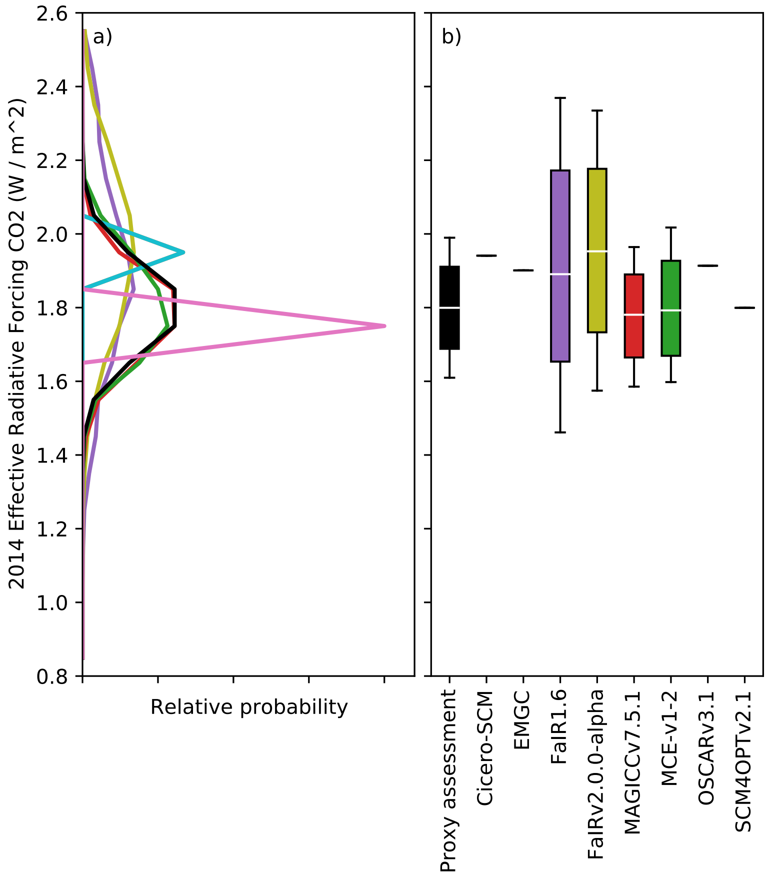

Figure S4. As in Figure 1, except for 2014 CO_2 effective radiative forcing.

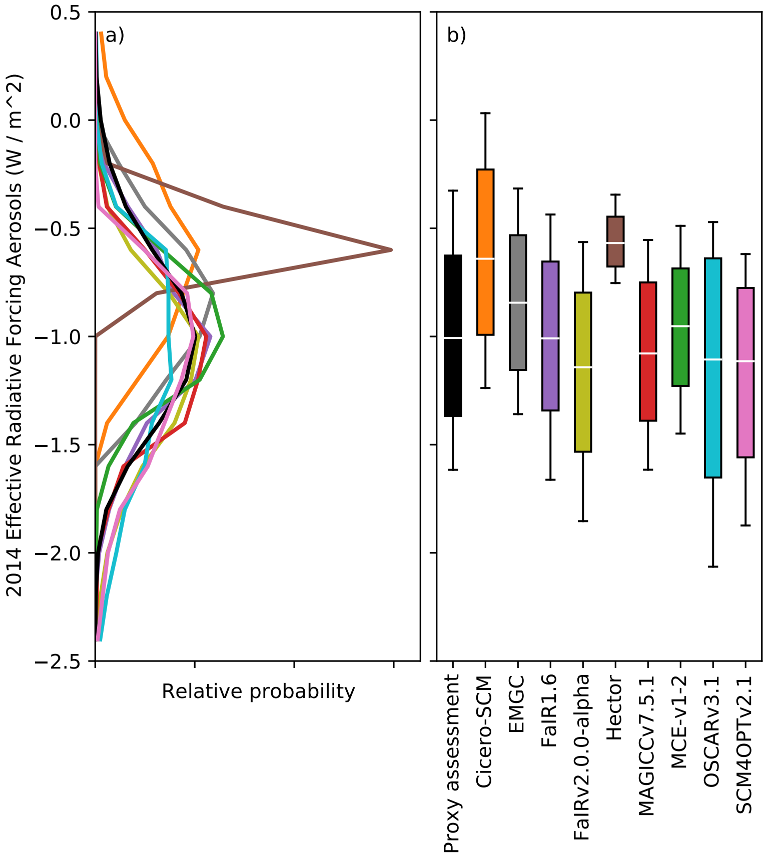

Figure S5. As in Figure 1, except for 2014 aerosol effective radiative forcing.

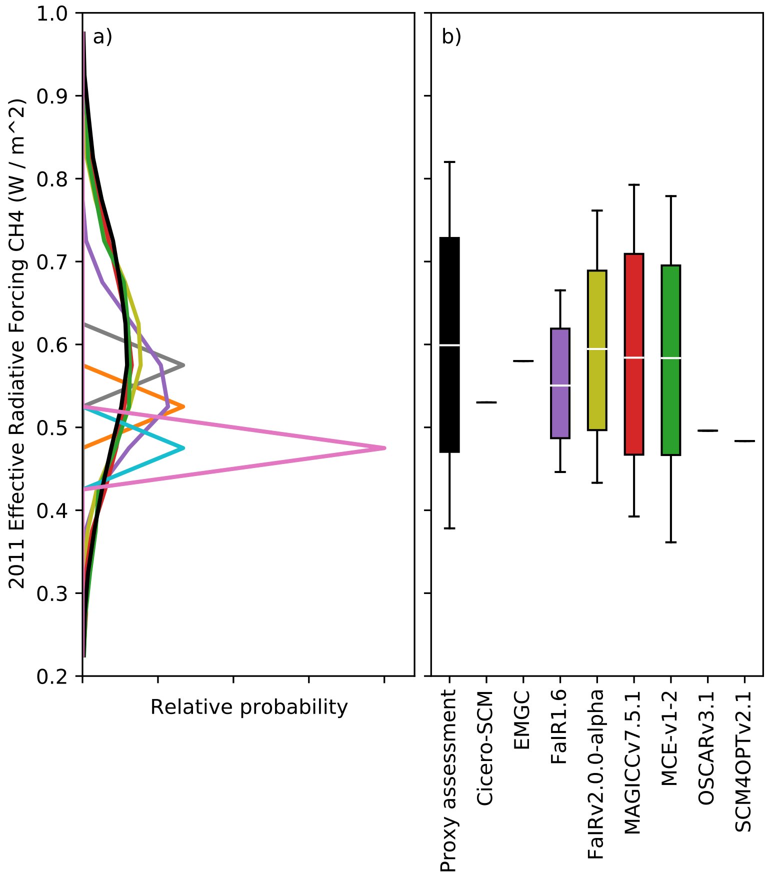

Figure S6. As in Figure 1, except for 2011 CH_4 effective radiative forcing.

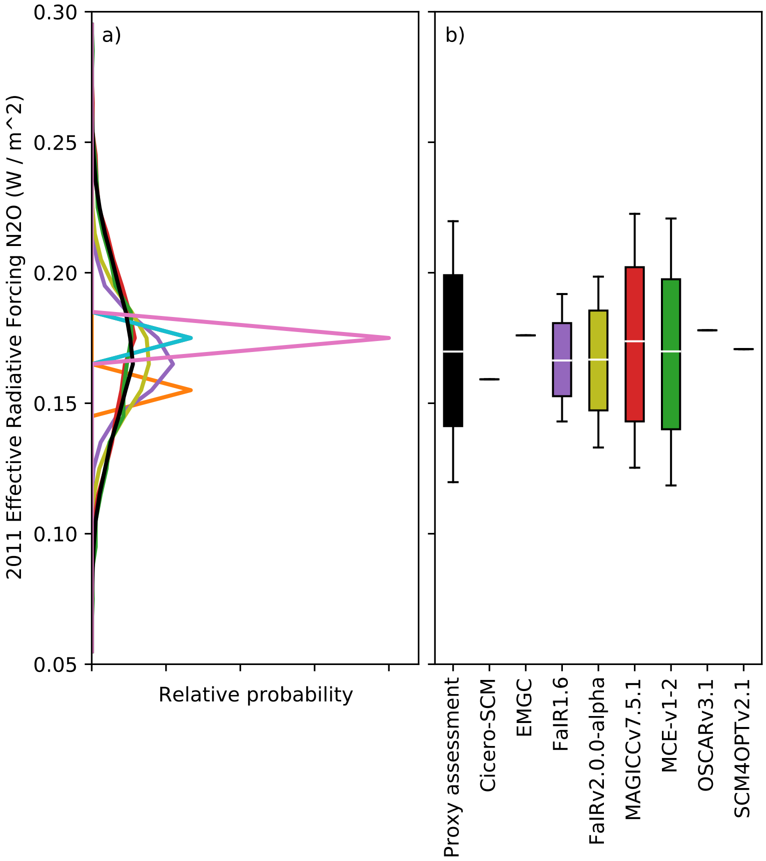

Figure S7. As in Figure 1, except for 2011 N_2O effective radiative forcing.

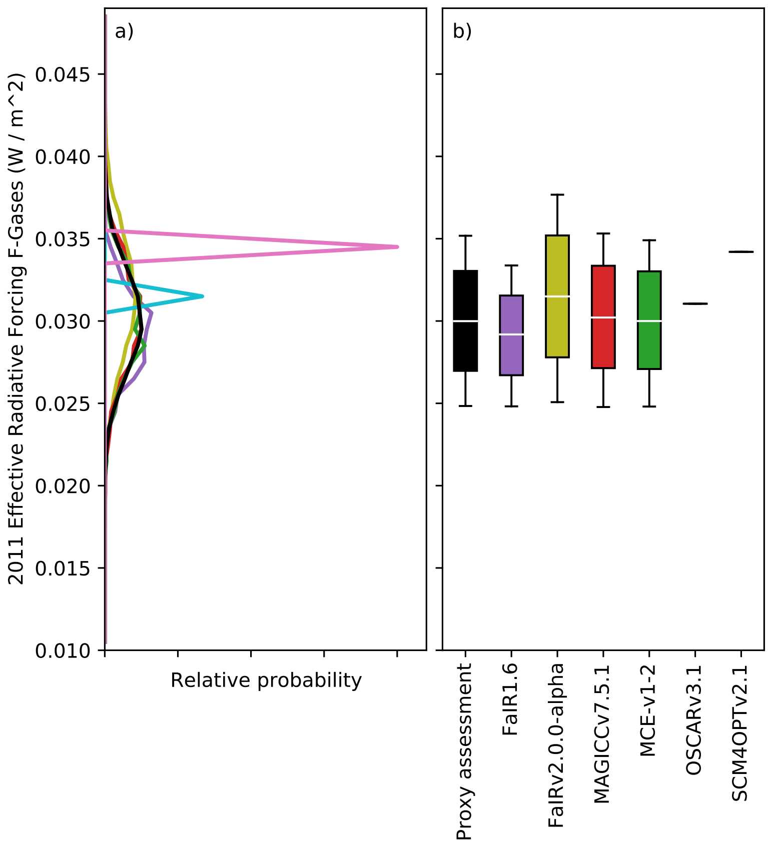

Figure S8. As in Figure 1, except for 2011 F-Gases effective radiative forcing.

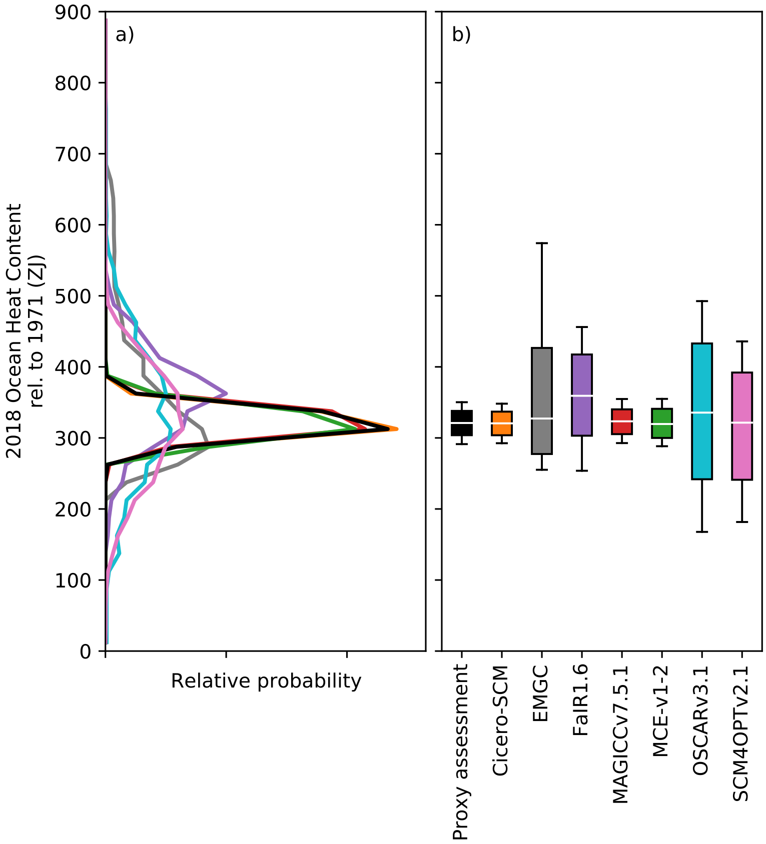

Figure S9. As in Figure 1, except for 2018 ocean heat content change relative to 1971.

