Errors in simple climate model emulations of past and future global temperature change

L. S. Jackson[1], A. C. Maycock[1], T. Andrews[2], H.-B. Fredriksen[3], C. J. Smith[4], and P. M. Forster[5]

[1] School of Earth and Environment, University of Leeds, Leeds, UK.

[2] Met Office Hadley Centre, Exeter, UK.

[3] Department of Physics and Technology, UiT the Arctic University of Norway, Tromsø, Norway.

[4] International Institute for Applied Systems Analysis, Laxenburg, Austria.

[5] Priestley International Centre for Climate, University of Leeds, Leeds, UK.

Corresponding author: L. S. Jackson (l.s.jackson@leeds.ac.uk)

Key Points:

- Emulators of global surface temperature calibrated to individual climate models can generate large errors in past and future predictions.
- Emulation errors are not systematically related to emulator parameters and vary between climate models meaning they cannot be predicted.
- Rigorous out-of-sample evaluation is necessary to characterize emulator performance.

Abstract

Climate model emulators are widely used to generate temperature projections for climate scenarios, including in the recent IPCC Sixth Assessment Report. Here we evaluate the performance of a two-layer energy balance model in emulating historical and future temperature projections from CMIP6 models. We find that emulation errors can be large (>0.5°C for SSP2-4.5) and differ markedly between climate models, forcing scenarios and time periods. Errors arise in emulating the near-surface temperature response to both greenhouse gas and aerosol forcing; in some periods the errors due to these forcings oppose one another, giving the spurious impression of better emulator performance. Climate feedbacks are assumed constant in the emulator, introducing time-varying or state dependent feedbacks may reduce prediction errors. Close emulations can be produced for a given period but, crucially, this does not guarantee reliable emulations of other scenarios and periods. Therefore, rigorous out-of-sample evaluation is necessary to characterize emulator performance.

Plain Language Summary

Complex climate models are state-of-the-art tools used to produce projections of future climate but they are expensive and take a long time to run. Climate model emulators are simple statistical or physically based models that can aim to reproduce the response of complex climate models to a prescribed climate change scenario at lower cost and more quickly. In this study, we use a climate model emulator to reproduce simulations of twentieth and twenty-first century temperatures for eight complex climate models. We show that close emulations can be produced for pre-defined climate scenarios and time periods. Close emulations are not guaranteed, however, when the emulator is used for other climate scenarios or other periods. This is important because climate model emulators are frequently used to produce projections that are not available from complex climate models. Evaluation of climate model emulators and characterization of their uncertainties, therefore, should use data not used in the calibration of the emulator.

1 Introduction

Climate model emulators are simplified physical or statistical models that are computationally efficient. Climate model emulators played a central role in producing future global near-surface temperature projections for Working Group I (Forster et al., 2021; Lee et al., 2021) and Working Group III (Riahi et al., 2022) of the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC AR6). The IPCC AR6 used climate model emulators to supplement simulations from coupled atmosphere-ocean general circulation models (AOGCMs) extending available simulations further into the future and projecting future climate scenarios not available from AOGCMs. It is important, therefore, that the simplifying assumptions used by emulators are rigorously tested so the robustness of their performance is understood.

Physically based climate model emulators, such as energy balance models (EBMs), use bulk physical relationships to emulate the large-scale behavior of Earth's climate system. For example, EBMs were used by Colman and Soldatenko (2020) to investigate links between climate variability and climate sensitivity and, by Modak and Mauritsen (2021) to investigate the probability of occurrence of the 2000-2012 global warming hiatus.

Two-layer EBMs produce close emulations of idealized abrupt-4xCO2 and 1pctCO2 simulations from AOGCMs (e.g., "EBM-ε" in Geoffroy et al. 2013b; "held-two-layer-uom" in Nicholls et al., 2020). Differences between emulations and AOGCM projections are generally greatest at times of pronounced change in the rate of temperature increase. Such changes are associated with time-varying feedbacks (Senior and Mitchell, 2000; Winton et al., 2010; Armour et al., 2013; Dong et al., 2020; Dunne et al., 2020; Rugenstein et al., 2020; Dong et al., 2021) which are caused by evolving spatial pattern effects in surface temperature (Stevens et al., 2016; Andrews et al., 2015; Rugenstein et al., 2016; Dong et al., 2021) and non-linear state dependences in climate feedbacks (Good et al., 2015; Rohrschneider et al., 2019; Bloch-Johnson et al., 2021). EBMs have been enhanced to capture time-varying feedbacks: the Geoffroy et al. (2013b) EBM includes an efficacy parameter for deep ocean heat uptake and the "held-two-layer-uom" EBM also includes a state dependent feedback parameter (Rohrschneider et al., 2019; Nicholls et al., 2020). These paradigms, however, do not precisely capture the feedback changes in AOGCMs and contribute to structural error which is irreducible unless the EBM structure is enhanced (e.g., extending a two-layer EBM to three or more layers (Cummins et al., 2020)).

Assessments of emulator performance are more trustworthy when projections are validated using data different from those used to calibrate the emulator parameters (out-of-sample validation). EBM parameters are frequently calibrated using idealized step-forcing experiments (e.g., abrupt-4xCO2) with the parameters estimated using analytical methods (Geoffroy et al., 2013a) or statistical methods (e.g., Cummins et al., 2020). The Coupled Model Intercomparison Project Phase 6 (CMIP6) (Eyring et al., 2016) historical and future shared socio-economic pathway (SSP) projections for AOGCMs, therefore, are well suited for assessing EBM emulator performance. They can be used to produce out-of-sample assessments using realistic climate scenarios. Although climate model emulators have been evaluated (e.g., Nicholls et al., 2020; Nicholls et al., 2021), it is not known how well emulators perform for the latest CMIP6 AOGCMs using realistic, out-of-sample climate projections and latest assessments of effective radiative forcing (ERF). Furthermore, the contribution of irreducible structural errors to total prediction error remains poorly understood.

In this study, we evaluate the performance of a two-layer energy balance model (EBM2) (Held et al., 2010; Geoffroy et al., 2013a, b) for emulating CMIP6 historical and future temperature trends using different EBM calibrations. We calibrate the EBM2 parameters for specific periods and ERFs, and evaluate the temperature projections for subsequent periods and alternative ERF scenarios. EBM2 is compared against an step-response emulator and a three-layer EBM.

