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

Lawrence Stephen Jackson[1], Amanda Maycock[1], Timothy Andrews[2], Christopher J Smith[1], and Piers Forster[1]

[1]University of Leeds
[2]Met Office

November 22, 2022

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 prediction errors can be large (greater than 0.5°C in a given year) 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. Time-varying and state-dependent feedbacks may contribute to 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.

L. S. Jackson[1], A. C. Maycock[1], T. Andrews[2], C. J. Smith[1], and P. M. Forster[3]

[1] School of Earth and Environment, University of Leeds, Leeds, UK. [2] Met Office Hadley Centre, Exeter, UK. [3] 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 model parameters 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 prediction errors can be large (greater than 0.5°C in a given year) 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. Time-varying and state-dependent feedbacks may contribute to 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 which can reproduce projections from complex climate models at lower cost and more quickly. In this study, we use a climate model emulator to reproduce projections of twentieth and twenty-first century temperatures for eight complex climate models. We show that close emulations can be produced for predefined 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 the Working Group I Sixth Assessment Report (Forster et al. 2021; Lee et al. 2021) 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 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 model 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 model 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 (Eyring et al. 2016) AOGCMs using realistic, out-of-sample climate projections and latest assessments of effective radiative forcing (ERF). Furthermore, the contribution of irreducible model 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 benchmarked against an impulse-response step model and a three-layer EBM.

2 Methods and data

2.1 Impulse-response step model

We use a step model (Good et al. 2011) to provide a benchmark 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 (Byrne & Goldblatt 2013). 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 model were produced by convolution of annual changes in ERF and the step-response functions.

2.2 Two-layer EBM

In the two-layer EBM (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 model layer is determined from:

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

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

Where C representations 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 (Table S1) were calibrated for each AOGCM using a single realization of a CMIP6 abrupt-4xCO2 simulation (Table S1).

2.3 Calibration of EBM2 using linear optimization

As an alternative to abrupt-4xCO2 calibration, we use a linear optimization algorithm (scipy.optimize.minimize v1.6.2) to optimize the λ and ε parameters by minimizing the root mean squared error (RMSE) of the emulated temperatures compared to the AOGCM. 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 and affirmed the calibrated parameter values of Geoffroy et al. (2013b).

2.4 Three-layer EBM

We use a three-layer EBM (EBM3) (Cummins et al. 2020) as a second benchmark for EBM2 performance. We follow the method of Cummins et al. (2020) to calibrate EBM3 parameters for each AOGCM 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. The Detection and Attribution Model Intercomparison Project (DAMIP) experiments (Gillett et al. 2016) are used for projections of temperature change attributed to different sources of ERF. RFMIP experiments (Pincus et al. 2016; Smith et al. 2021) are used for estimates of ERF during the historical period and ERF projected to 2100 under SSP2-4.5. Following Forster et al. (2013), unforced drift is removed from the AOGCM projections using the preindustrial control simulation.

3 Results

3.1 Historical period

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 three members in each AOGCM historical ensemble). The performance of EBM2, however, varies substantially between AOGCMs. 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, GFDL-ESM4 and NorESM2-LM). For IPSL-CM6A-LR, temperatures are overestimated by the emulators throughout 1915-2014. Intriguingly, close emulation of temperatures in abrupt-4xCO2 does not guarantee close emulation for the historical period (e.g. GFDL-ESM4), and a relatively poor emulation of abrupt-4xCO2 does not prohibit close emulation for the historical period (e.g. CNRM-CM6-1) (Figure S1).

The step model produces emulations with RMSEs equivalent to or less than emulations from EBM2 in seven of the eight AOGCMs. The exception is NorESM2-LM which has relatively large inter-annual variability and is the only model to show an apparent cooling trend during years 20-50 of its abrupt-4xCO2 simulation (Figure S1).

EBM3 performs better than EBM2 for abrupt-4xCO2, which is expected given the additional timescales resolved by the third layer. The additional degrees of freedom enable a much closer emulation of temperatures during years 10-40 of the abrupt-4xCO2 experiment, a period when the rate of temperature increase weakens rapidly (Figure S1). However, the improvement of EBM3 over EBM2 in the abrupt-4xCO2 experiment does not consistently translate to the historical experiment. Indeed there are two AOGCMs for which EBM2 has smaller RMSEs than EBM3 (MIROC6 and IPSL-CM6A-LR). Both EBMs overestimate temperatures for 1990-2014 in four of the eight AOGCMs and generally produce larger RMSEs than the step model.