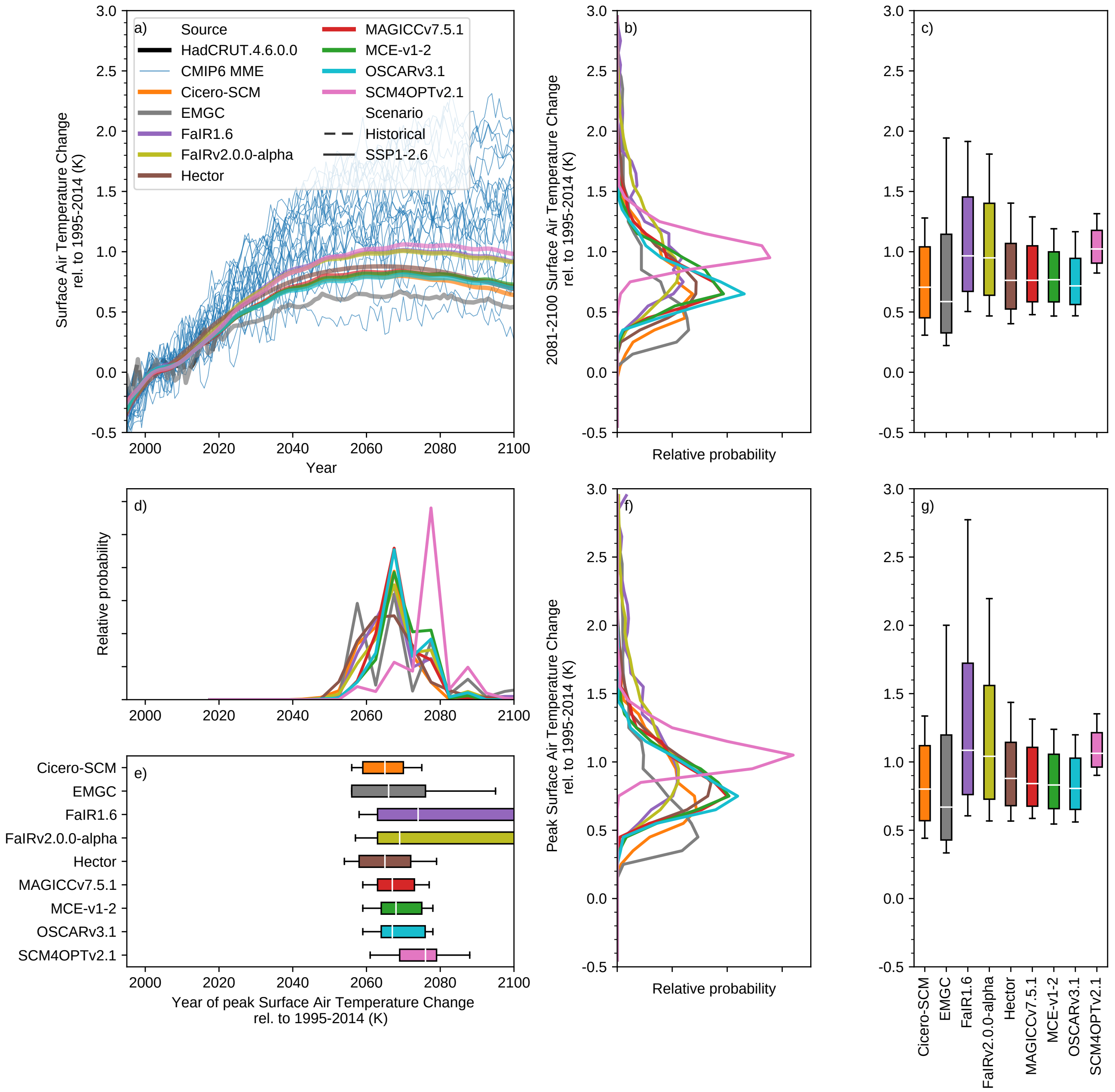

Figure S10. As in Figure 1 except for the low-emissions SSP1-2.6 scenario.

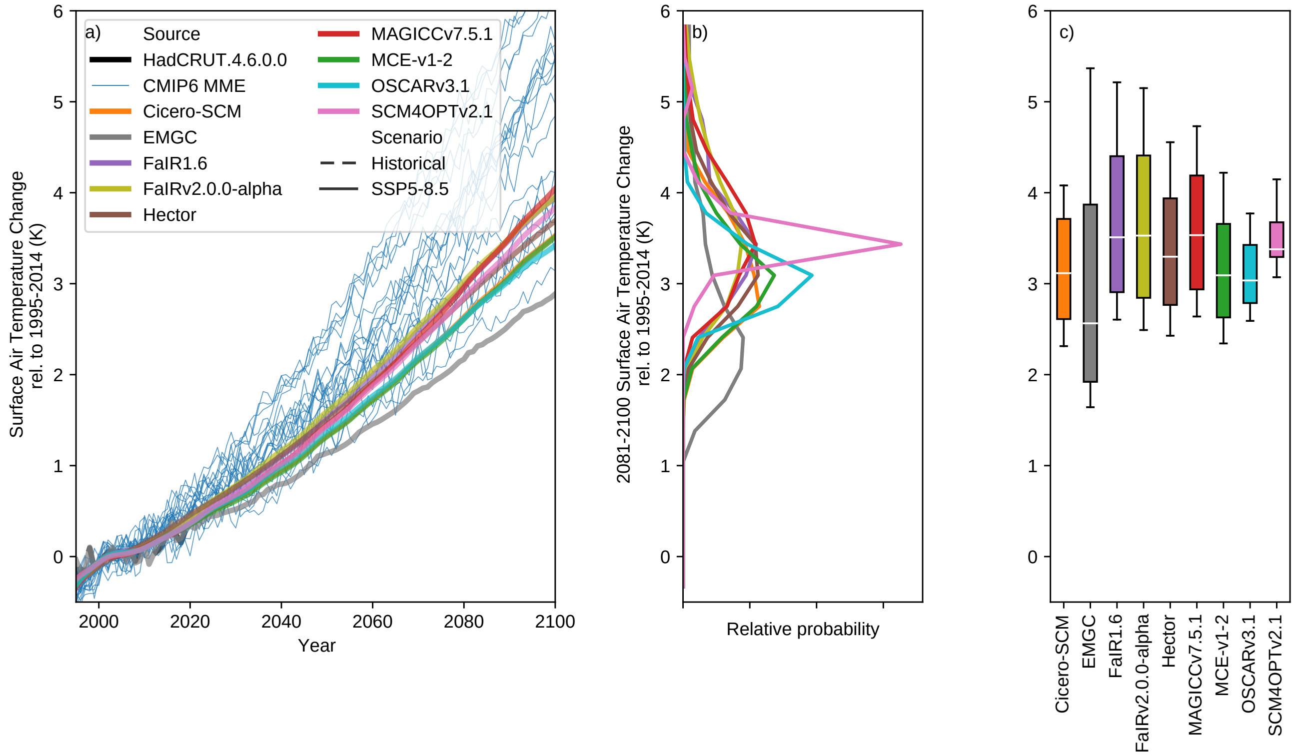

Figure S11. As in panels a), b) and c) of Figure 1 except for the high-emissions SSP5-8.5 scenario.

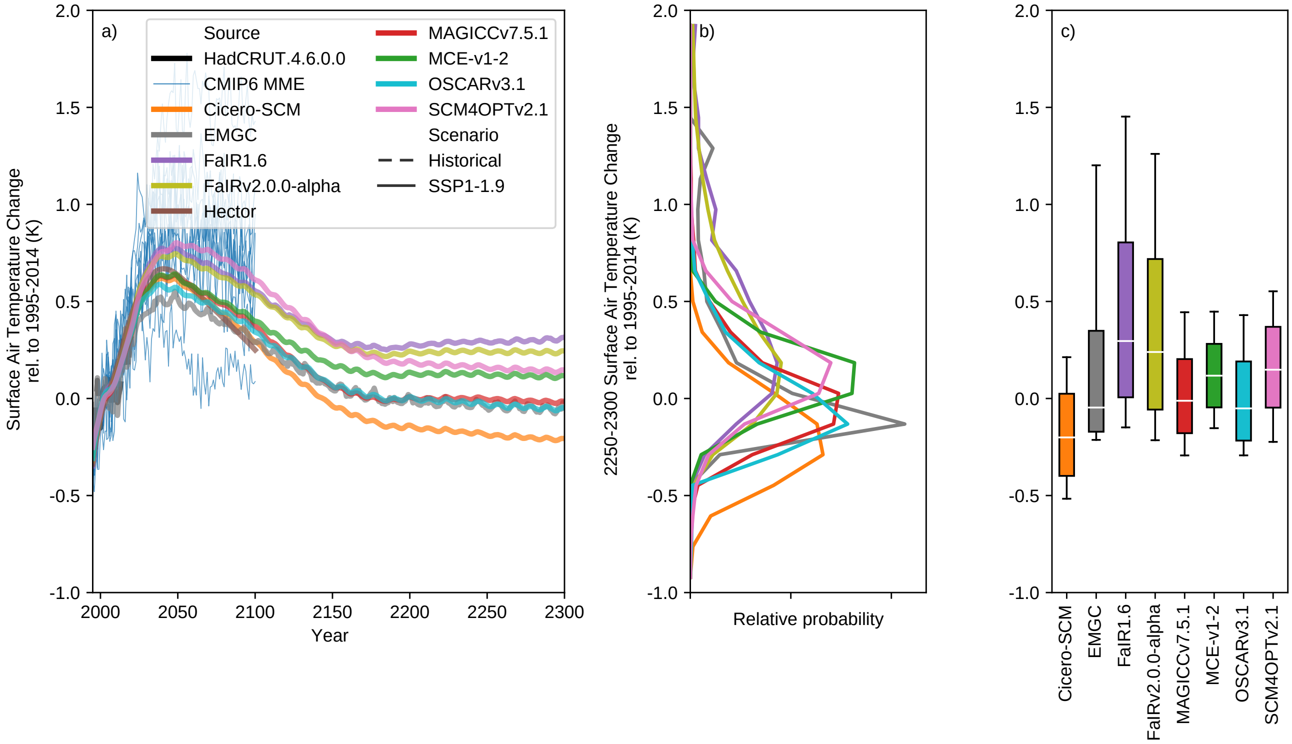

Figure S12. As in Figure 2 except for the low-emissions SSP1-1.9 scenario.

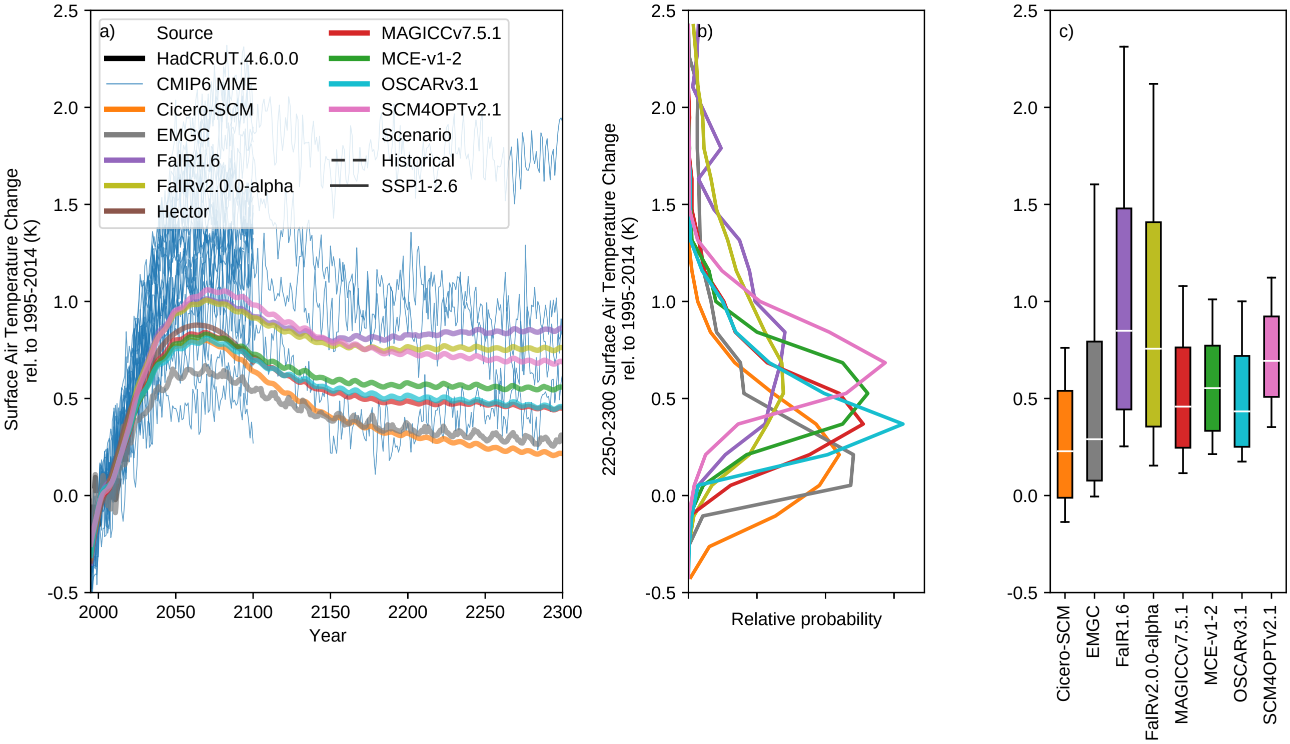

Figure S13. As in Figure 2 except for the low-emissions SSP1-2.6 scenario.