2 Methods and data

2.1 Step-response emulator

We use a step-response emulator (Good et al., 2011) to provide a comparator of EBM emulator performance for temperature projections. The step-response function for each AOGCM was derived by dividing the projected temperature changes from a single realization of a CMIP6 abrupt-4xCO2 simulation by the radiative forcing for 4xCO2 (Smith et al., 2020). The step-response function was smoothed using cubic splines, and linear regession (years 121-150) was used for extrapolation beyond the 150 years of the abrupt-4xCO2 simulations. Temperature projections from the step-response emulator were produced by convolution of annual changes in ERF and the step-response functions.

2.2 Two-layer EBM emulator (EBM2)

In EBM2 (Held et al., 2010; Geoffroy et al., 2013a) the upper layer represents the Earth's atmosphere, land surface and ocean mixed layer, and the lower layer represents the deep ocean. The rate of temperature change in each EBM2 layer is determined from:

$$C_1 \frac{dT_1}{dt} = F + \lambda T_1 - \varepsilon\gamma(T_1 - T_0) \quad (1)$$

$$C_0 \frac{dT_0}{dt} = \gamma(T_1 - T_0) \quad (2)$$

Where C represents heat capacity, T temperature, F ERF, λ the climate feedback parameter and γ the heat transfer coefficient between the upper layer (layer 1) and the lower layer (layer 0). We follow the formulation of Geoffroy et al. (2013b) which includes an efficacy parameter for deep ocean heat uptake (ε) to account for the forced pattern effect in surface temperature (Stevens et al., 2016). As is commonplace (Geoffroy et al., 2013a, b; Gregory et al., 2015; Cummins et al., 2020), the EBM2 parameters were calibrated for each AOGCM using a single realization of a CMIP6 abrupt-4xCO2 simulation. Radiative forcing for 4xCO2 was taken from Smith et al. (2020). See Tables S1 and S2 for further details.

2.3 Calibration of EBM2 using linear optimization

As an alternative to calibration using the abrupt-4xCO2 experiment, we use linear optimization (the L-BFGS-B algorithm in scipy.optimize.minimize v1.6.2) to optimize the λ and ε parameters by minimizing the root mean square error (RMSE) of the emulated temperatures compared to the AOGCM over a defined time period (e.g., historical) (Table S3, S4). Lower bounds of -0.5 W m^{-2} K^{-1} and 0.5 were imposed for λ and ε respectively, and upper bounds of -2.0 W m^{-2} K^{-1} and 2.0 respectively. These bounds are broadly based on the range of parameter values from the abrupt-4xCO2 calibration. The temperature projections are less sensitive to changes in the other EBM2 parameters (i.e., C_0, C_1, and γ), so these parameters are unchanged from their abrupt-4xCO2 calibrations. We also applied the linear optimization methodology to the abrupt-4xCO2 simulations, which produced very similar parameter values to the Geoffroy et al. (2013b) methodology used in the abrupt-4xCO2 calibration.

2.4 Three-layer EBM

We use a three-layer EBM (EBM3) (Cummins et al., 2020) as a second comparator for EBM2 performance. We follow the method of Cummins et al. (2020) to calibrate the EBM3 parameters (including ERF for 4xCO2) using a single realization of a CMIP6 abrupt-4xCO2 simulation.

2.5 Data

We use projections of global annual mean near-surface temperature and radiative fluxes at the top of atmosphere (TOA) from the CMIP6 archive. We emulate temperatures for eight AOGCMs selected because data was available for the CMIP6 experiments of interest. For projections of recent and future climate change, the Historical and SSP experiments were used. Projections of temperature change attributed to specific sources of ERF are taken from the Detection and Attribution Model Intercomparison Project (DAMIP) experiments (Gillett et al., 2016). The emulations are driven by time series of total annual ERF; estimates of ERF are taken from the Radiative Forcing Model Intercomparison Project (RFMIP) experiments (Pincus et al., 2016; Smith et al., 2021). The ERF for GFDL-CM4 was used for GFDL-ESM4 (RFMIP ERF being unavailable for GFDL-ESM4). Following Forster et al. (2013), unforced drift is removed from the AOGCM projections using the preindustrial control experiment.

3 Results

3.1 Historical period using the abrupt-4xCO2 calibration

EBM2 captures the increasing temperature trend during the twentieth century and distinguishes between high and low climate sensitivity AOGCMs (Figure 1). In all EBM2 emulations, a proportion of the RMSE (~ 0.07 K) arises from interannual variations in the AOGCM ensemble means that is not captured in the emulations (there are up to three members in each AOGCM historical ensemble). The performance of EBM2, however, varies substantially between AOGCMs. The emulation errors are not strongly correlated with parameter values though there is a weak correlation between smaller RMSEs and large relative deep ocean heat capacity (i.e., C_o/C_1) (Figure S1). The sensitivity of emulation errors to changes in λ and ε varies between AOGCMs (Figure S2). There are instances of both large and small RMSE emulations for both high and low climate sensitivity AOGCMs. For AOGCMs where there are substantial differences between the emulations and the AOGCM projections, the differences occur over different time periods. Differences are large for 1925-1950 (HadGEM3-GC31-LL), for 1950-1975 (NorESM2-LM) and for 2000-2015 (HadGEM3-GC31-LL, IPSL-CM6A-LR, and GFDL-ESM4). For IPSL-CM6A-LR, temperatures are overestimated by the emulators throughout 1915-2014. Close emulation of temperatures in abrupt-4xCO2 does not guarantee close emulation for the historical period (e.g. GFDL-ESM4, although using ERF from GFDL-CM4 likely introduces some emulation error for GFDL-ESM4). Similarly, a relatively poor emulation of abrupt-4xCO2 does not prohibit close emulation for the historical period (e.g. CNRM-CM6-1) (Figure S3). These results are important because they show that there is no a priori way to know if an AOGCM will be closely emulated.

The step-response emulator produces emulations with RMSEs broadly equivalent to or less than emulations from EBM2. The treatment of time-varying feedbacks in the step-response emulator (i.e., implicitly in the step-response function) differs from the treatment in EBM2 (i.e., based on ε) and may contribute to the good performance of the step-response emulator.

EBM3 performs better than EBM2 for abrupt-4xCO2, which is expected given the additional timescales resolved by the third layer which facilitates closer emulation of temperatures during years 10-40 of the abrupt-4xCO2 experiment, a period when the rate of temperature increase weakens rapidly (Figure S3). However, the improvement of EBM3 over EBM2 in the abrupt-4xCO2 experiment does not consistently translate to the historical experiment. Indeed there are three AOGCMs for which EBM2 has smaller RMSEs than EBM3 (HadGEM3-GC31-LL, MIROC6 and IPSL-CM6A-LR). EBM3, similar to EBM2, overestimates temperatures for 2000-2014 in three of the eight AOGCMs and produces larger RMSEs than the step-response emulator for some AOGCMs.