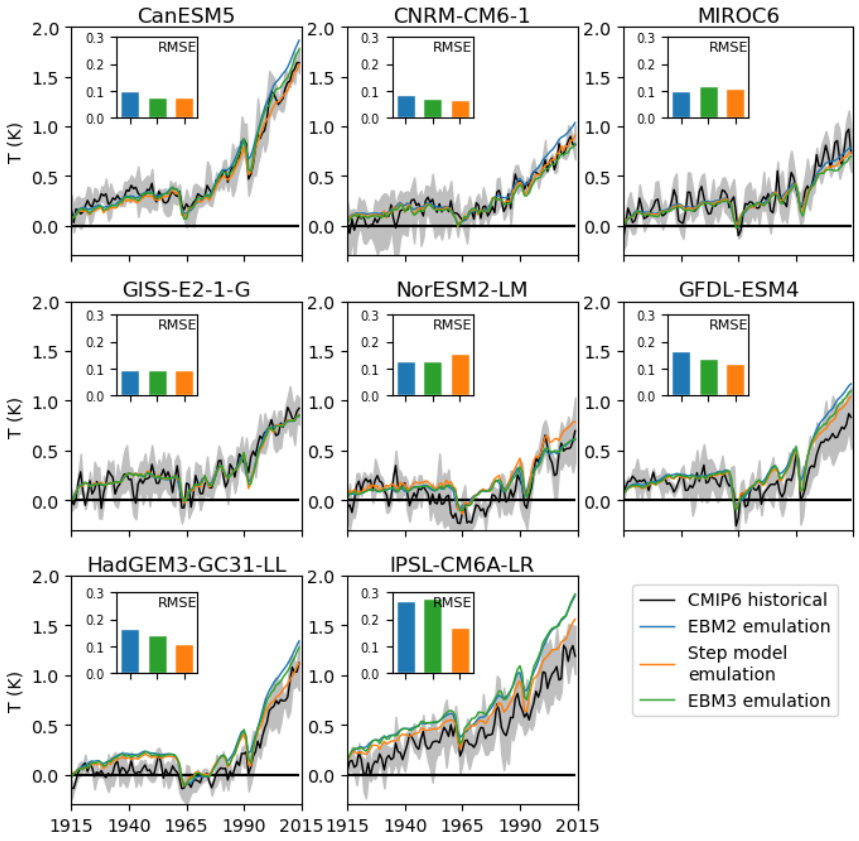

Figure 1. Global mean temperature anomalies from a 1850-1900 baseline for CMIP6 AOGCMs. The range between the ensemble maximum and minimum temperature changes is shown by gray shading. Changes in temperatures are forced by historical forcings during 1850-2014 and are shown for the period 1915-2014. RMSEs are calculated over 1915-2014.

3.2 Roles of different forcings for near-surface temperature change

In 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 contrast with the other AOGCMs (NorESM2-LM).

Although EBM2 was calibrated using abrupt-4xCO2, errors predominantly arise from emulation of the response to GHG forcing; in part because GHG has the largest ERF. The EBM2 emulations overestimate the temperature increase due to GHGs for HadGEM3-GC31-LL, IPSL-CM6A-LR and CanESM5 (even though the CanESM5 historical fit is good). In contrast, the EBM2 emulation underestimates the temperature response to GHGs for NorESM2-LM.

Emulation of the temperature response to aerosol forcing is the largest source of error in one model (NorESM2-LM). For all models, 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 model produces closer emulations of temperature for both GHG and aerosol forcings.

Emulation of the temperature response to natural forcings is a small source of error for the eight AOGCMs and the emulations are mostly within the spread of the AOGCM ensemble (Figures 2 and S2). 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 contribute to closer emulations (Held et al. 2010; Gregory et al. 2016).

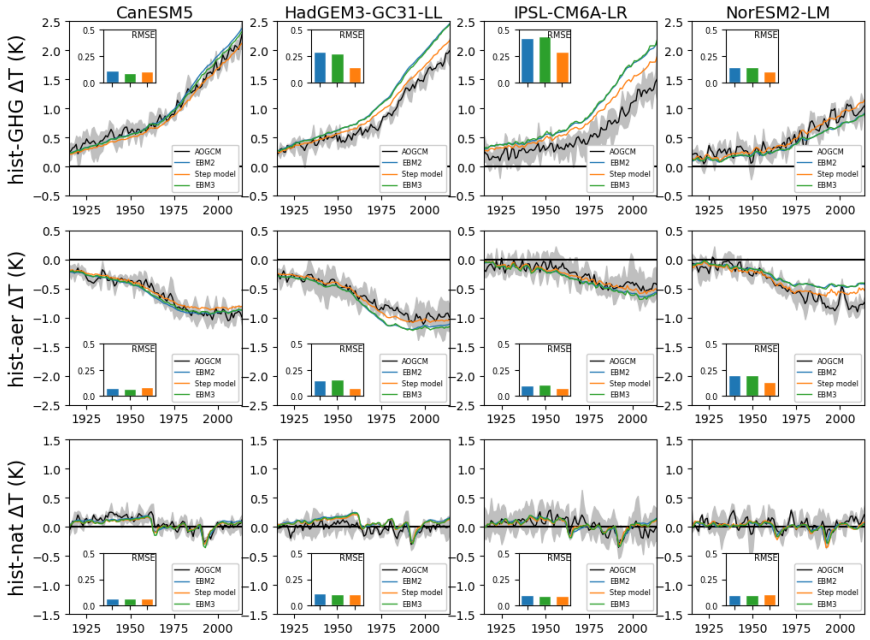

Figure 2. As Figure 1, except that temperature changes are forced by historical greenhouse gas (top row), anthropogenic aerosol (middle row), and natural (bottom row) forcings from RFMIP.

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 (Figure 3, Tables S2 and S3).

This improves the emulations for all 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 radiative flux 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 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 and IPSL-CM6-LR 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 constrained the λ and ε parameters separately for GHG and aerosol forcing using the DAMIP experiments. The constrained parameter values differ for the two types of forcing (Tables S2 and S3). Constrained parameter values also vary when RMSE is minimized over different periods of time.

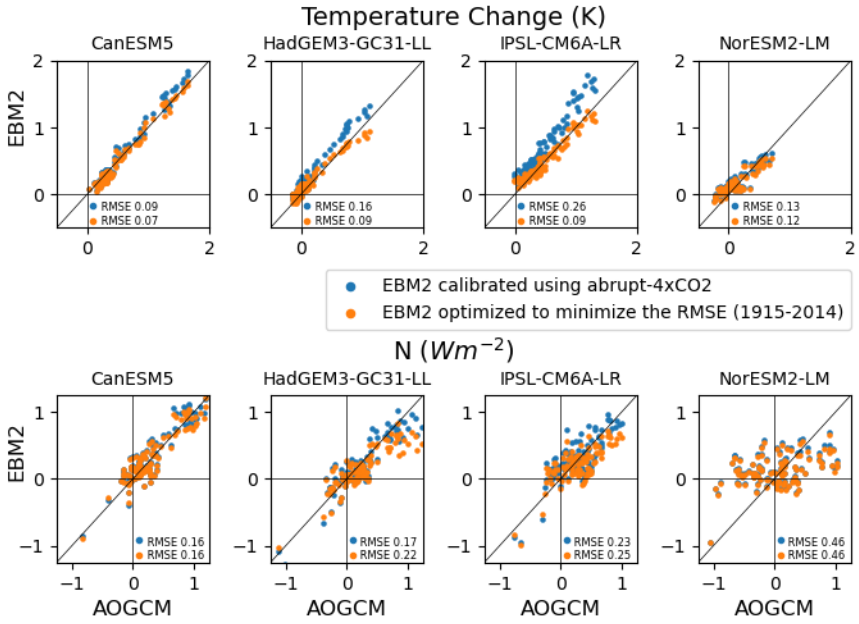

Figure 3. Projected changes in global mean temperature (top row) and energy balance at the TOA (N) (bottom row). Each panel shows changes in the AOGCM (x-axis) against the EBM2 emulation (y-axis). Each point represents an annual mean during 1915-2014.