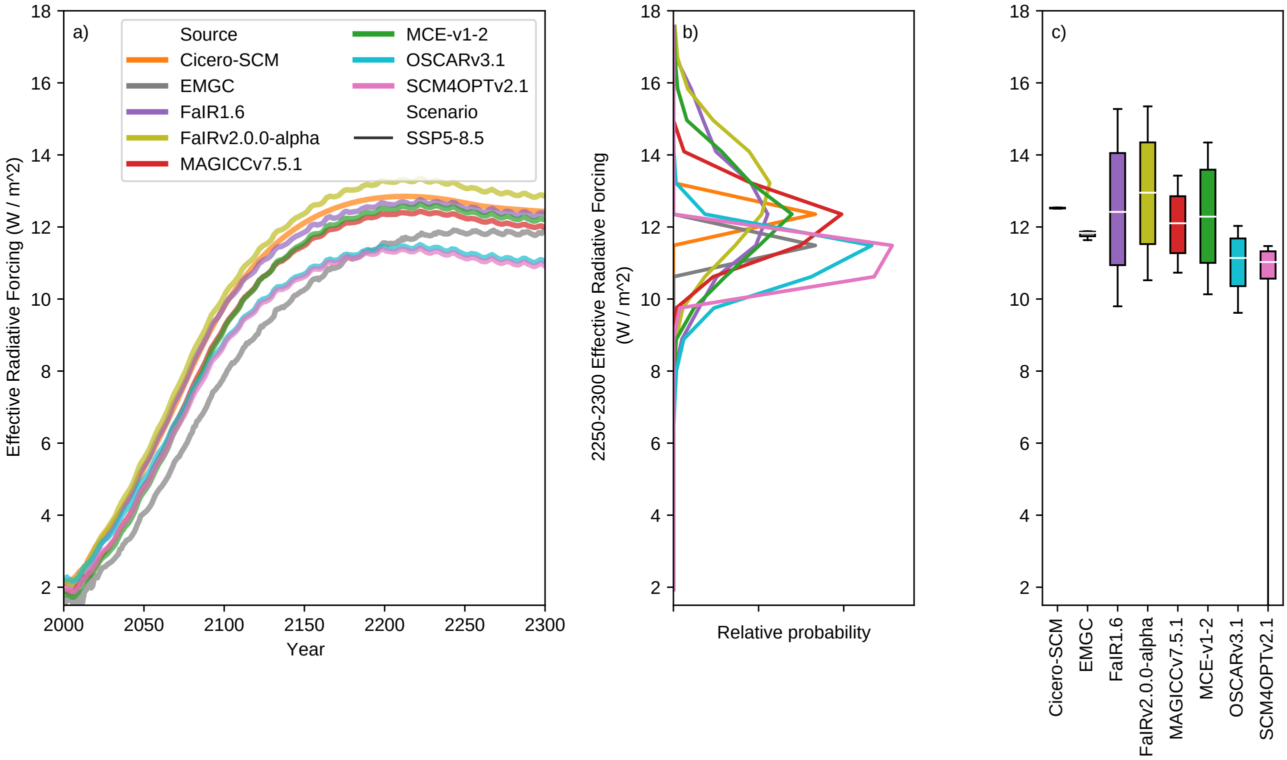

Figure S14. Long-term effective radiative forcing under the high emissions SSP5-8.5 scenario. a) Effective radiative forcing projections from 1995 to 2300 for each RCM; b) distribution of 2250-2300 mean effective radiative forcing from each RCM; c) very likely (whiskers), likely (box) and central (white line) 2250-2300 mean effective radiative forcing estimate from each RCM.

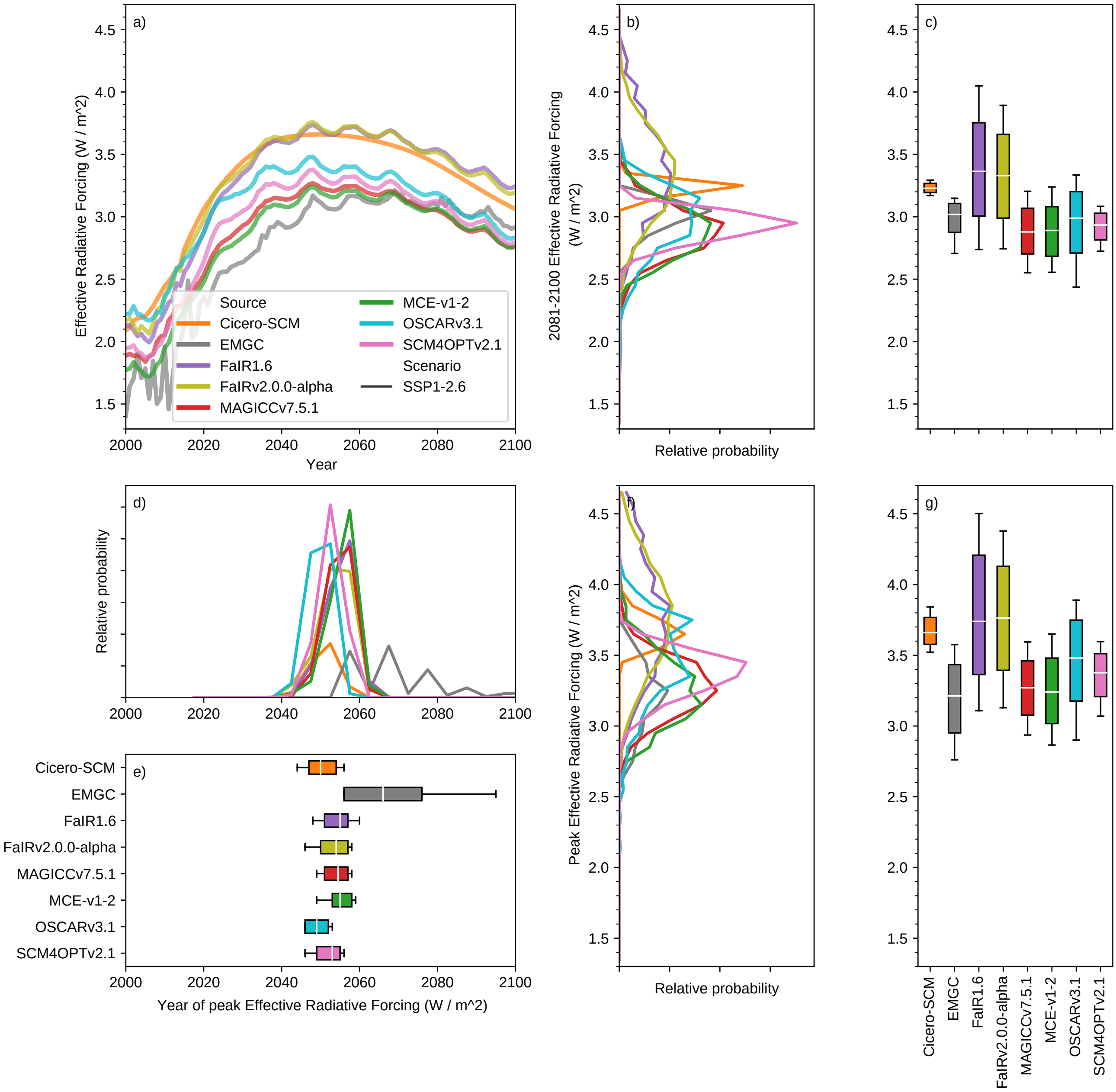

Figure S15. As in Figure 3 except for the low-emissions SSP1-2.6 scenario.

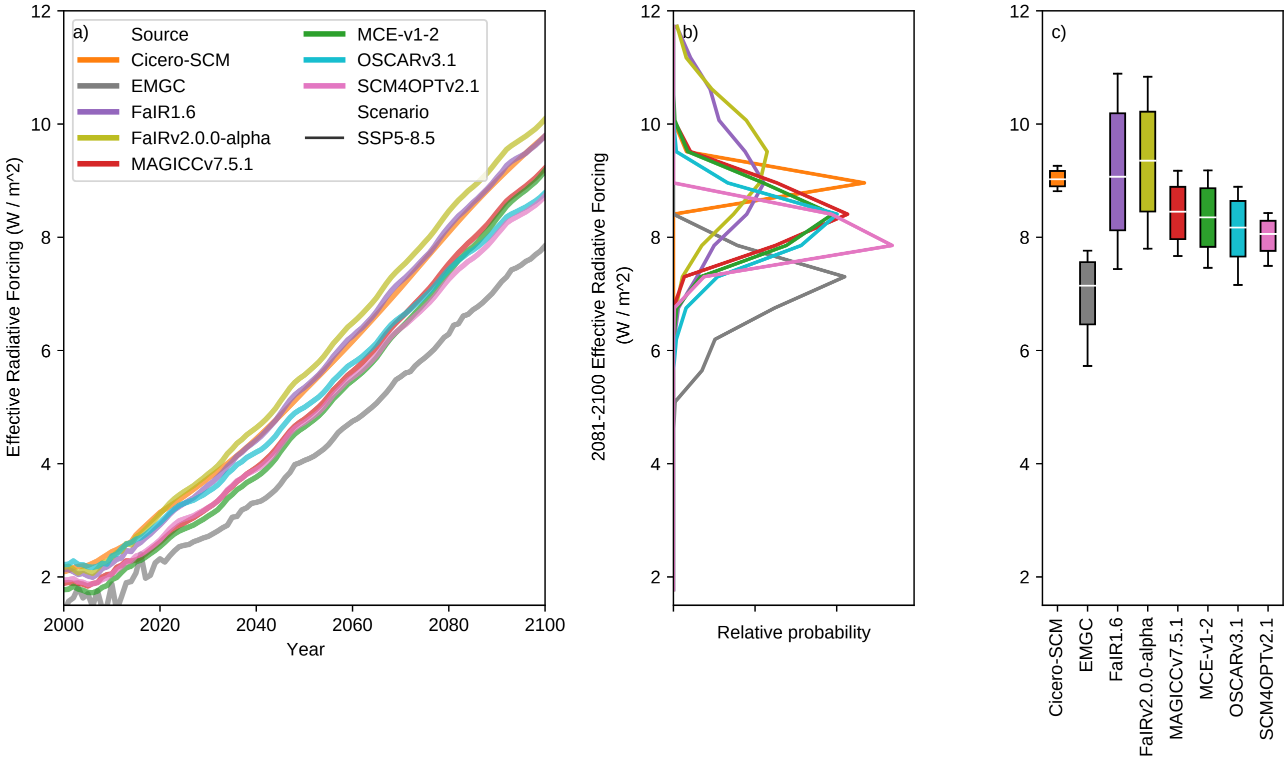

Figure S16. As in panels a), b) and c) of Figure 3 except for the high-emissions SSP5-8.5 scenario.

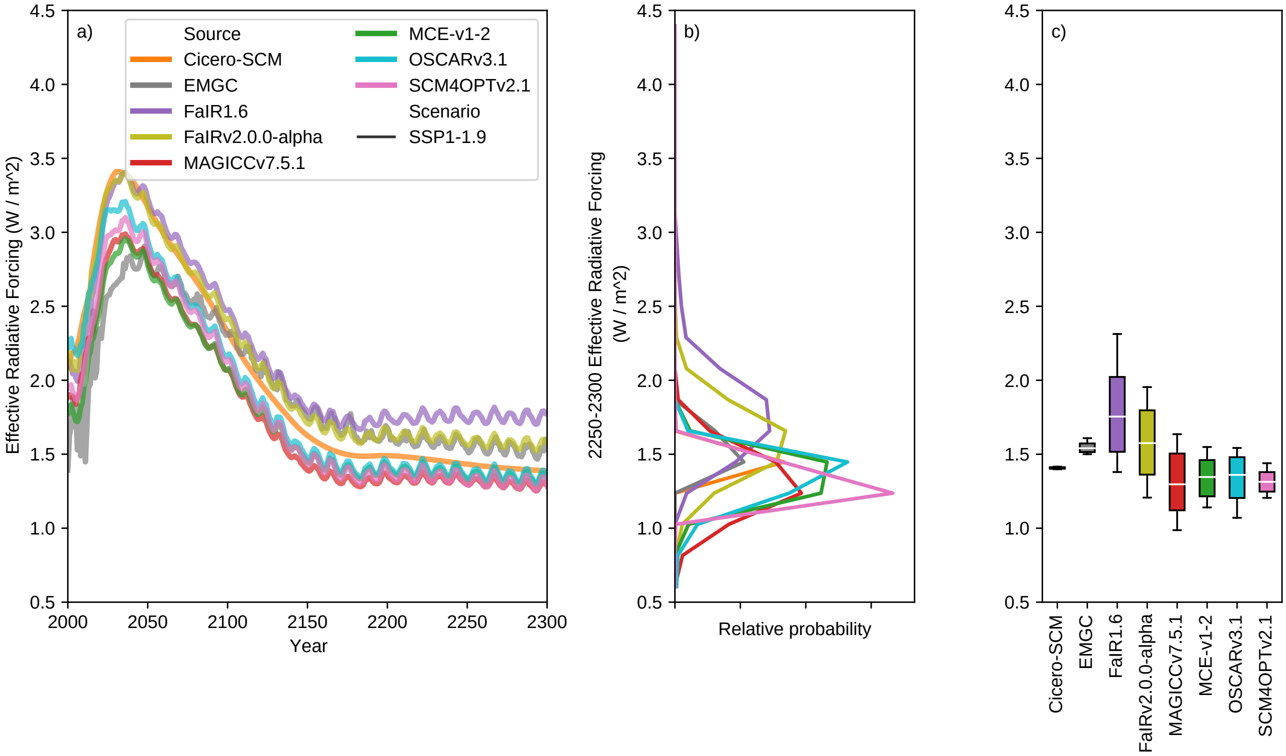

Figure S17. As in Figure S14 except for the low-emissions SSP1-1.9 scenario.