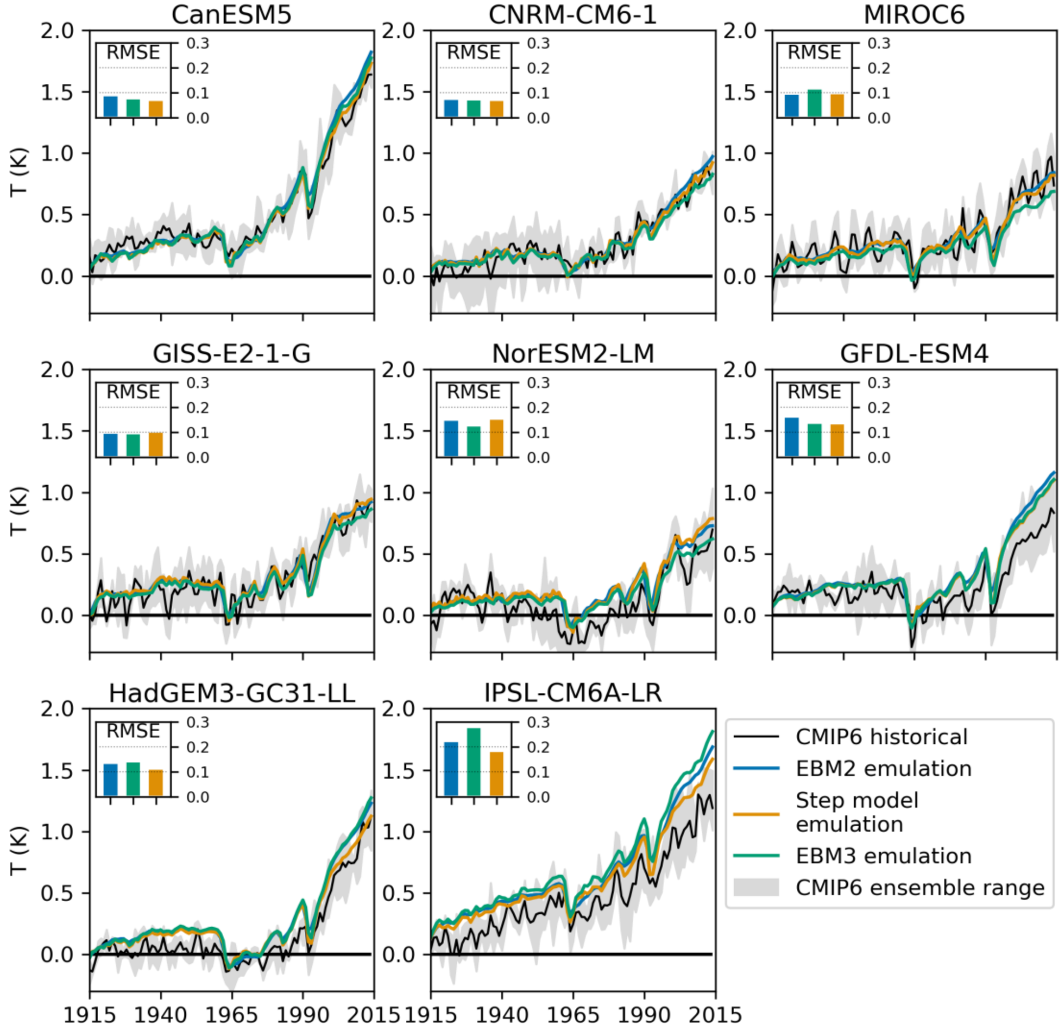

Figure 1. Global mean temperature anomalies from a 1850-1900 baseline for CMIP6 AOGCMs. Changes in temperatures are forced by historical forcings during 1850-2014 and are shown for the period 1915-2014. RMSEs are calculated over 1915-2014.

3.2 Roles of different forcings for near-surface temperature change

We used EBM2 to emulate the responses to historical greenhouse gas (hist-GHG), anthropogenic aerosol (hist-aer) and natural (hist-nat) forcings only. EBM2 was calibrated using abrupt-4xCO2 simulations and the AOGCM projections are from DAMIP (Gillett et al., 2016) (Figure 2). We focus on two AOGCMs with relatively large errors in their emulations for the historical period (HadGEM3-GC31-LL and IPSL-CM6A-LR), one AOGCM with relatively small errors (CanESM5), and one AOGCM whose responses to GHG and aerosol forcings contrast with the other AOGCMs (NorESM2-LM).

Although EBM2 was calibrated using abrupt-4xCO2, errors predominantly arise from the emulation of the response to GHG forcing; in part because GHGs have the largest ERF. The EBM2 emulations overestimate the temperature increase due to GHGs for HadGEM3-GC31-LL and IPSL-CM6A-LR.

Emulation of the temperature response to aerosol forcing is the largest source of error in one climate model (NorESM2-LM). For HadGEM3-GC31-LL and IPSL-CM6A-LR, errors associated with aerosol forcing offset errors associated with GHG forcing. This cancellation of errors gives a spurious impression of better performance for the historical simulations. As shown for the combined forcings (Figure 1), the step-response emulator produces closer emulations of temperature for GHG forcing. For anthropogenic aerosol forcing, the step-response emulator produces emulations of temperature very similar to EBM2.

Emulation of the temperature response to natural forcings is a small source of error and the emulations are mostly within the spread of each AOGCM ensemble (Figures 2 and S4). Although larger ensembles and longer simulations are required to robustly assess the emulated response to volcanic forcing, thermal inertia of the EBM2 layers and allowance for rapid cloud adjustments within RFMIP ERFs will likely have contributed to the close emulations for natural forcings (Held et al., 2010; Gregory et al., 2016).