3.4 Future near-surface temperature projections

We compare temperature emulations for the twenty-first century from EBM2 based on the different methods for calibrating λ and ε (Figure 4). Results are shown for five of the eight AOGCMs where the most complete CMIP6 data is available. Results for other models and experiments are shown in Figure S4.

The performance of the abrupt-4xCO2 calibration varies greatly between the AOGCMs and typically performs worse than the step model (Figure S4). For four of the AOGCMs, the emulations of SSP2-4.5 deteroriate 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. The exception is MIROC6 for which the abrupt-4xCO2 calibrated EBM2 performs well throughout 1850-2100 and across the three simulations. For NorESM2-LM, SSP2-4.5 is relatively closely emulated but SSP2-4.5-AER is not. Optimization of the λ and ε parameters (the "1850-2100" calibration in Figure 4) yielded close emulations for all of the AOGCMs and across the three 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.

The "1850-2014" calibration yields a close emulation of temperatures to 2014 but errors increase strongly after the calibration period. Extending the calibration period from 1850-2014 to 1850-2040 (not shown) does improve 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 "1850-2100" calibration from SSP2-4.5 was applied to the SSP2-4.5-GHG and SSP2-4.5-AER experiments (the "SSP2-4.5" calibration in Figure 4). For both SSP2-4.5-GHG and SSP2-4.5-AER, the error for the "SSP2-4.5" calibration is greater than for the "1850-2100" calibration. The impact also varies between models and experiments in terms of the size of the impact and its temporal behaviour. For CanESM5 for instance, the difference in temperature emulation is evident early in the twentieth century for SSP2-4.5-AER compared to early in the twenty-first century for SSP2-4.5-GHG. 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 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 models (Figure 4 "Ensemble mean") has relatively small RMSEs (except for the 1850-2014 calibration). 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 (Figure S5). For the same RMSE, there are solutions with a strong feedback (λ) with weak pattern effect (ε), and solutions with a weak feedback with strong pattern effect.This shows that optimized values for the λ and ε parameters may not be robust estimates of climate feedback or the AOGCM pattern effect.

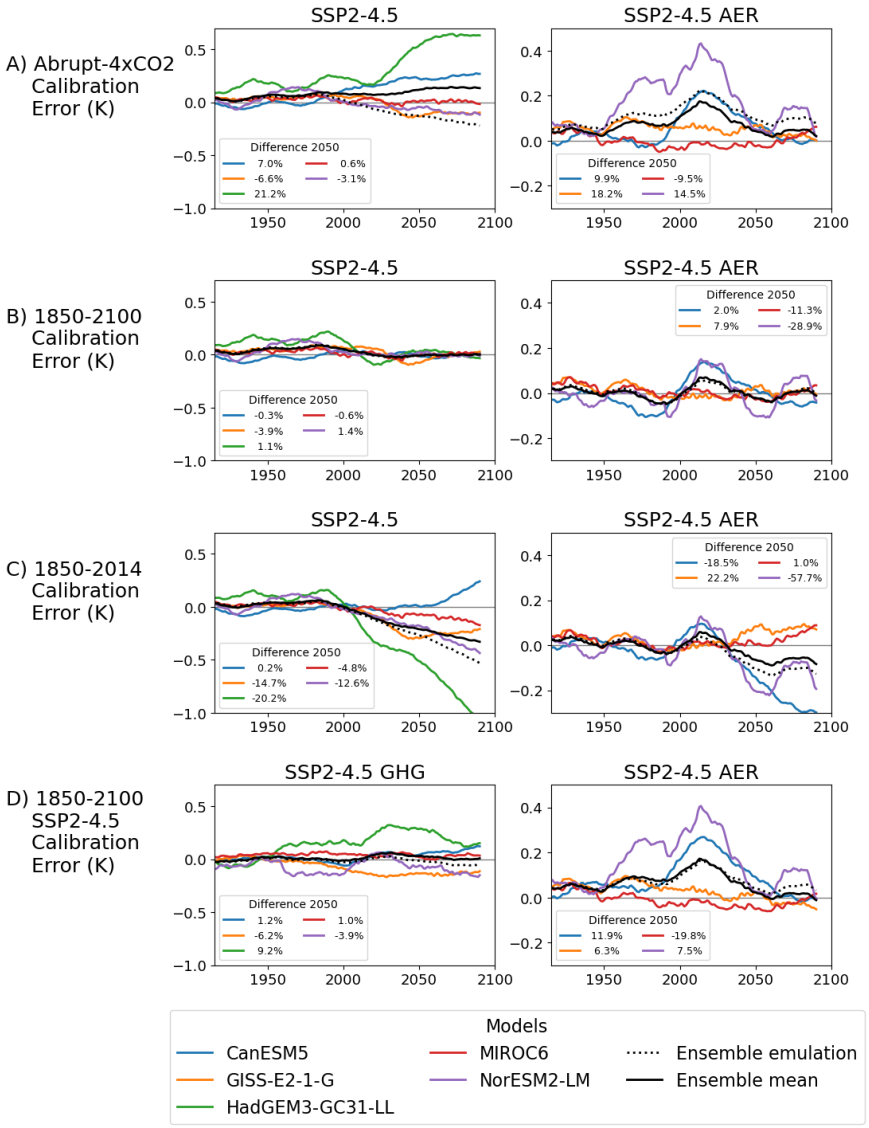

Figure 4. Differences between EBM2 emulations and AOGCM temperature projections. Rows show results for four calibrations of EBM2. Row B uses λ and ε parameter values which minimize the RMSE for temperatures during 1850-2100. Similarly, row C uses parameter values which minimize the RMSE during 1850-2014. In row D, EBM2 is 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. For plotting, 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 errors can be large, in many cases exceeding 20%. In this section, we discuss: the implications of our results; how emulations from EBM2 might be improved; and, the real-world relevance of our results.