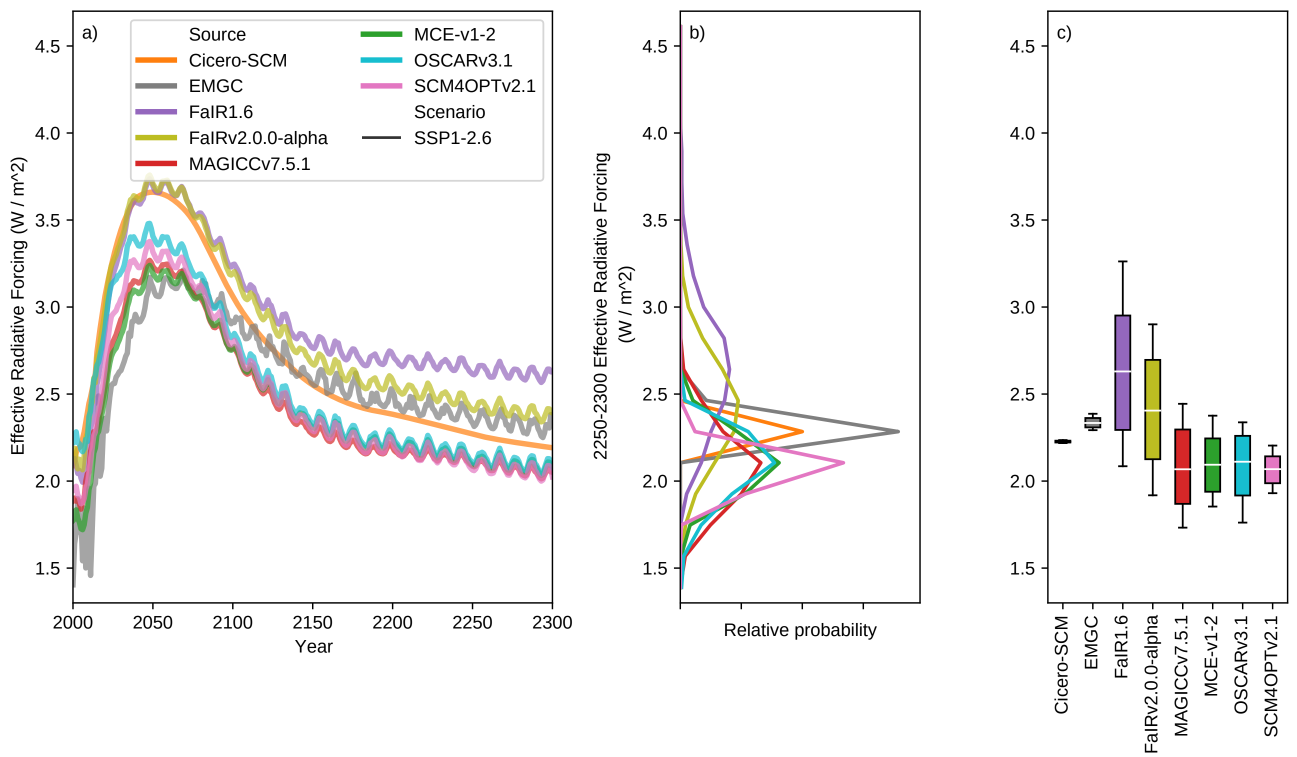

Figure S18. As in Figure S14 except for the low-emissions SSP1-2.6 scenario.

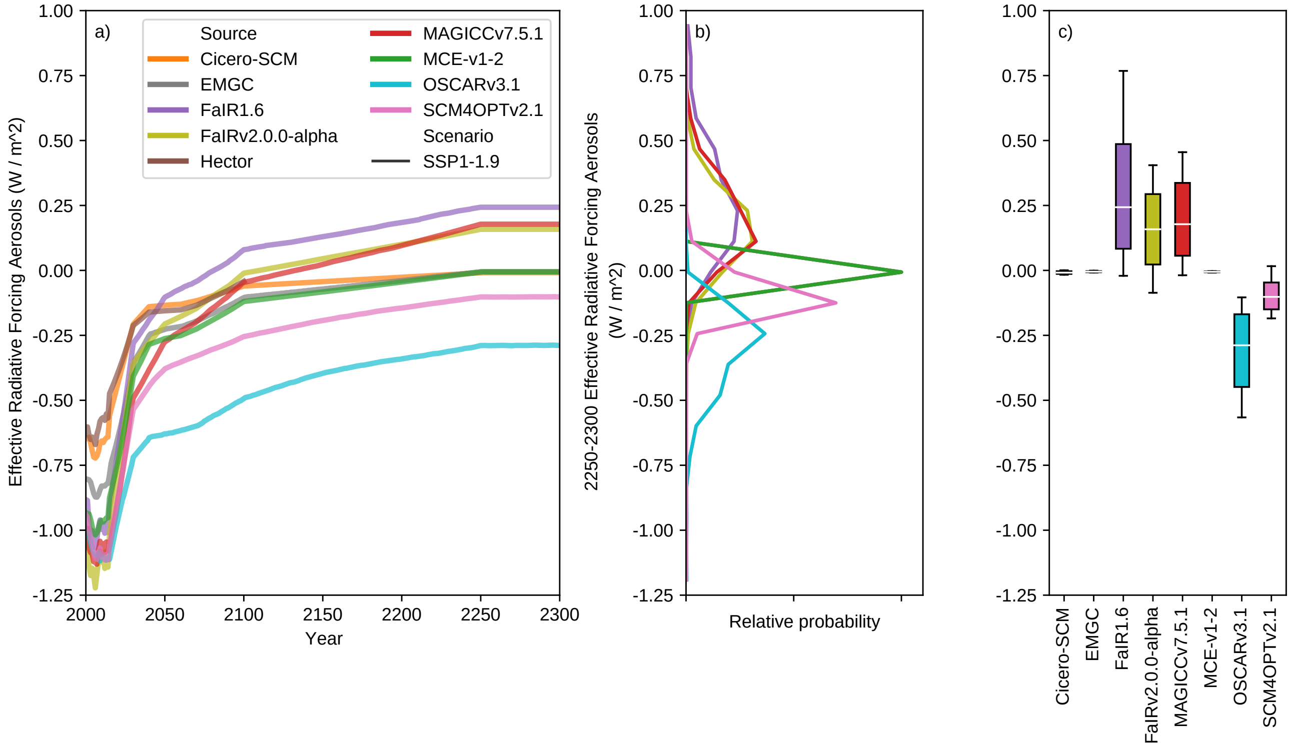

Figure S19. As in Figure S14 except for effective radiative forcing due to aerosols in the low-emissions SSP1-1.9 scenario.

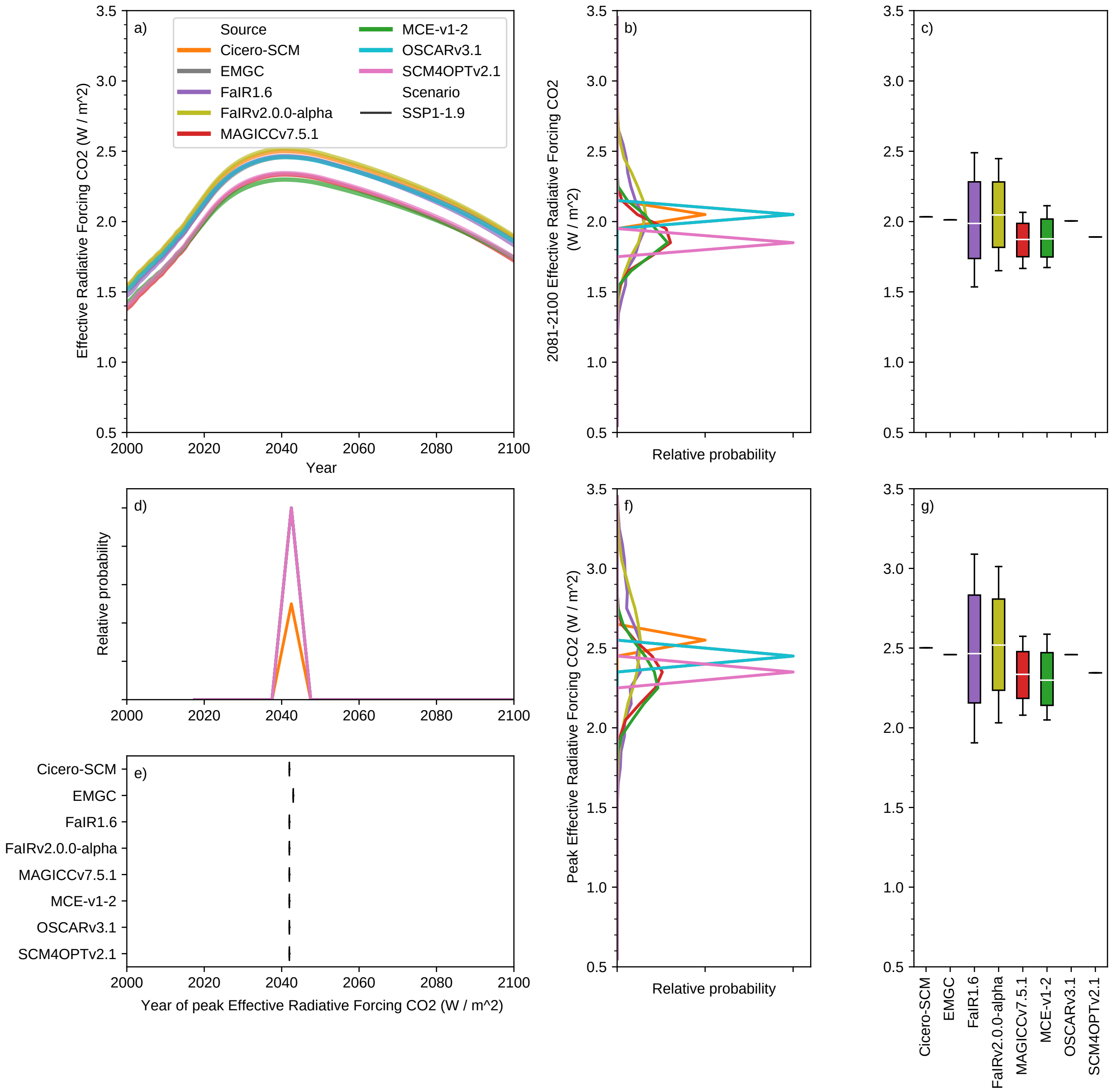

Figure S20. As Figure 3, except for effective radiative forcing due to CO_2.

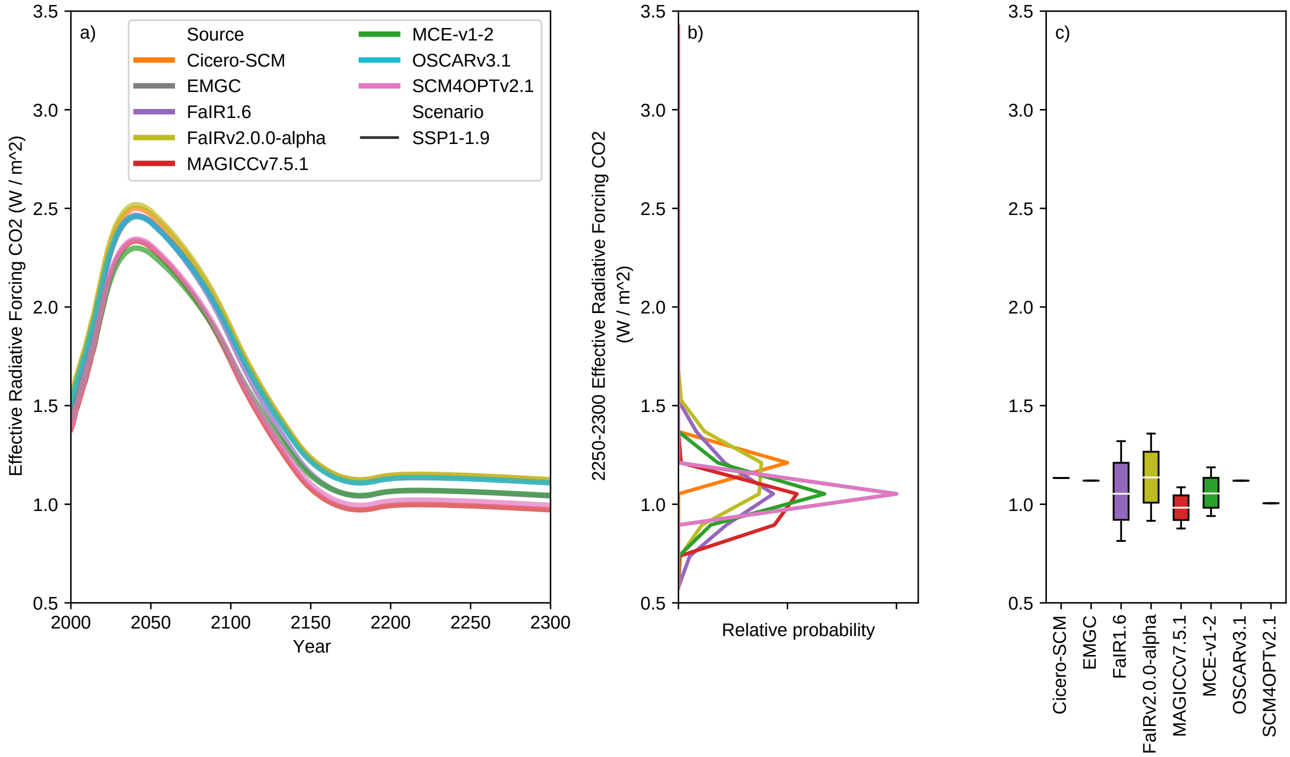

Figure S21. As in Figure S14 except for effective radiative forcing due to CO_2 in the low-emissions SSP1-1.9 scenario.