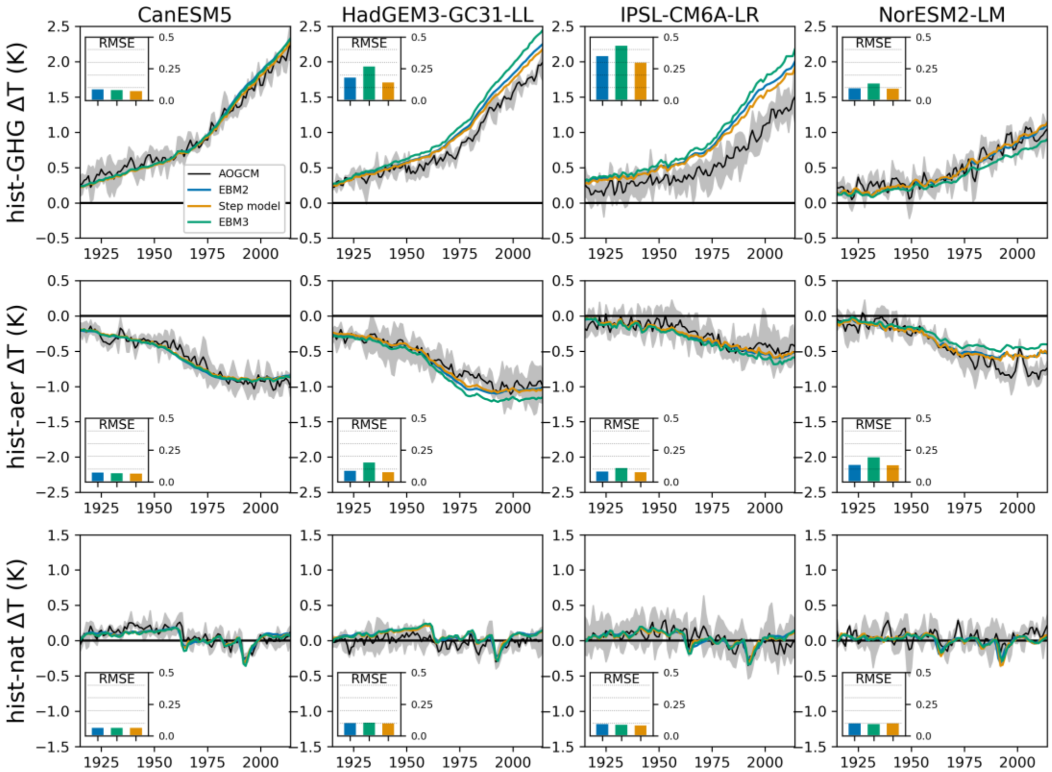

Figure 2. As Figure 1, except that temperature changes are forced by historical greenhouse gas (top row), anthropogenic aerosol (middle row), and natural (bottom row) forcings from RFMIP. The AOGCM projections are from DAMIP.

3.3 Alternative calibration of EBM2

To determine whether temperature emulations from EBM2 for the historical period can be improved by changes to the fitted parameters alone, we apply optimization (Section 2.3) to calibrate the λ and ε parameters (Figures 3 and S5).

This improves the emulations for all climate models. The greatest improvement occurs during 1980-2014 and the emulation of temperature during this period is improved further if the optimization is amended to minimize the RMSE specifically over this period. The spread in emulated temperatures about the 1:1 line is mainly driven by the small AOGCM ensemble sizes and is, therefore, similar for both EBM2 calibrations. Interannual variability is particularly large for NorESM2-LM and the emulated temperatures have a low correlation with the AOGCM temperatures for years prior to the 1980s when the climate response to forcing is relatively weak.

The emulations of the net radiation at the TOA (N) (Figure 3) show that close emulations of near-surface temperature can be produced despite poor emulations of N. There is a large spread in the emulations of N about the 1:1 line for all climate models. The emulation of N during the late twentieth/early twenty-first century is poor for HadGEM3-GC31-LL and emulated N has a weak correlation with its AOGCM for NorESM2-LM. Optimization does not improve the emulation of N. There are small changes in emulated N for CanESM5 and NorESM2-LM. The improved temperature emulations from the optimization method for HadGEM3-GC31-LL come at the expense of poorer emulations of N. This result is important because it demonstrates that climate model emulators can produce reasonable simulations of near-surface temperature change, but the evolution of ocean heat uptake and TOA energy imbalance is incorrect demonstrating limitations to physical interpretation.

We also used optimization to calibrate the λ and ε parameters separately for GHG and aerosol forcing using the DAMIP experiments. The calibrated parameter values differ for the two types of forcing (Table S3) and also vary when RMSE is minimized over different periods.

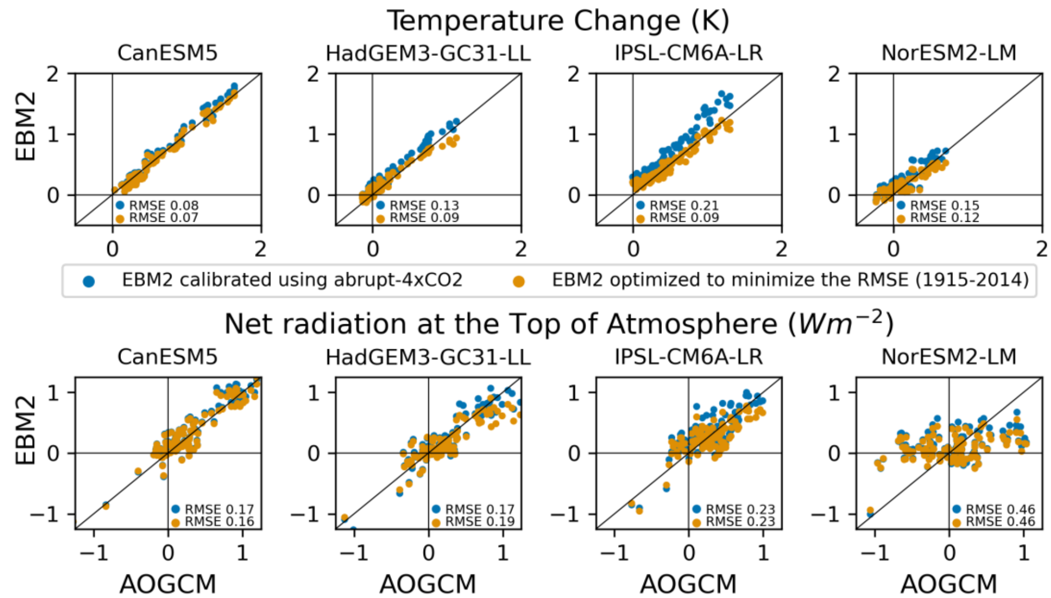

Figure 3. Projected changes in global mean temperature (top row) and net radiation at the TOA (*N*) (bottom row). Each panel shows changes in the AOGCM (x-axis) against the EBM2 emulation (y-axis). Each point represents an annual mean during 1915-2014.

3.4 Future near-surface temperature projections

We compare temperature emulations for the twenty-first century from EBM2 based on different methods for calibrating λ and ε (Figure 4). Results are shown for the AOGCMs where the most complete CMIP6 data are available. Results for other experiments are shown in Figure S6 and Table S1 describes the calibrations.