We agree with Nicholls et al. (2021) that close emulation of the historical period is not sufficient to guarantee reliable emulation of future temperature changes. Late twentieth-century warming is suppressed by strong aerosol cooling (Smith and Forster 2021) and opposing errors in the emulation of GHG and aerosol forcings give a misleading impression of the accuracy of emulator performance. Further, opposing trends in GHG and aerosol forcings during the twenty-first century can cause a large divergence between AOGCM and EBM2 projections. Nicholls et al. (2021) found that many climate model emulators do not reliably emulate future projections from AOGCMs for high emissions scenarios. Our results also suggest that strong mitigation scenarios may not be reliably emulated.

EBM2 calibration using the abrupt-4xCO2 simulation 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 time periods where an AOGCM simulation is available, optimization of these parameters does not guarantee reliable out-of-sample projections. Further, without an AOGCM projection for a given AOGCM and scenario, it is not knowable if the EBM2 future projection will be reliable. This undermines trust in the EBM2 future projections.

Incorporating time varying feedbacks and an unforced pattern effect into EBM2 could reduce emulation errors and improve the reliability of future projections. 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) and future warming could be affected by future 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). To include 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)).

Improvements in the emulations by optimization of the λ and ε parameters could be implicitly compensating for errors arising from being unable to cleanly separate forcing and climate feedbacks in AOGCMs, as forcing estimates are dependent on the method used (Forster et al. 2013; Sherwood et al. 2015; Larson and Portmann 2016; Fredriksen et al. 2021). 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).

We optimized the λ and ε parameters by minizing the RMSE for temperature. Using the Hector emulator, Dorheim et al. (2020) show that minimizing errors for temperature and ocean heat flux produces more physically plausible parameter tunings than minimizing errors in temperature projections alone. Our initial investigations minimizing RMSE for temperature and N, however, showed that the emulation of historical temperatures was substantially worse than minimizing RMSE for temperature alone. Incorporating time varying feedbacks may mitigate this issue. Machine learning could also provide new techniques for calibrating and designing climate model emulators (Strobach and Bel, 2020; Watson-Parris, 2020).

There are several reasons why some AOGCMs are closely emulated and others not. First, some AOGCMs have greater symmetry in their responses to GHG and aerosol forcings (Figure 2) and EBM2 assumes symmetric responses to opposing forcings. Second, optimization of the λ and ε parameters (for temperature) yields closer emulations of N for some AOGCMs (Figure 3). Third, if EBM2 has a good representation of time varying feedbacks and the evolution of pattern effects in a AOGCM, model structural error is smaller. Finally, with small ensemble sizes, some of the variation in emulation errors arises from chance.

One approach for managing the variability in emulation errors between AOGCMs is to use a multi-model ensemble. Multi-model ensembles can be used to estimate structural uncertainty (e.g., Tebaldi and Knutti, 2007) and typically offer improved skill over individual climate models (e.g., Hagedorn et al. 2005). Our AOGCM ensemble is small, however, and we find that the ensemble mean of AOGCM emulations does not perform as well as the best AOGCM (Figure 4).

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 characterize uncertainty in emulating past climate. For future projections, AOGCMs remain an essential tool for estimating out-of-sample prediction errors, as done in this study, and enable the use of optimization techniques for emulator calibration.

Acknowledgments

LSJ, ACM, TA and PMF were supported by the European Union's Horizon 2020 research and innovation 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.

Data Availability Statement

CMIP6 data were downloaded from the ESGF; publically available from https://esgf-node.llnl.gov/search/cmip6/. Code will be publically available with a DOI in a Zenodo repository.

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.

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.

Byrne, B., & Goldblatt, C. (2013). Radiative forcing at high concentrations of well-mixed greenhouse gases. *Geophys. Res. Lett.* , 41, 152–160, doi:10.1002/2013GL058456.

Cummins, D. P., Stephenson, D. B., & 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.

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.

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.

Hagedorn, R., Doblas-Reyes, F. J. & Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. Tellus 57A, 219–233.

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. Pean, 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.Yelekci, 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.

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., Frolicher, 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., 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.

Smith, C. J., & Forster, P. M. (2021). Suppressed Late-20th Century Warming in CMIP6 Models Explained by Forcing and Feedbacks.*Geophysical Research Letters* , *48* (19). https://doi.org/10.1029/2021GL094948.

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.

Strobach, E., & Bel, G. (2020). Learning algorithms allow for improved reliability and accuracy of global mean surface temperature projections.*Nature Communications* , *11* (1). https://doi.org/10.1038/s41467-020-14342-9.

Tebaldi, C., & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. Phil. Trans. R. Soc. A (2007) 365, 2053–2075, doi:10.1098/rsta.2007.2076.

Watson-Parris D. (2021). Machine learning for weather and climate are worlds apart. *Phil. Trans. R. Soc. A.* **379:** 20200098, doi.org/10.1098/rsta.2020.0098.

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.

Hosted file

`suppinfo_draft_initial_submission.docx` available at https://authorea.com/users/539495/articles/599935-errors-in-simple-climate-model-emulations-of-past-and-future-global-temperature-change

L. S. Jackson[1], A. C. Maycock[1], T. Andrews[2], C. J. Smith[1], and P. M. Forster[3]

[1] School of Earth and Environment, University of Leeds, Leeds, UK. [2] Met Office Hadley Centre, Exeter, UK. [3] 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 model parameters 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 prediction errors can be large (greater than 0.5°C in a given year) 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. Time-varying and state-dependent feedbacks may contribute to 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 which can reproduce projections from complex climate models at lower cost and more quickly. In this study, we use a climate model emulator to reproduce projections 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 the Working Group I Sixth Assessment Report (Forster et al. 2021; Lee et al. 2021) 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- " 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 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 model 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 model 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 (Eyring et al. 2016) AOGCMs using realistic, out-of-sample climate projections and latest assessments of effective radiative forcing (ERF). Furthermore, the contribution of irreducible model 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 benchmarked against an impulse-response step model and a three-layer EBM.