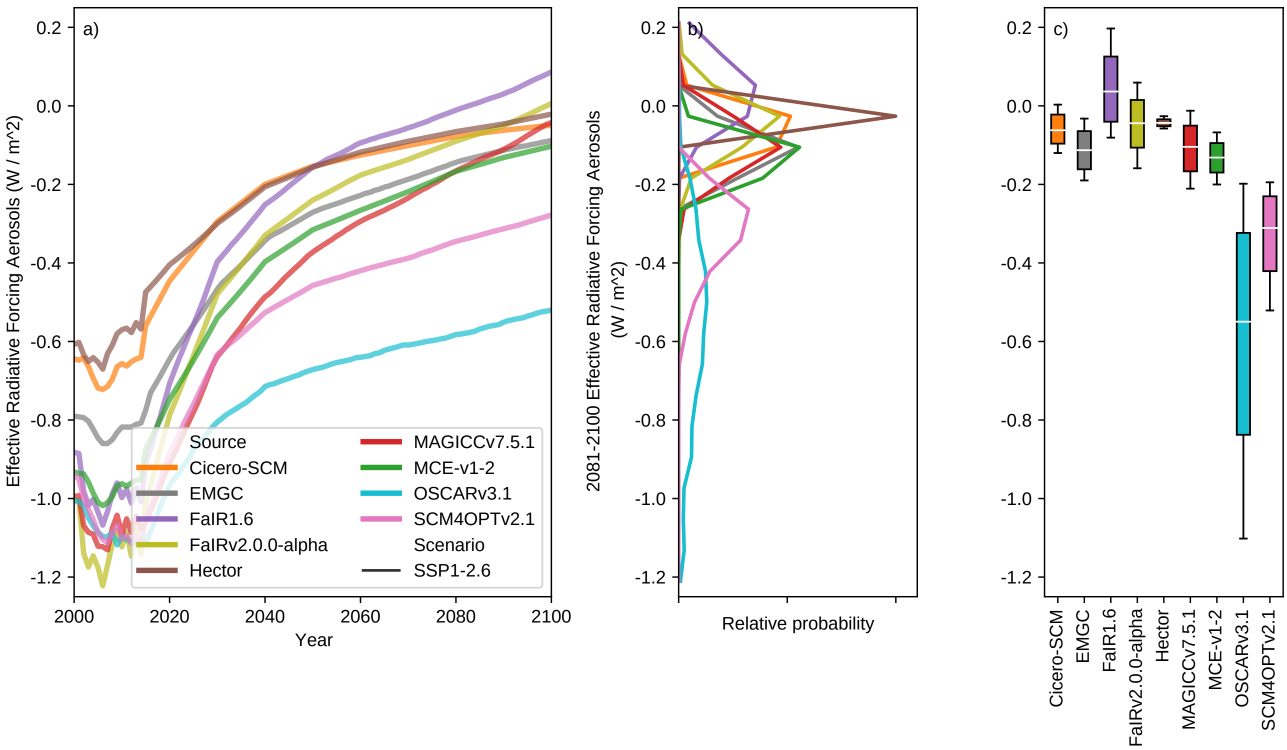

Figure S22. As in panels a), b) and c) of Figure 3, except for effective radiative forcing due to aerosols under the low-emissions SSP1-2.6 scenario.

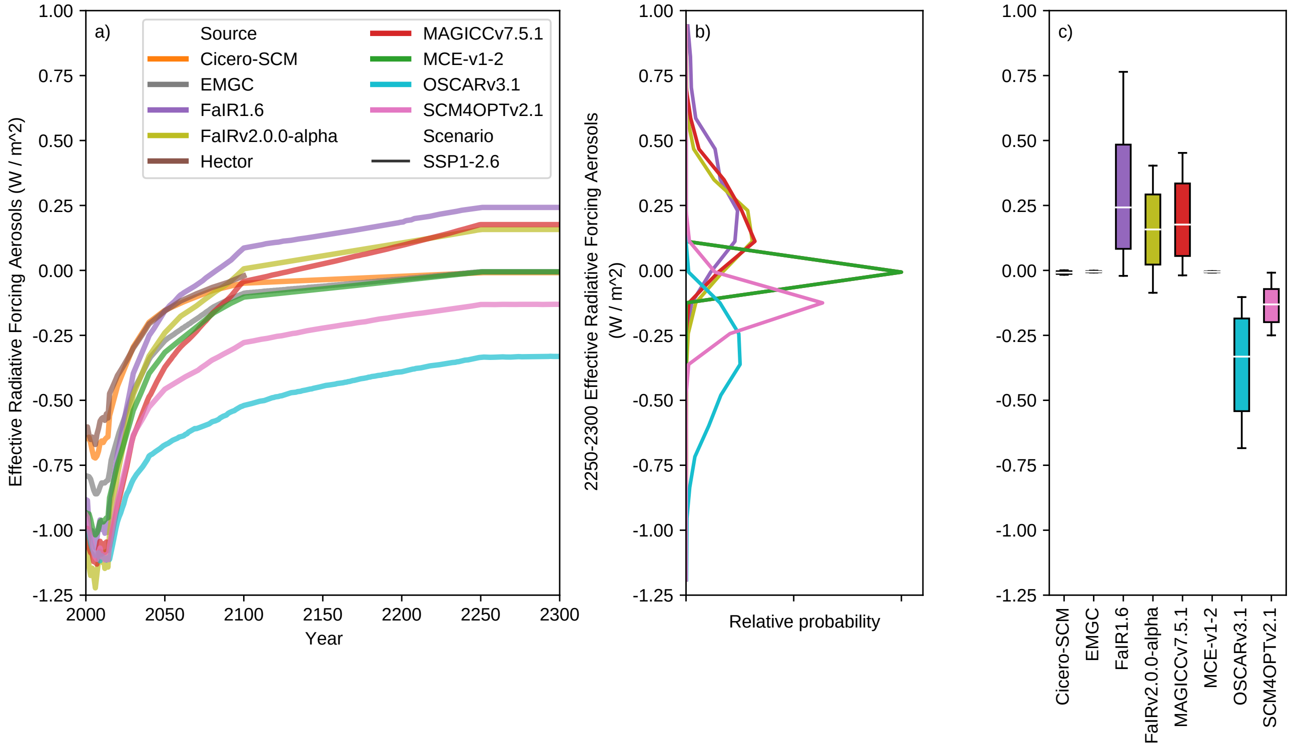

Figure S23. As in Figure S14 except for effective radiative forcing due to aerosols in the low-emissions SSP1-2.6 scenario.

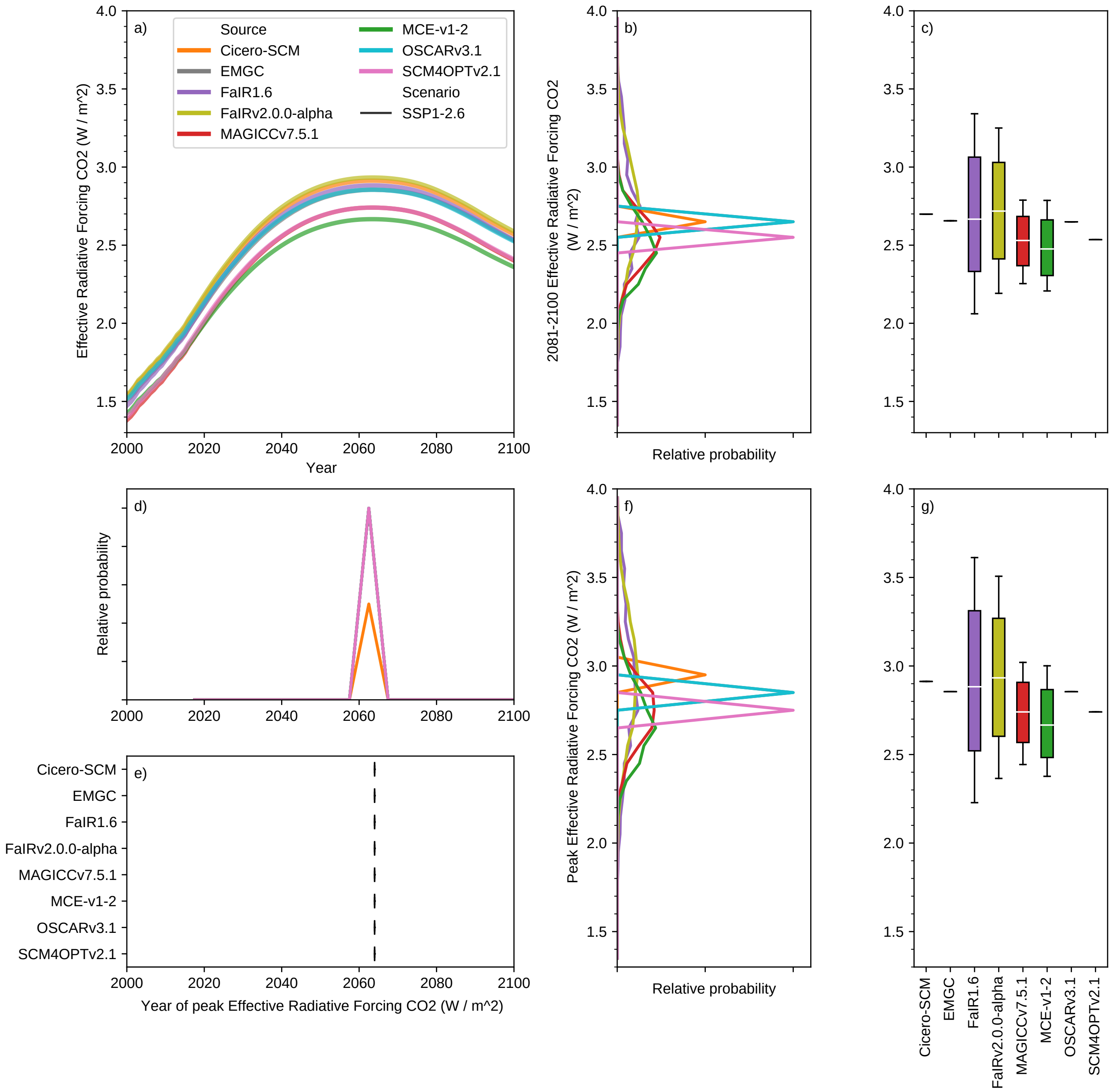

Figure S24. As Figure 3, except for effective radiative forcing due to CO_2 under the low-emissions SSP1-2.6 scenario.

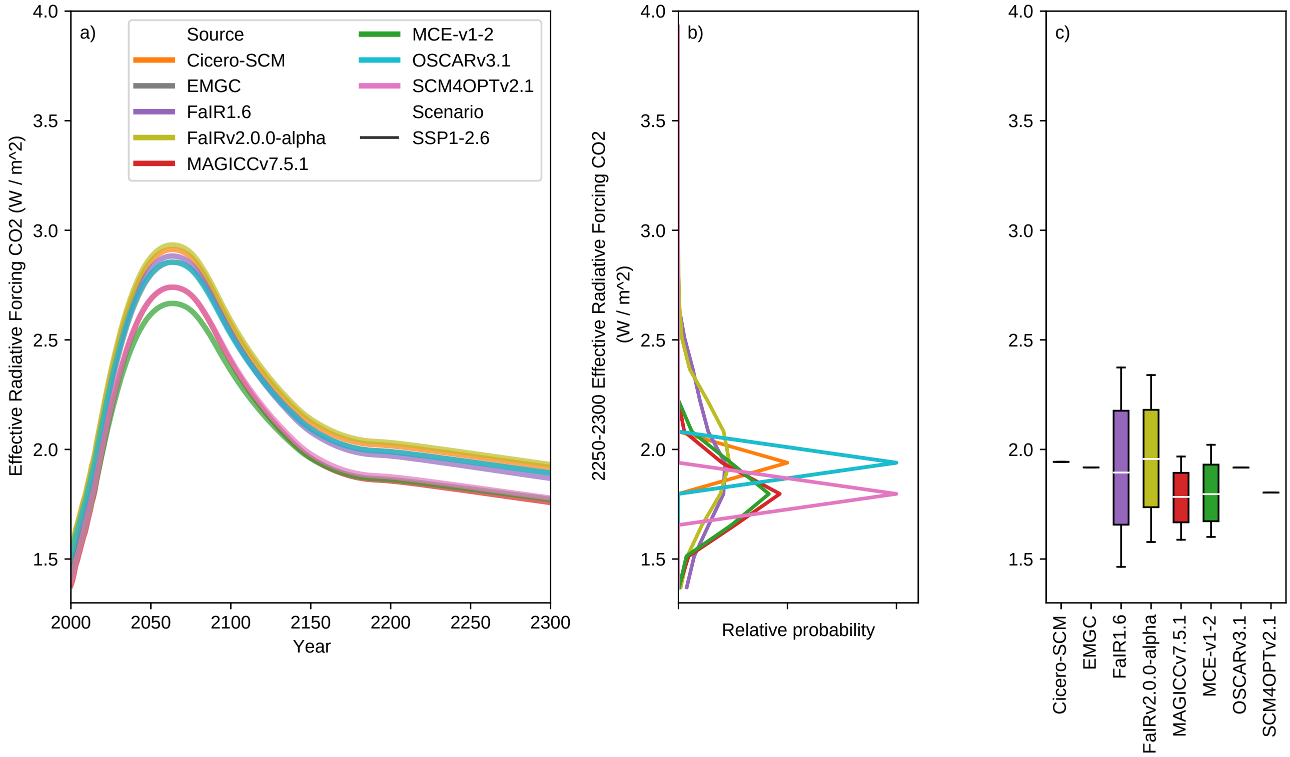

Figure S25. As in Figure S14 except for effective radiative forcing due to CO_2 in the low-emissions SSP1-2.6 scenario.