The performance of the abrupt-4xCO2 calibration varies greatly between the AOGCMs (Figures 4a, b) and typically performs worse than the step-response emulator. The emulations of SSP2-4.5 deteriorate during the twenty-first century. The errors in the emulations are correlated with the magnitude of the forcing and peak near the end of the twenty-first century for total and GHG forcing and early in the twenty-first century for aerosol forcing.

Calibration by optimization of the λ and ε parameters over 1850-2100 (Figures 4c, d) yielded close emulations for all of the AOGCMs and across all experiments. Similarly close emulations were also achieved by minimizing the RMSE over 2015-2100 (not shown). Minimizing the RMSE for the later years of the projection, when the temperature anomalies are largest, is key.

Emulations of temperature to 2100 based on optimizing the λ and ε parameters using the 1850-2014 period yields close emulations of temperature to 2014 but errors increase after the calibration period (Figures 4e, f). Extending the calibration period from 1850-2014 to 1850-2040 (not shown) improves the emulation to 2040 but not always after 2040. Importantly, it does not mitigate the risk of large emulation errors outside the calibration period and its impact varies greatly between AOGCMs and between different experiments for the same AOGCM.

To investigate the impact of using a calibration from one experiment for a different experiment, the "Hist-SSP245_1850-2100" calibration (which uses SSP2-4.5 all forcings) was applied to the GHG only (SSP2-4.5-GHG) and the anthropogenic aerosol only (SSP2-4.5-AER) experiments from DAMIP (Figures 4g, h). For both SSP2-4.5-GHG and SSP2-4.5-AER, the error for the "Hist-SSP245_1850-2100" calibration is greater than for the Hist-SSP245-GHG_1850-2100 and Hist-SSP245-AER_1850-2100 calibrations respectively. The impact also varies between climate models and experiments in terms of the size of the impact and its temporal behaviour. Similar results were also found for the high mitigation scenario SSP1-1.9 (Figure S7). Bespoke parameter calibrations for different ERF scenarios are necessary, therefore, to achieve close emulations throughout 1850-2100. This result is important because it demonstrates that emulator performance can be poor for out-of-sample predictions, yet there is no clear a priori way to know if this will be the case. This poses a problem since some of the value of emulators lies in their use for creating out-of-sample scenarios where AOGCM simulations do not exist and cannot be readily performed.

The average of the emulations for individual climate models (Figure 4 "Ensemble mean") has relatively small RMSEs (except for the SSP2-4.5 1850-2014 calibration in Figure 4e). This is due, in part, to averaging of interannual variability across the ensemble of emulations. Further, the ensemble mean generally has smaller RMSEs than an emulation in which the ensemble mean ERF is used to emulate the ensemble temperature projection (Figure 4 "Ensemble emulation").

Finally, while the optimization method yields unique parameter solutions there is a near linear trade-off between the λ and ε parameters when minimizing the RMSE (demonstrated for historical/SSP2-4.5 in Figure S2 and for the first 150 years of abrupt-4xCO2 in Figure S8). In EBM2, changes in the climate feedback parameter (λ) are compensated for by changes in the efficacy of deep ocean heat uptake (ε) and the transient temperature response is largely unchanged. This shows that optimized values for the λ and ε parameters may not be robust estimates of climate feedback or the AOGCM pattern effect and the physical interpretation of parameter value changes when optimizing the calibration is hindered by the linear trade-off between the λ and ε parameters.

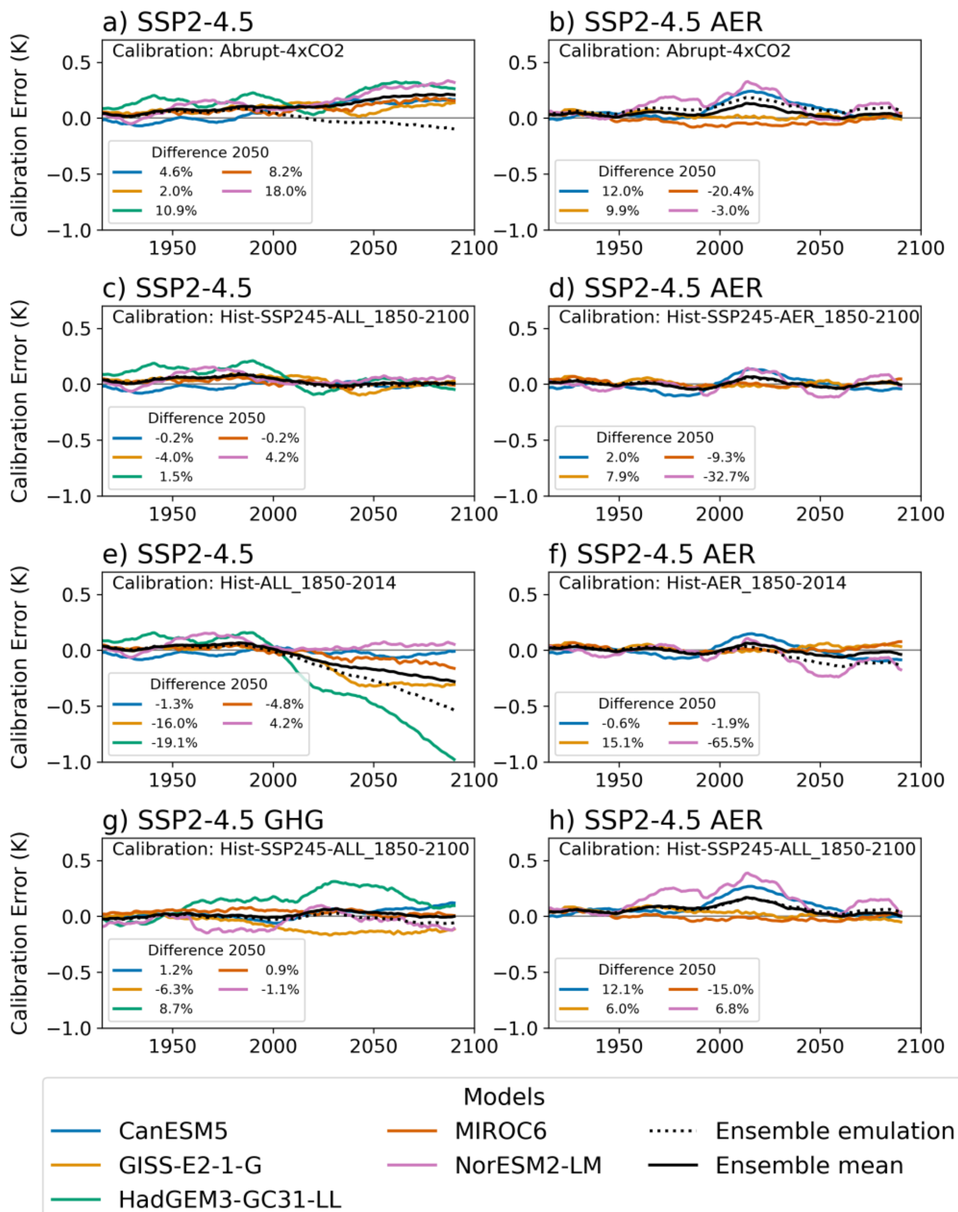

Figure 4. Differences between EBM2 emulations and AOGCM temperature projections. Rows show results for four calibrations of EBM2. Row B uses λ and ε parameter values which minimize the RMSE for temperatures during 1850-2100. Row C uses parameter values which minimize the RMSE during 1850-2014. Row D shows EBM2 calibrated to minimize the RMSE