2 Methods and data

2.1 Impulse-response step model

We use a step model (Good et al. 2011) to provide a benchmark 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 (Byrne & Goldblatt 2013). 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 model were produced by convolution of annual changes in ERF and the step-response functions.

2.2 Two-layer EBM

In the two-layer EBM (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 model layer is determined from:

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

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

Where C representations heat capacity, T temperature, F ERF, the climate feedback parameter and 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 ( ) 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 (Table S1) were calibrated for each AOGCM using a single realization of a CMIP6 abrupt-4xCO2 simulation (Table S1).

2.3 Calibration of EBM2 using linear optimization

As an alternative to abrupt-4xCO2 calibration, we use a linear optimization algorithm (scipy.optimize.minimize v1.6.2) to optimize the and parameters by minimizing the root mean squared error (RMSE) of the emulated temperatures compared to the AOGCM. The temperature projections are less sensitive to changes in the other EBM2 parameters (i.e., C_0, C_1, and), so these parameters are unchanged from their abrupt-4xCO2 calibrations. We also applied the linear optimization methodology to the abrupt-4xCO2 simulations and affirmed the calibrated parameter values of Geoffroy et al. (2013b).

2.4 Three-layer EBM

We use a three-layer EBM (EBM3) (Cummins et al. 2020) as a second benchmark for EBM2 performance. We follow the method of Cummins et al. (2020) to calibrate EBM3 parameters for each AOGCM 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. The Detection and Attribution Model Intercomparison Project (DAMIP) experiments (Gillett et al. 2016) are used for projections of temperature change attributed to different sources of ERF. RFMIP experiments (Pincus et al. 2016; Smith et al. 2021) are used for estimates of ERF during the historical period and ERF projected to 2100 under SSP2-4.5. Following Forster et al. (2013), unforced drift is removed from the AOGCM projections using the preindustrial control simulation.

3 Results

3.1 Historical period

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 ($\sim$ 0.07 K) arises

from interannual variations in the AOGCM ensemble means that is not captured in the emulations (there are three members in each AOGCM historical ensemble). The performance of EBM2, however, varies substantially between AOGCMs. 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, GFDL-ESM4 and NorESM2-LM). For IPSL-CM6A-LR, temperatures are overestimated by the emulators throughout 1915-2014. Intriguingly, close emulation of temperatures in abrupt-4xCO2 does not guarantee close emulation for the historical period (e.g. GFDL-ESM4), and a relatively poor emulation of abrupt-4xCO2 does not prohibit close emulation for the historical period (e.g. CNRM-CM6-1) (Figure S1).

The step model produces emulations with RMSEs equivalent to or less than emulations from EBM2 in seven of the eight AOGCMs. The exception is NorESM2-LM which has relatively large inter-annual variability and is the only model to show an apparent cooling trend during years 20-50 of its abrupt-4xCO2 simulation (Figure S1).

EBM3 performs better than EBM2 for abrupt-4xCO2, which is expected given the additional timescales resolved by the third layer. The additional degrees of freedom enable a much closer emulation of temperatures during years 10-40 of the abrupt-4xCO2 experiment, a period when the rate of temperature increase weakens rapidly (Figure S1). However, the improvement of EBM3 over EBM2 in the abrupt-4xCO2 experiment does not consistently translate to the historical experiment. Indeed there are two AOGCMs for which EBM2 has smaller RMSEs than EBM3 (MIROC6 and IPSL-CM6A-LR). Both EBMs overestimate temperatures for 1990-2014 in four of the eight AOGCMs and generally produce larger RMSEs than the step model.