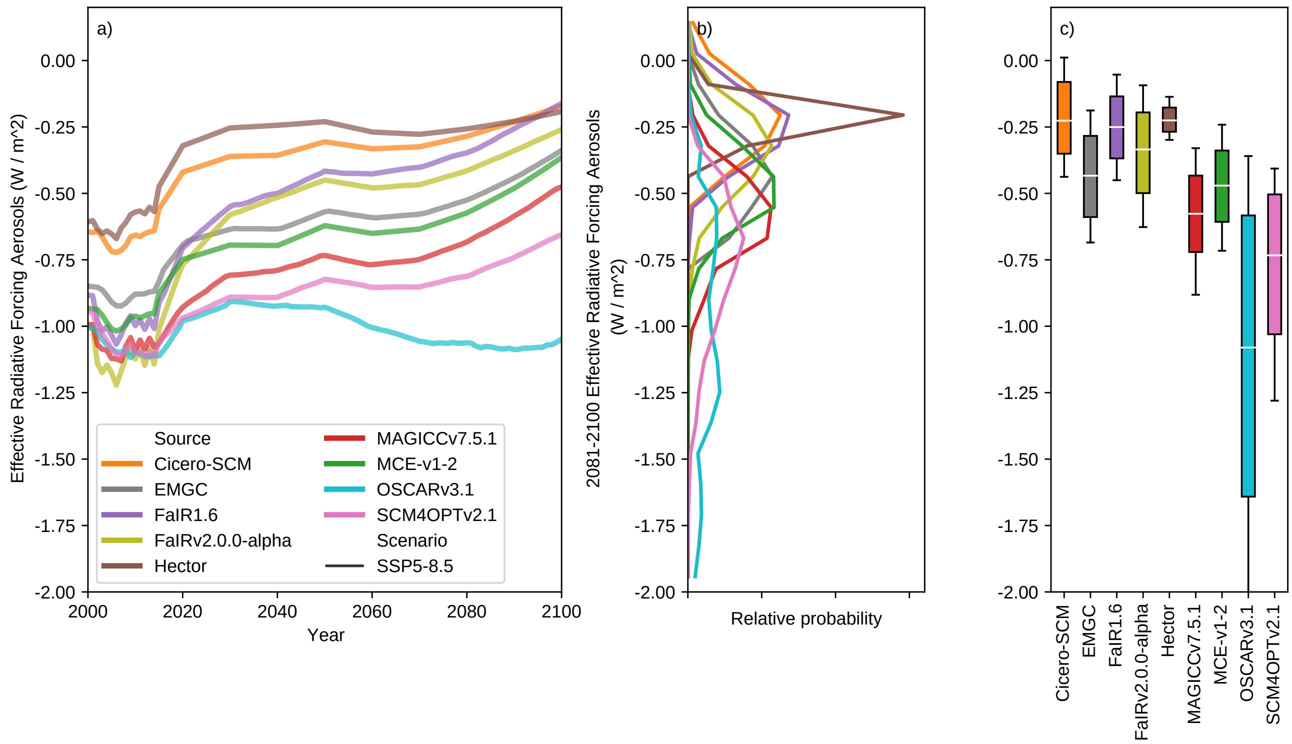

Figure S26. As in panels a), b) and c) of Figure 3, except for effective radiative forcing due to aerosols under the high-emissions SSP5-8.5 scenario.

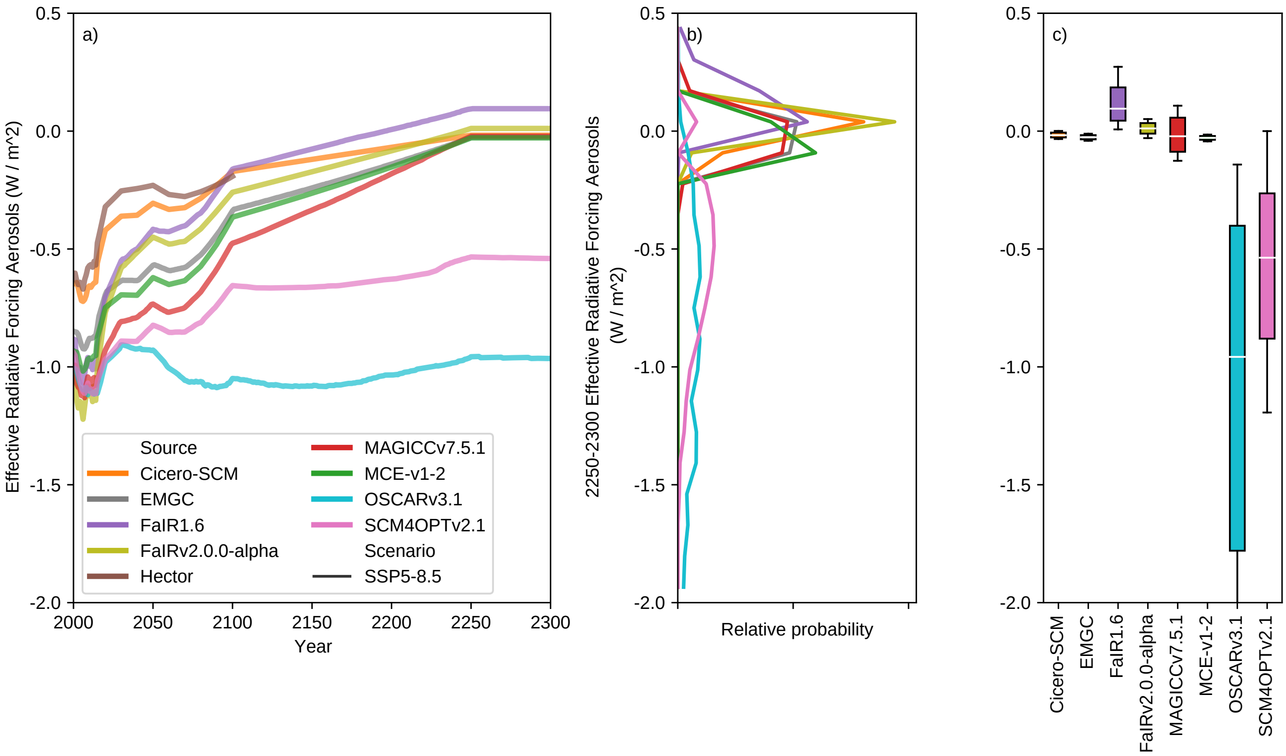

Figure S27. As in Figure S14 except for effective radiative forcing due to aerosols in the high-emissions SSP5-8.5 scenario.

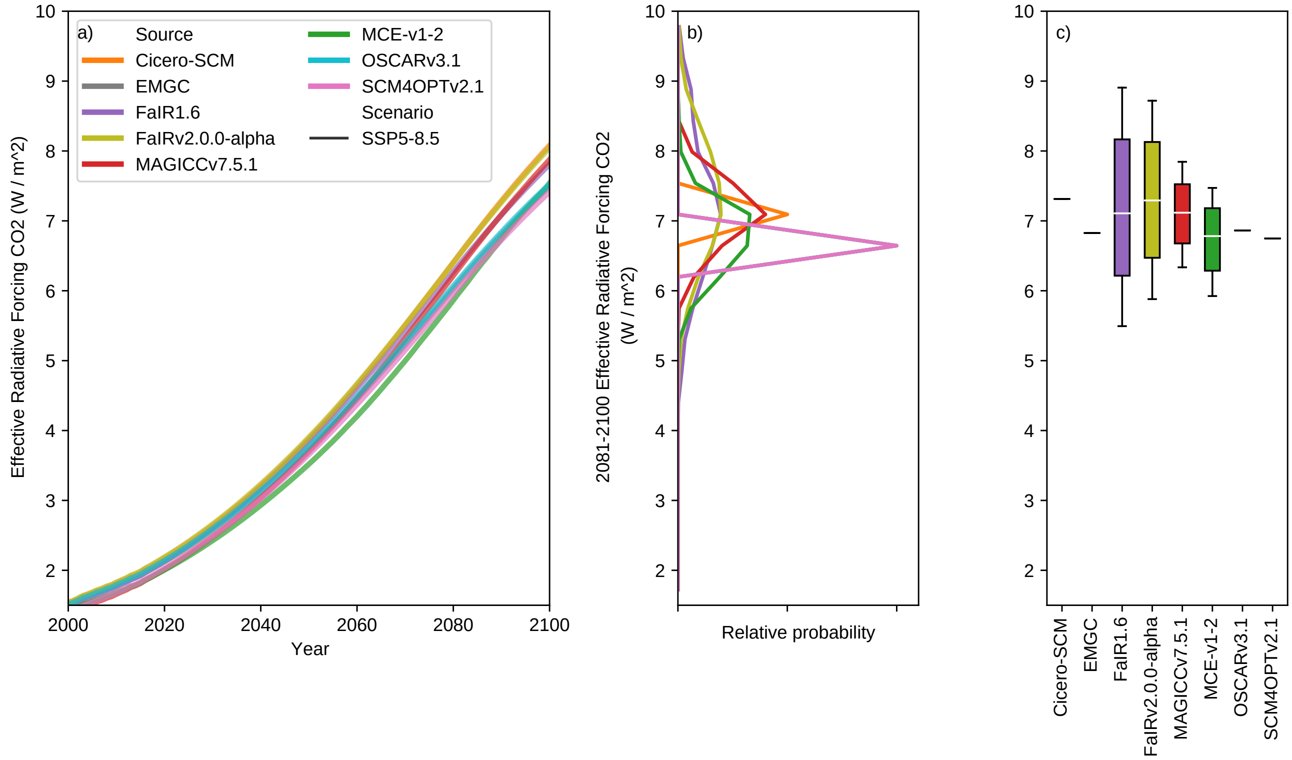

Figure S28. As in panels a), b) and c) of Figure 3, except for effective radiative forcing due to CO_2 under the high-emissions SSP5-8.5 scenario.

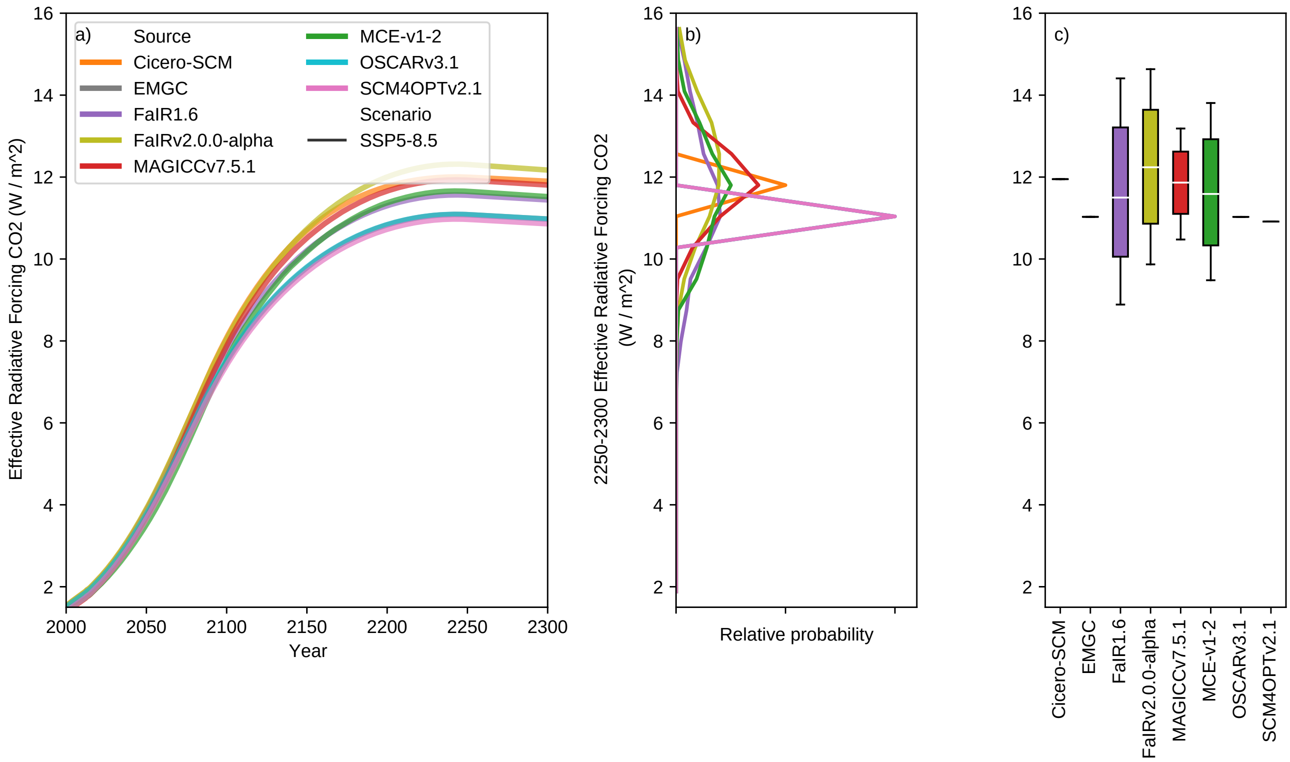

Figure S29. As in Figure S14 except for effective radiative forcing due to CO_2 in the high-emissions SSP5-8.5 scenario.