during 1850-2100 for SSP2-4.5 and this calibration is used to emulate SSP2-4.5-GHG and SSP2-4.5-AER. Annual means are smoothed using a 21-year moving average.

4 Discussion and conclusions

Our results show prediction errors in EBM2 for future global temperature projections vary greatly between AOGCMs, forcings, time periods and methods of emulator calibration.

The EBM2 calibration using the abrupt-4xCO2 experiment does not produce reliable projections of historical warming for several AOGCMs. Although calibration of the λ and ε parameters using optimization substantially reduces emulation errors for periods where an AOGCM simulation is available, optimization of these parameters does not guarantee reliable out-of-sample projections. Without an AOGCM projection for a given AOGCM and scenario, it is not knowable if the EBM2 future projection will be reliable.

Close emulation of the historical period is not sufficient to guarantee reliable emulation of future temperature changes (Figure 4; Nicholls et al., 2021). Opposing errors in the emulation of GHG and aerosol forcings give a misleading impression of emulator performance. Many climate model emulators do not reliably emulate AOGCM projections for high emissions scenarios (Nicholls et al., 2021); our results suggest that strong mitigation scenarios may not be reliably emulated.

How could the EBM2 emulator be changed to improve the out-of-sample emulations? First, an efficacy factor could be introduced to account for asymmetry in AOGCM responses to GHG and aerosol forcings. Second, EBM2 could be developed to incorporate variations in climate feedbacks and the evolution of AOGCM pattern effects. Late twentieth-century warming has been suppressed by changes in the observed sea surface temperature (SST) patterns and associated cloud feedbacks (Andrews et al., 2018; Dong et al., 2021; Fueglistaler and Silvers, 2021). Future warming could be affected by changes in the pattern effect (Zhou et al., 2021). Climate model simulations show that climate feedbacks weaken through time in response to step-forcings and changes in feedbacks are associated with changes in SST patterns (e.g., Dong et al., 2020; Dunne et al., 2020). Incorporating time-varying feedbacks in EBM2, however, requires further research to distinguish forced changes in feedbacks from unforced climate noise and to explicitly link global feedback changes to variations in SST patterns (e.g., using SST anomalies for regions of tropical deep convection (Fueglistaler and Silvers (2021)).

EBM2 out-of-sample emulations could potentially be improved without changes to the emulator. First, when available, larger AOGCM ensembles could be used to reduce the contribution to emulation errors from chance. Second, more physically plausible parameter tunings could be achieved by using optimization to jointly minimize RMSEs for temperature and ocean heat flux (Dorheim et al., 2020). Our initial investigations minimizing RMSE for temperature and *N*, however, showed that the emulation of historical temperatures was substantially worse than when minimizing RMSE for temperature alone.

Emulations could also be improved through advances in the separation of forcing and climate feedbacks in AOGCMs. We used the latest estimates of ERF derived from fixed-SST simulations but substantial uncertainty in ERF remains (Forster et al., 2016; Dong et al., 2021). Correcting for land warming in abrupt-4xCO2 fixed-SST experiments increases the ERF (Andrews et al., 2021) and leads to a weaker temperature response per unit forcing in EBM2. If the abrupt-4xCO2 ERF without corrections happens to be more underestimated than the historical ERF, the historical EBM2 responses will be overestimated. Forcing estimates remain dependent on the method used (Forster et al., 2013; Sherwood et al., 2015; Larson and Portmann, 2016; Fredriksen et al., 2021).

Our findings are relevant to observationally contrained climate model emulators aiming to simulate real-world changes (e.g., Forster et al., 2021). Emulator structural errors and uncertainties in inputs (e.g., ERF) are as relevant to real-world emulations as to emulations of AOGCMs. Indeed, there are additional challenges. There is only one realization of past climate and future climate is unknown. Observational large ensembles (McKinnon et al., 2017) could be used to help characterize uncertainty in emulating past climate.

Acknowledgments

LSJ, ACM, TA and PMF were supported by the European Union's Horizon 2020 programme under grant agreement No 820829 (CONSTRAIN). TA was supported by the Met Office Hadley Centre Climate Programme funded by BEIS. CJS was supported by a joint NERC-IIASA Collaborative Research Fellowship (NE/T009381/1). ACM was supported by The Leverhulme Trust (PLP-2018-278). We acknowledge: the World Climate Research Programme and its Working Group on Coupled Modeling for coordinating and promoting CMIP6; the climate modeling groups for producing their model output; the Earth System Grid Federation (ESGF) for archiving the data and providing access; and the funding agencies who support CMIP6 and ESGF. We thank Nicholas Leach and an anonymous reviewer for their useful review comments.

Open Research

The CMIP6 data were downloaded from the publicly available Earth System Grid Federation archive at https://esgf-node.llnl.gov/projects/cmip6/. The R package for the three-layer model (Cummins et al. 2020) was downloaded on 29 July 2021 from https://github.com/donaldcummins/EBM and is available from https://doi.org/10.5281/zenodo.5217603. Processed data produced for this paper are available on Zenodo at https://doi.org/10.5281/zenodo.6646804.

References

Andrews, T., Gregory, J. M., & Webb, M. J. (2015). The dependence of radiative forcing and feedback on evolving patterns of surface temperature change in climate models. *Journal of Climate*, *28*(4), 1630–1648. https://doi.org/10.1175/JCLI-D-14-00545.1.

Andrews, T., Gregory, J. M., Paynter, D., Silvers, L. G., Zhou, C., Mauritsen, T., Webb, M. J., Armour, K. C., Forster, P. M., & Titchner, H. (2018). Accounting for changing temperature patterns increases historical estimates of climate sensitivity. Geophysical Research Letters, 45, 8490–8499. https://doi.org/10.1029/2018GL078887.