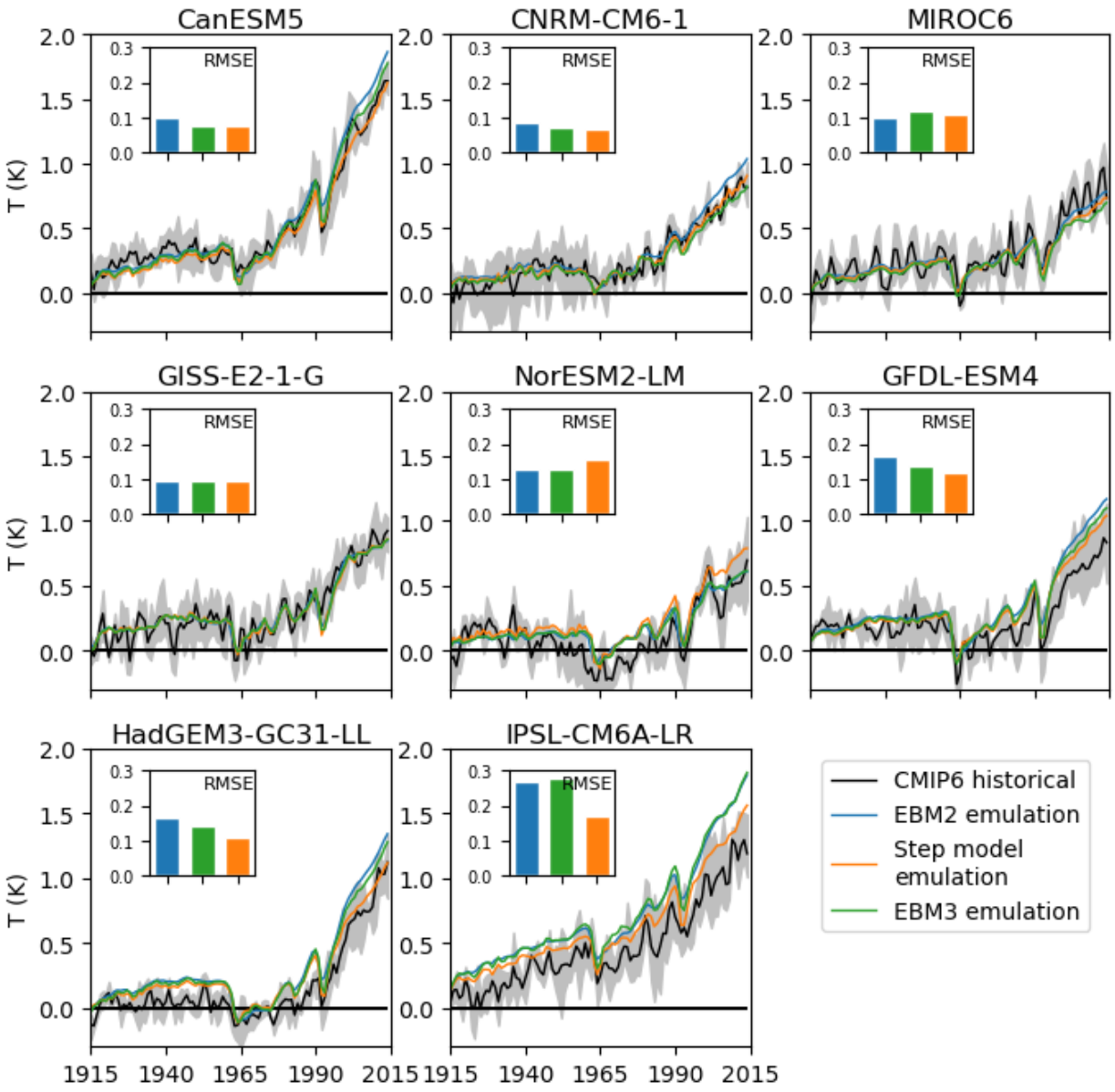

Figure 1. Global mean temperature anomalies from a 1850-1900 baseline for CMIP6 AOGCMs. The range between the ensemble maximum and minimum temperature changes is shown by gray shading. Changes in temperatures are forced by historical forcings during 1850-2014 and are shown for the period 1915-2014. RMSEs are calculated over 1915-2014.

3.2 Roles of different forcings for near-surface temperature change

In 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 contrast with the other AOGCMs (NorESM2-LM).

Although EBM2 was calibrated using abrupt-4xCO2, errors predominantly arise from emulation of the response to GHG forcing; in part because GHG has the largest ERF. The EBM2 emulations overestimate the temperature increase due to GHGs for HadGEM3-GC31-LL, IPSL-CM6A-LR and CanESM5 (even though the CanESM5 historical fit is good). In contrast, the EBM2 emulation underestimates the temperature response to GHGs for NorESM2-LM.

Emulation of the temperature response to aerosol forcing is the largest source of error in one model (NorESM2-LM). For all models, 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 model produces closer emulations of temperature for both GHG and aerosol forcings.

Emulation of the temperature response to natural forcings is a small source of error for the eight AOGCMs and the emulations are mostly within the spread of the AOGCM ensemble (Figures 2 and S2). 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 contribute to closer emulations (Held et al. 2010; Gregory et al. 2016).

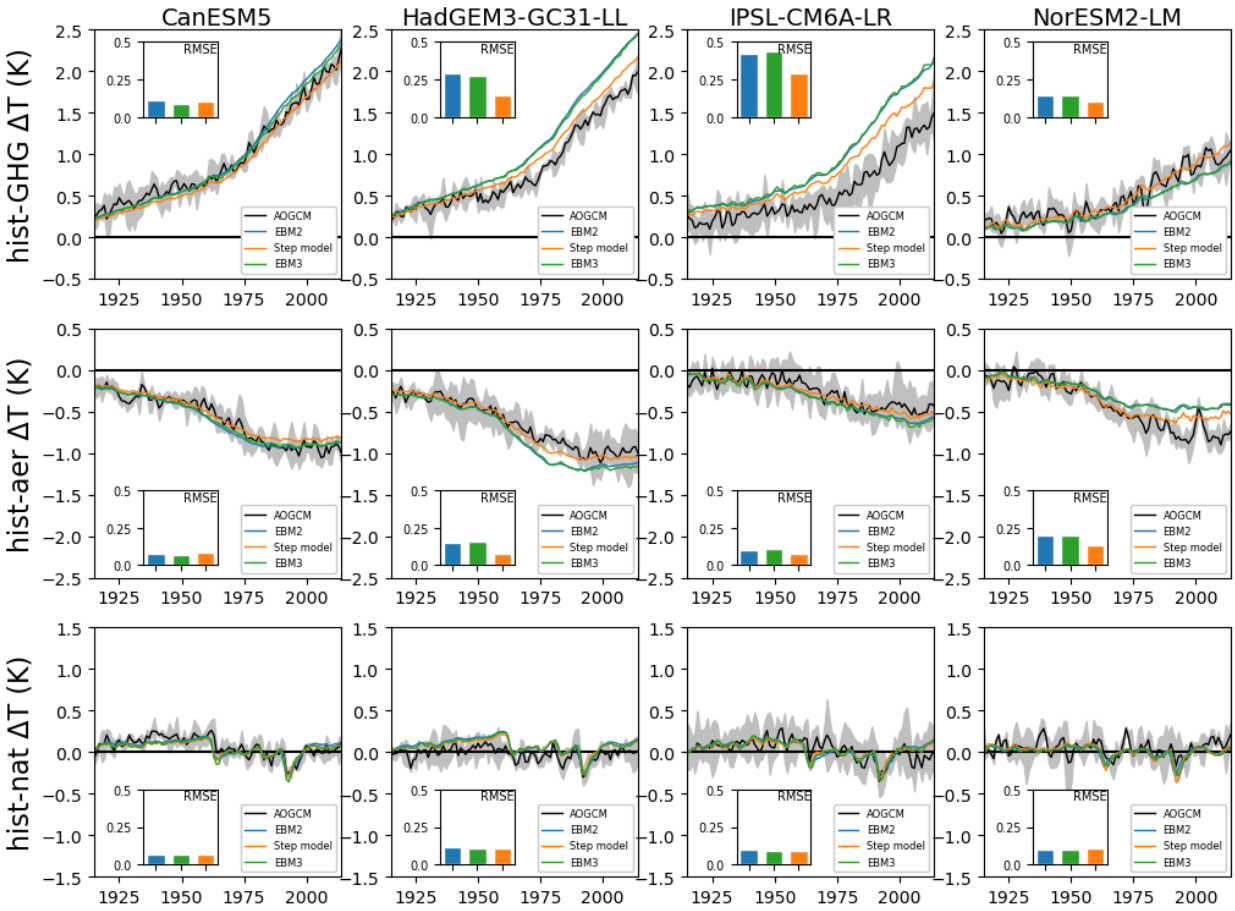

Figure 2. As Figure 1, except that temperature changes are forced by historical greenhouse gas (top row), anthropogenic aerosol (middle row), and natural (bottom row) forcings from RFMIP.