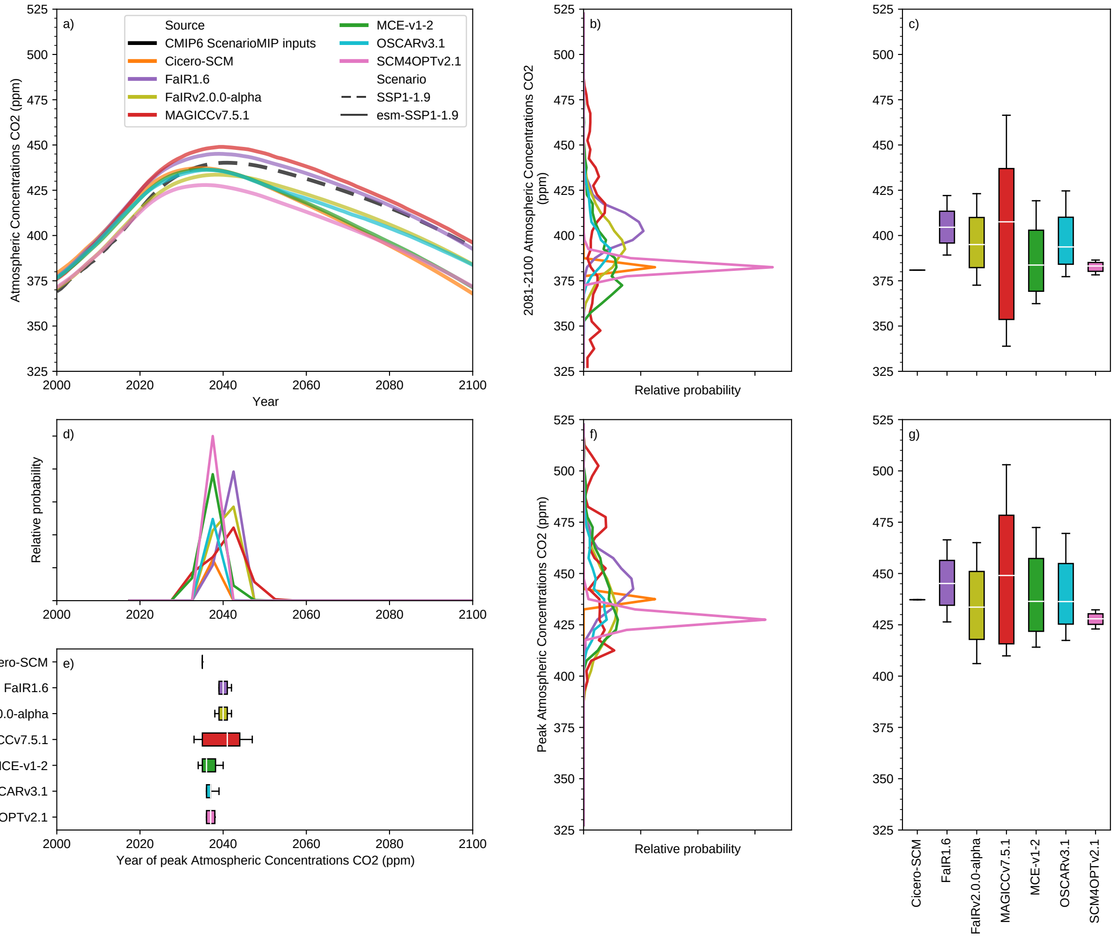

Figure S30. Atmospheric CO_2 concentration projections in the esm-SSP1-1.9 experiment. a) Atmospheric CO_2 concentration projections from 1995 to 2100. We show the median RCM projections (coloured lines) and prescribed CMIP6 ScenarioMIP input concentrations from the SSP1-1.9 concentration-driven experiment (dashed black line); 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. d) as in b) except for the year in which atmospheric CO_2 concentrations peak; e) as in c) except for the year in which atmospheric CO_2 concentrations peak; f) as in b) except for the peak atmospheric CO_2 concentrations; g) as in c) except for the peak atmospheric CO_2 concentrations. Note that FaIR1.6 data is taken from the esm-SSP1-1.9-allGHG simulations because esm-SSP1-1.9 simulations are not available.

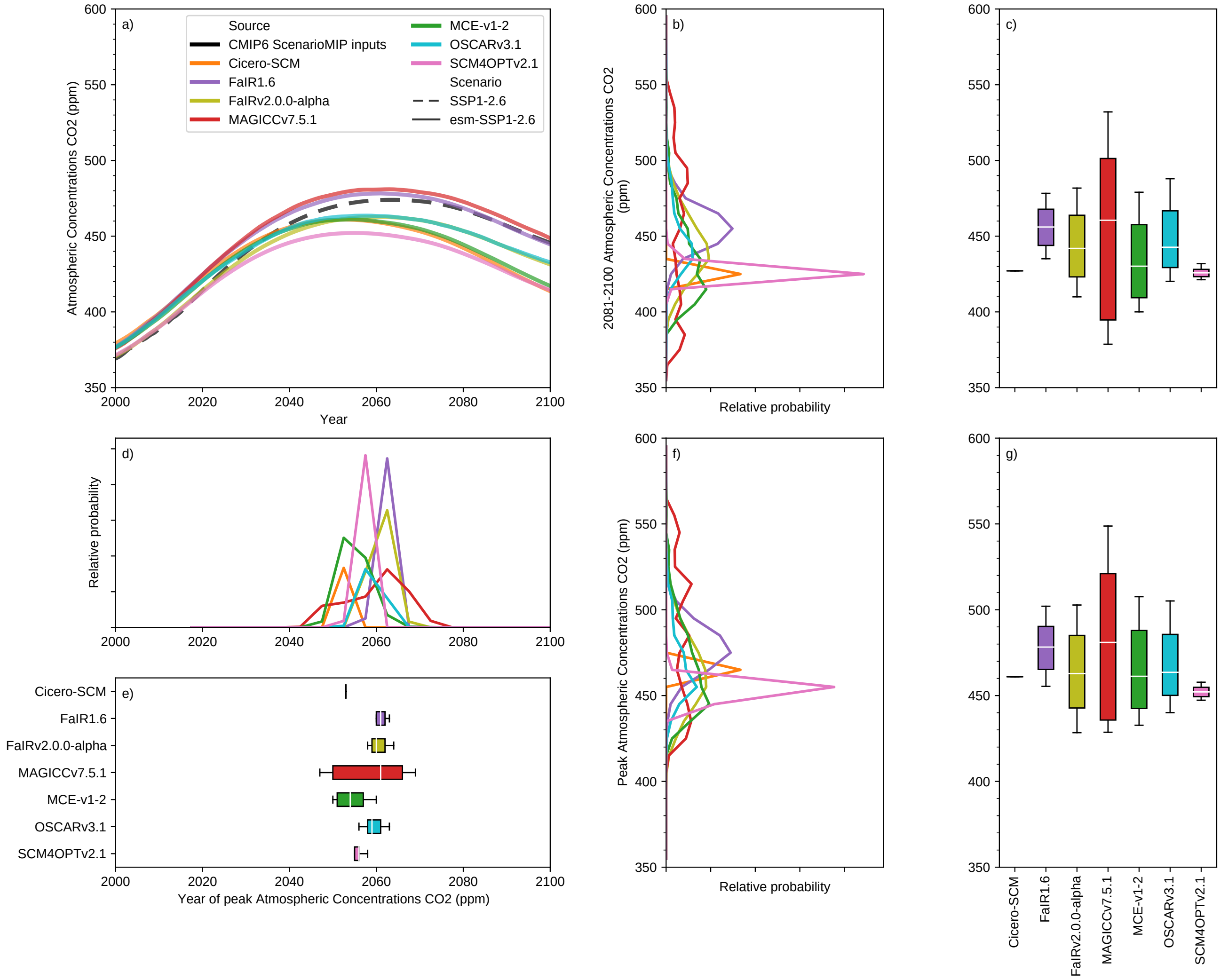

Figure S31. As in Figure S30 except for the esm-SSP1-2.6 experiment. Note that FaIR data is taken from the esm-SSP1-2.6-allGHG simulations because esm-SSP1-2.6 simulations are not available.

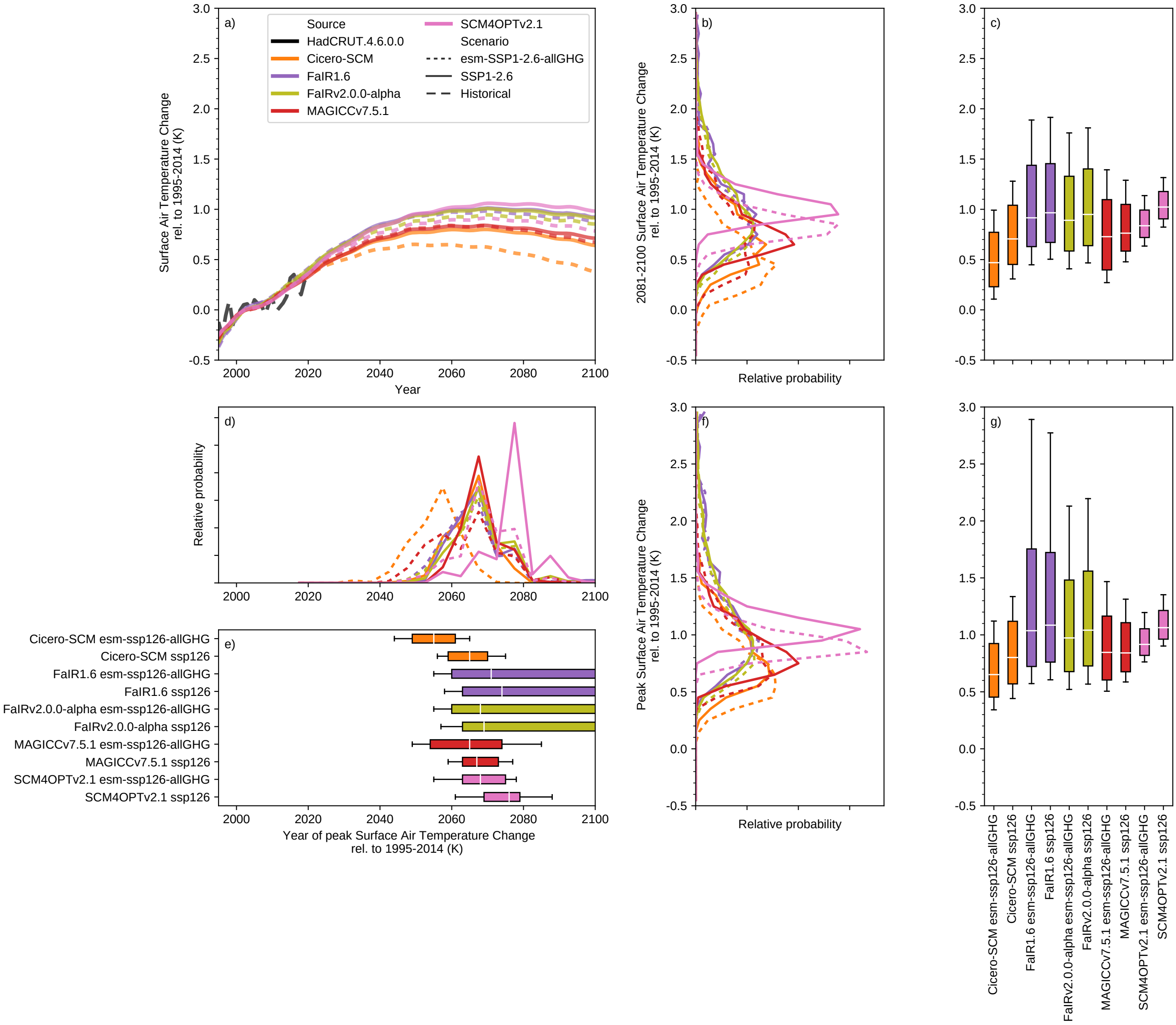

Figure S32. As in Figure 6 except for the SSP1-2.6, esm-SSP1-2.6-allGHG scenario pair.

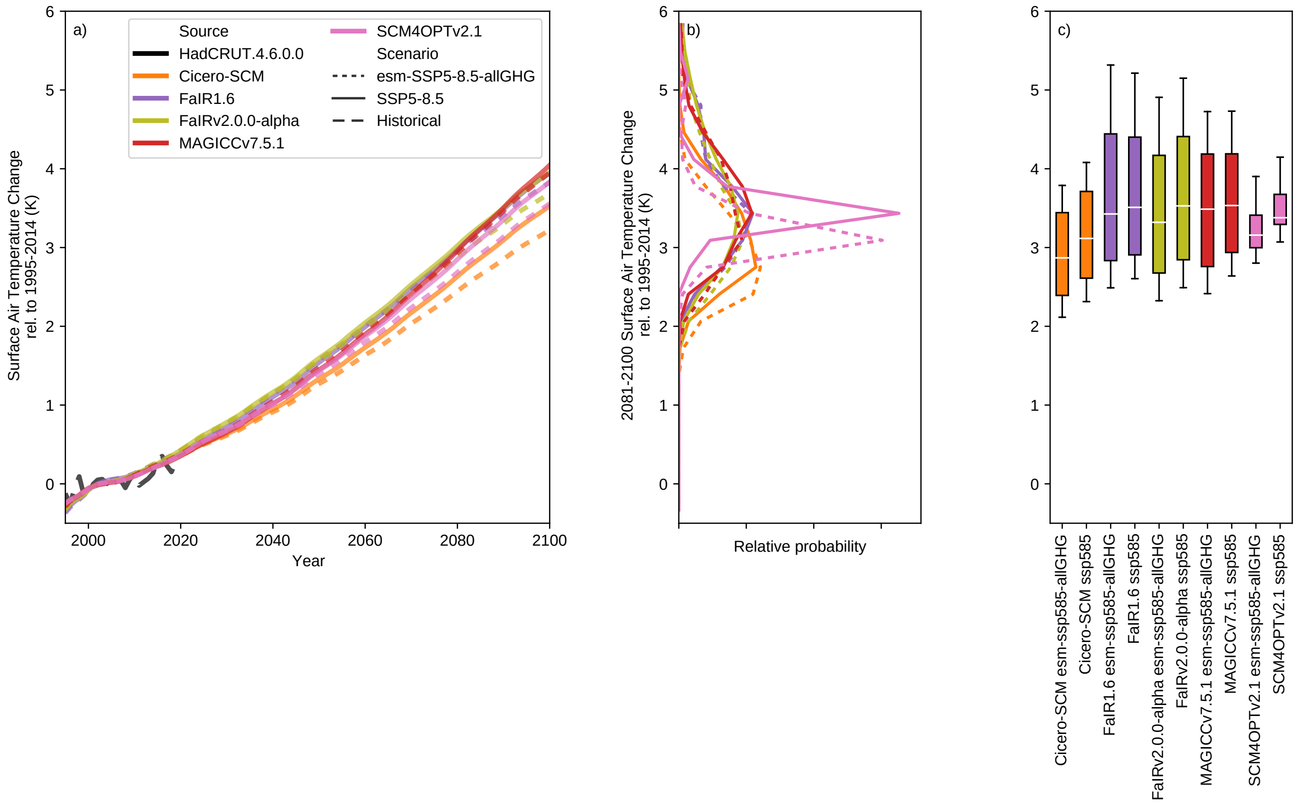

Figure S33. As in panels a), b) and c) of Figure 6 except for the SSP5-8.5, esm-SSP5-8.5-allGHG scenario pair.