Andrews, T., Smith, C. J., Myhre, G., Forster, P. M., Chadwick, R., & Ackerley, D. (2021). Effective radiative forcing in a GCM with fixed surface temperatures. Journal of Geophysical Research: Atmospheres, 126, e2020JD033880. https://doi.org/10.1029/2020JD033880.

Armour, K. C., Bitz, C. M., & Roe, G. H. (2013). Time-Varying Climate Sensitivity from Regional Feedbacks. *Journal of Climate*, *26*(13), 4518–4534. https://doi.org/10.1175/JCLI-D-12-00544.1.

Bloch-Johnson, J., Rugenstein, M., Stolpe, M. B., Rohrschneider, T., Zheng, Y., & Gregory, J. M. (2021). Climate Sensitivity Increases Under Higher CO2 Levels Due to Feedback Temperature Dependence. In *Geophysical Research Letters* (Vol. 48, Issue 4). Blackwell Publishing Ltd. https://doi.org/10.1029/2020GL089074.

Colman, R., & Soldatenko, S. (2020). Understanding the links between climate feedbacks, variability and change using a two-layer energy balance model. Climate Dynamics, 54, 3441–3459, https://doi.org/10.1007/s00382-020-05189-3.

Cummins, D. P., Stephenson, D. B., and Stott, P. A. (2020). Optimal Estimation of Stochastic Energy Balance Model Parameters. Journal of Climate, 33, 7909-7926, DOI: 10.1175/JCLI-D-19-0589.1

Dong, Y., Armour, K. C., Zelinka, M. D., Proistosescu, C., Battisti, D. S., Zhou, C., & Andrews, T. (2020). Intermodel spread in the pattern effect and its contribution to climate sensitivity in CMIP5 and CMIP6 models. *Journal of Climate*, *33*(18), 7755–7775. https://doi.org/10.1175/JCLI-D-19-1011.1.

Dong, Y., Armour, K. C., Proistosescu, C., Andrews, T., Battisti, D. S., Forster, P. M., Paynter, D., Smith, C. J., & Shiogama, H. (2021). Biased estimates of Equilibrium Climate Sensitivity and Transient Climate Response derived from historical CMIP6 simulations. *Geophysical Research Letters*. https://doi.org/10.1029/2021GL095778.

Dorheim, K., Link, R., Hartin, C., Kravitz, B., & Snyder, A. (2020). Calibrating Simple Climate Models to Individual Earth System Models: Lessons Learned From Calibrating Hector. *Earth and Space Science*, *7*(11). https://doi.org/10.1029/2019EA000980.

Dunne, J. P., Winton, M., Bacmeister, J., Danabasoglu, G., Gettelman, A., Golaz, J. C., Hannay, C., Schmidt, G. A., Krasting, J. P., Leung, L. R., Nazarenko, L., Sentman, L. T., Stouffer, R. J., & Wolfe, J. D. (2020). Comparison of Equilibrium Climate Sensitivity Estimates From Slab Ocean, 150-Year, and Longer Simulations. *Geophysical Research Letters*, *47*(16). https://doi.org/10.1029/2020GL088852.

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. *Geosci. Model Dev.*, 9, 1937–1958, doi:10.5194/gmd-9-1937-2016.

Forster, P. M., Andrews, T., Good, P., Gregory, J. M., Jackson, L. S., & Zelinka, M. (2013). Evaluating adjusted forcing and model spread for historical and future scenarios in the CMIP5 generation of climate models. *J. Geophys. Res. Atmos.*, 118, 1139–1150. https://doi.org/10.1002/jgrd.50174.

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

Forster, P., T. Storelvmo, K. Armour, W. Collins, J. L. Dufresne, D. Frame, D. J. Lunt, T. Mauritsen, M. D. Palmer, M. Watanabe, M. Wild, & H. Zhang (2021). The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity. In: *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* [Masson-Delmotte, V., P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K.Leitzell, E. Lonnoy, J.B.R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu and B. Zhou (eds.)]. Cambridge University Press. In Press.

Fredriksen, H., Rugenstein, M., & Graversen, R. (2021). Estimating Radiative Forcing With a Nonconstant Feedback Parameter and Linear Response. *Journal of Geophysical Research: Atmospheres*, *126*(24). https://doi.org/10.1029/2020jd034145.

Fueglistaler, S., & Silvers, L. G. (2021). The Peculiar Trajectory of Global Warming. *Journal of Geophysical Research: Atmospheres*, *126*(4). https://doi.org/10.1029/2020JD033629.

Geoffroy, O., Saint-Martin, D., Olivié, D. J. L., Voldoire, A., Bellon, G., & S. Tytéca, S. (2013a). 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, 1841-1857. doi: 10.1175/JCLI-D-12-00195.1.

Geoffroy, O., Saint-martin, D., Bellon, G., & Voldoire, A. (2013b). 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, 1859-1876. doi: 10.1175/JCLI-D-12-00196.1.

Gillett, N. P., Shiogama, H., Funke, B., Hegerl, G., Knutti, R., Matthes, K., Santer, B. D., Stone, D., & Tebaldi, C. (2016). The Detection and Attribution Model Intercomparison Project (DAMIP v1.0) contribution to CMIP6. Geosci. Model Dev., 9, 3685–3697. doi:10.5194/gmd-9-3685-2016.

Good, P., Gregory, J. M., & Lowe, J. A. (2011). A step-response simple climate model to reconstruct and interpret AOGCM projections. *Geophysical Research Letters*, 38, L01703. doi:10.1029/2010GL045208.

Good, P., Lowe, J. A., Andrews, T., Wiltshire, A., Chadwick, R., Ridley, J. K., Menary, M. B., Bouttes, N., Dufresne, J. L., Gregory, J. M., Schaller, N., & Shiogama, H. (2015). Nonlinear regional warming with increasing CO2 concentrations. Nature Climate Change, 5(2), 138–142. doi.org/10.1038/nclimate2498.

Gregory, J. M., Andrews, T., & Good, P. (2015). The inconstancy of the transient climate response parameter under increasing CO2. Phil. Trans. R. Soc. A 373: 20140417. http://dx.doi.org/10.1098/rsta.2014.0417.

Gregory, J. M., Andrews, T., Good, P., Mauritsen, T., & Forster, P. M. (2016). Small global-mean cooling due to volcanic radiative forcing. Clim. Dyn., 47, 3979–3991. DOI 10.1007/s00382-016-3055-1.