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 and parameters (Figure 3, Tables S2 and S3).

This improves the emulations for all 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 radiative flux 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 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 and IPSL-CM6-LR 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 constrained the and parameters separately for GHG and aerosol forcing using the DAMIP experiments. The constrained parameter values differ for the two types of forcing (Tables S2 and S3). Constrained parameter values also vary when RMSE is minimized over different periods of time.

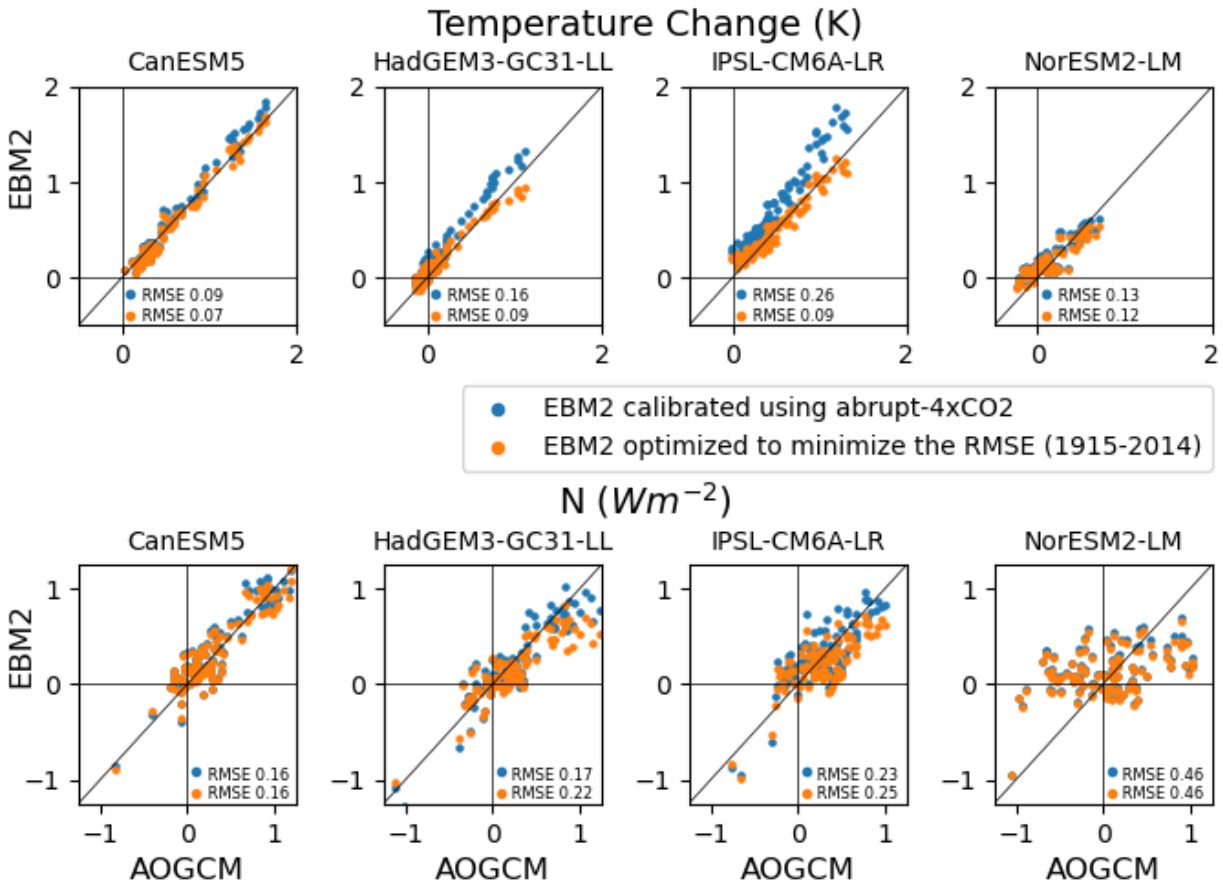

Figure 3. Projected changes in global mean temperature (top row) and energy balance at the TOA (N) (bottom row). Each panel shows changes in the AOGCM (x-axis) against the EBM2 emulation (y-axis). Each point represents an annual mean during 1915-2014.

3.4 Future near-surface temperature projections

We compare temperature emulations for the twenty-first century from EBM2 based on the different methods for calibrating and (Figure 4). Results are shown for five of the eight AOGCMs where the most complete CMIP6 data is available. Results for other models and experiments are shown in Figure S4.

The performance of the abrupt-4xCO2 calibration varies greatly between the AOGCMs and typically performs worse than the step model (Figure S4). For four of the AOGCMs, the emulations of SSP2-4.5 deteroriate 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. The exception is MIROC6 for which the abrupt-4xCO2 calibrated EBM2 performs well throughout 1850-2100 and across the three simulations. For NorESM2-LM, SSP2-4.5 is relatively closely emulated but SSP2-4.5-AER is not. Optimization of the and parameters (the "1850-2100" calibration in Figure 4) yielded close emulations for all of the AOGCMs and across the three 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.

The "1850-2014" calibration yields a close emulation of temperatures to 2014 but errors increase strongly after the calibration period. Extending the calibration period from 1850-2014 to 1850-2040 (not shown) does improve 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 "1850-2100" calibration from SSP2-4.5 was applied to the SSP2-4.5-GHG and SSP2-4.5-AER experiments (the "SSP2-4.5" calibration in Figure 4). For both SSP2-4.5-GHG and SSP2-4.5-AER, the error for the "SSP2-4.5" calibration is greater than for the "1850-2100" calibration. The impact also varies between models and experiments in terms of the size of the impact and its temporal behaviour. For CanESM5 for instance, the difference in temperature emulation is evident early in the twentieth century for SSP2-4.5-AER compared to early in the twenty-first century for SSP2-4.5-GHG. 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 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 models (Figure 4 "Ensemble mean") has relatively small RMSEs (except for the 1850-2014 calibration). 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 and parameters when minimizing the RMSE (Figure S5). For the same RMSE, there are solutions with a strong feedback ( ) with weak pattern effect ( ), and solutions with a weak feedback with strong pattern effect.This shows that optimized values for the and parameters may not be robust estimates of climate feedback or the AOGCM pattern effect.