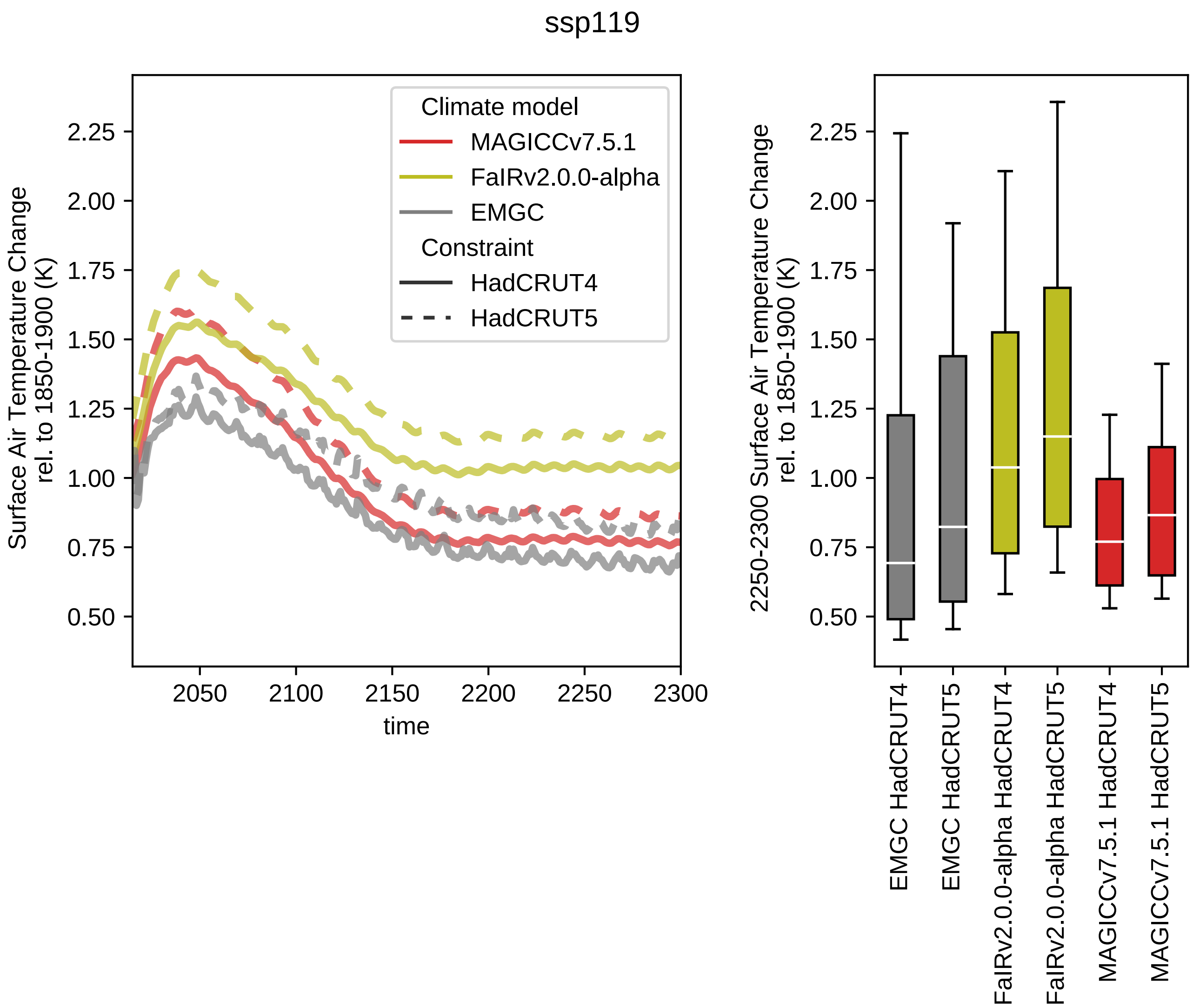

Figure S34. Surface air temperature (also referred to as global-mean surface air temperature, GSAT) change in the concentration-driven SSP1-1.9 experiment. For each model, two different probabilistic distributions are shown. One is constrained to HadCRUT.4.6.0.0 (Morice et al., 2012, as used in the main study), the second is constrained to HadCRUT.5.0.1.0 (Morice et al., 2021), which makes higher estimates of historical-warming. a) GSAT projections from 1995 to 2300. We show the median RCM projections for the probabilistic distributions which used HadCRUT.4.6.0.0 (solid lines) and HadCRUT.5.0.1.0 (dashed lines) as constraints; b) very likely (whiskers), likely (box) and central (white line) 2250-2300 mean GSAT for each RCM for each probabilistic distribution. All results are shown relative to the 1850-1900 reference period.

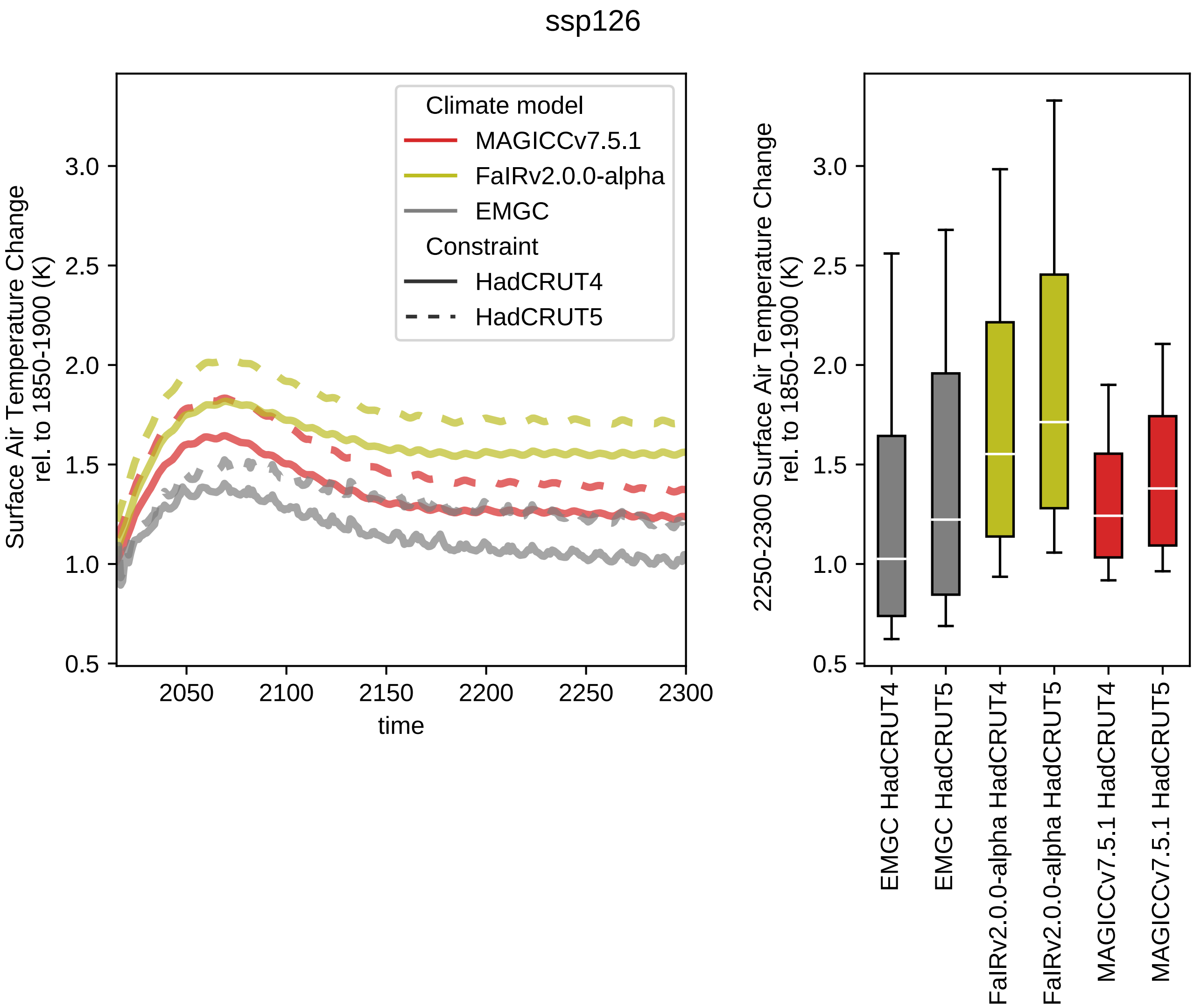

Figure S35. As in Figure S35, except for the concentration-driven SSP1-2.6 experiment.

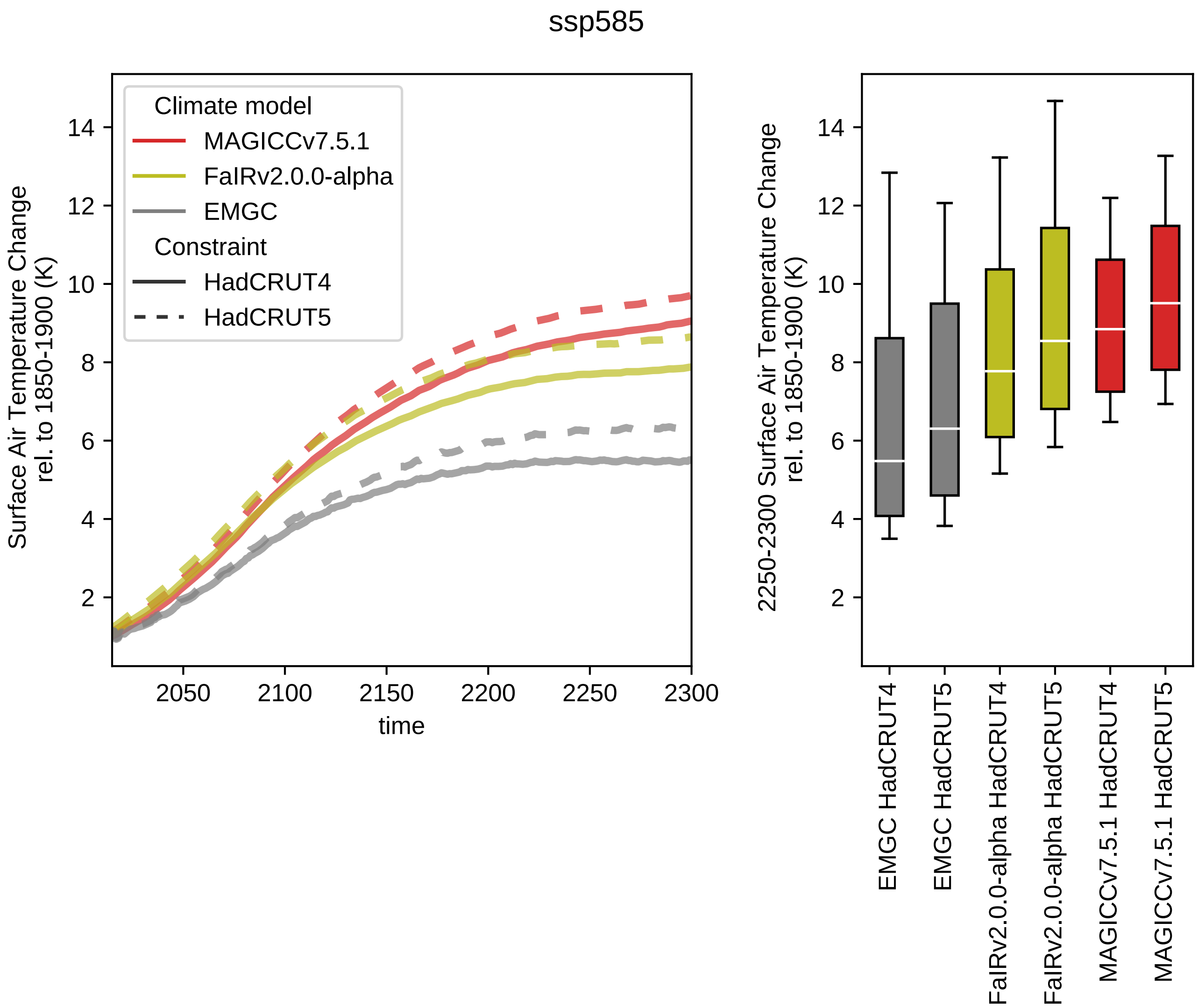

Figure S36. As in Figure S36, except for the concentration-driven SSP5-8.5 experiment.