Held, I. M., Winton, M., Takahashi, K., Delworth, T., Zeng, F., & Vallis, G. K. (2010). Probing the Fast and Slow Components of Global Warming by Returning Abruptly to Preindustrial Forcing. Journal of Climate, 23, 2418-2427. Doi: 10.1175/2009JCLI3466.1.

Larson, E. J. L., & Portmann, R. W. (2016). A Temporal Kernel Method to Compute Effective Radiative Forcing in CMIP5 Transient Simulations. Journal of Climate, 29, 1497–1509. https://doi.org/10.1175/JCLI-D-15-0577.1.

Lee, J. Y., J. Marotzke, G. Bala, L. Cao, S. Corti, J. P. Dunne, F. Engelbrecht, E. Fischer, J. C. Fyfe, C. Jones, A. Maycock, J. Mutemi, O. Ndiaye, S. Panickal, & T. Zhou (2021). Future Global Climate: Scenario-Based Projections and Near-Term Information. In: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Masson-Delmotte, V., P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O.Yelekçi, R. Yu and B. Zhou (eds.)]. Cambridge University Press. In Press.

McKinnon, K. A., Poppick, A., Dunn-Sigouin, E., & Deser, C. (2017). An "Observational Large Ensemble" to Compare Observed and Modeled Temperature Trend Uncertainty due to Internal Variability. Journal of Climate, 30, 7585–7598. https://doi.org/10.1175/JCLI-D-16-0905.1.

Modak, A., & Mauritsen, T. (2021). The 2000–2012 global warming hiatus more likely with a low climate sensitivity. Geophysical Research Letters, 48, e2020GL091779. https://doi.org/10.1029/2020GL091779.

Nicholls, Z. R. J., Meinshausen, M., Lewis, J., Gieseke, R., Dommenget, D., Dorheim, K., Fan5, C. S., Fuglestvedt, J. S., Gasser, T., Goluke, U., Goodwin, P., Hartin, C., P. Hope, A., Kriegler, E., J. Leach, N., Marchegiani, D., A. McBride, L., Quilcaille, Y., Rogelj, J., & 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. https://doi.org/10.5194/gmd-13-5175-2020.

Nicholls, Z., Meinshausen, M., Lewis, J., Corradi, M. R., Dorheim, K., Gasser, T., Gieseke, R., Hope, A. P., Leach, N. J., McBride, L. A., Quilcaille, Y., Rogelj, J., Salawitch, R. J., Samset, B. H., Sandstad, M., Shiklomanov, A., Skeie, R. B., Smith, C. J., Smith, S. J., Su, X., Tsutsui, J., Vega-Westhoff, B., & Woodard, D. L. (2021). Reduced complexity Model Intercomparison Project Phase 2: Synthesizing Earth system knowledge for probabilistic climate projections. Earth's Future, 9, e2020EF001900. https://doi.org/10.1029/2020EF001900.

Pincus, R., Forster, P. M., & Stevens, B. (2016), The Radiative Forcing Model Intercomparison Project (RFMIP): experimental protocol for CMIP6. Geosci. Model Dev., 9, 3447–3460. doi:10.5194/gmd-9-3447-2016.

Riahi, K., Schaeffer, R., Arango, J., Calvin, K., Guivarch, C., Hasegawa, T., et al. (2022): Mitigation pathways compatible with long-term goals. In IPCC, 2022*: Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* [P.R. Shukla, J. Skea, R. Slade, A. Al Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, J. Malley, (eds.)]. Cambridge University Press, Cambridge, UK and New York, NY, USA. doi: 10.1017/9781009157926.005

Rohrschneider, T., Stevens, B., & Mauritsen, T. (2019). On simple representations of the climate response to external radiative forcing. *Climate Dynamics*, *53*(5–6), 3131–3145. https://doi.org/10.1007/s00382-019-04686-4.

Rugenstein, M. A. A., Caldeira, K., & Knutti, R. (2016). Dependence of global radiative feedbacks on evolving patterns of surface heat fluxes. *Geophysical Research Letters*, *43*(18), 9877–9885. https://doi.org/10.1002/2016GL070907.

Rugenstein, M., Bloch-Johnson, J., Gregory, J., Andrews, T., Mauritsen, T., Li, C., Frölicher, T. L., Paynter, D., Danabasoglu, G., Yang, S., Dufresne, J. L., Cao, L., Schmidt, G. A., Abe-Ouchi, A., Geoffroy, O., & Knutti, R. (2020). Equilibrium Climate Sensitivity Estimated by Equilibrating Climate Models. *Geophysical Research Letters*, *47*(4).https://doi.org/10.1029/2019GL083898.

Senior, C. A., & Mitchell, J. F. B. (2000). The time-dependence of climate sensitivity. *Geophysical Research Letters*, *27*(17), 2685–2688. https://doi.org/10.1029/2000GL011373.

Sherwood, S. C., Bony, S., Boucher, O., Bretherton, C., Forster, P. M., Gregory, J. M., & Stevens, B. (2015). Adjustments in the forcing-feedback framework for understanding climate change. *Bulletin of the American Meteorological Society*, *96*(2), 217–228. https://doi.org/10.1175/BAMS-D-13-00167.1.

Smith, C. J., Kramer, R. J., Myhre, G., Alterskjær, K., Collins, W., Sima, A., et al. (2020). Effective radiative forcing and adjustments in CMIP6 models. Atmos. Chem. Phys., 20, 9591–9618, https://doi.org/10.5194/acp-20-9591-2020.

Smith, C. J., Harris, G. R., Palmer, M. D., Bellouin, N., Collins, W., Myhre, G., Schulz, M., Golaz, J.-C., Ringer, M., Storelvmo, T., & Forster, P. M. (2021). Energy Budget Constraints on the Time History of Aerosol Forcing and Climate Sensitivity. Journal of Geophysical Research: Atmospheres, 126, e2020JD033622. https://doi.org/10.1029/2020JD033622.

Stevens, B., Sherwood, S. C., Bony, S., & Webb, M. J. (2016). Prospects for narrowing bounds on Earth's equilibrium climate sensitivity, *Earth's Future*, 4, 512–522. doi:10.1002/2016EF000376.

Winton, M., Takahashi, K., & Held, I. M. (2010). Importance of Ocean Heat Uptake Efficacy to Transient Climate Change. Journal of Climate, 23, 2333-2344, DOI: 10.1175/2009JCLI3139.1.

Zhou, C., Zelinka, M. D., Dessler, A. E., & Wang, M. (2021). Greater committed warming after accounting for the pattern effect. *Nature Climate Change*, *11*(2), 132–136. https://doi.org/10.1038/s41558-020-00955-x.