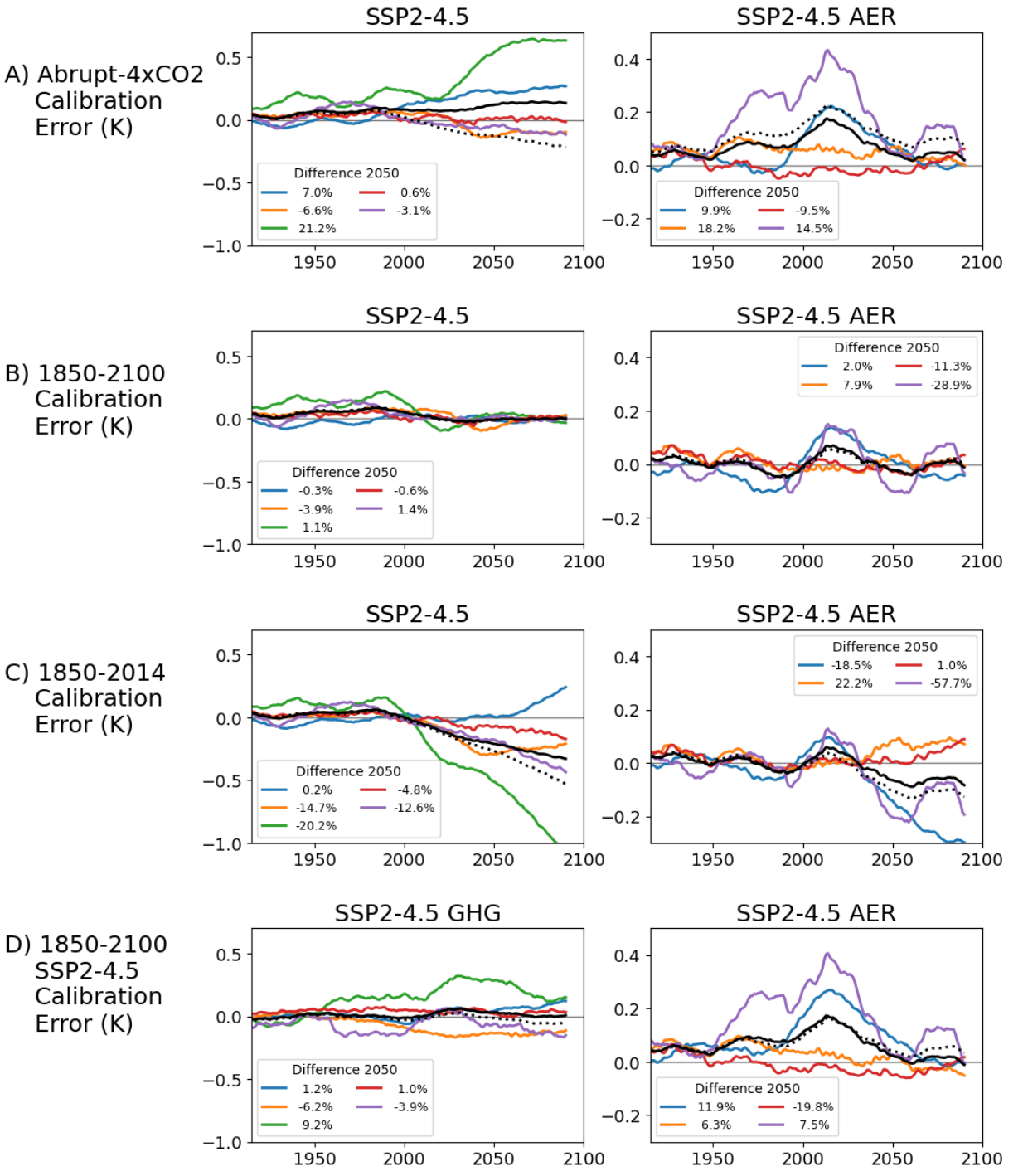
A) Abrupt-4xCO2 Calibration Error (K)
SSP2-4.5
SSP2-4.5 AER
Difference 2050
7.0% 0.6%
-6.6% -3.1%
21.2%
Difference 2050
9.9% -9.5%
18.2% 14.5%
B) 1850-2100 Calibration Error (K)
SSP2-4.5
SSP2-4.5 AER
Difference 2050
-0.3% -0.6%
-3.9% 1.4%
1.1%
Difference 2050
2.0% -11.3%
7.9% -28.9%
C) 1850-2014 Calibration Error (K)
SSP2-4.5
SSP2-4.5 AER
Difference 2050
0.2% -4.8%
-14.7% -12.6%
-20.2%
Difference 2050
-18.5% 1.0%
22.2% -57.7%
D) 1850-2100 SSP2-4.5 Calibration Error (K)
SSP2-4.5 GHG
SSP2-4.5 AER
Difference 2050
1.2% 1.0%
-6.2% -3.9%
9.2%
Difference 2050
11.9% -19.8%
6.3% 7.5%
1950
2000
2050
2100
Models
CanESM5
GISS-E2-1-G
HadGEM3-GC31-LL
MIROC6
NorESM2-LM
Ensemble emulation
Ensemble mean

Figure 4. Differences between EBM2 emulations and AOGCM temperature projections. Rows show results for four calibrations of EBM2. Row B uses and parameter values which minimize the RMSE for temperatures during 1850-2100. Similarly, row C uses parameter values which minimize the RMSE during 1850-2014. In row D, EBM2 is 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. For plotting, 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 errors can be large, in many cases exceeding 20%. In this section, we discuss: the implications of our results; how emulations from EBM2 might be improved; and, the real-world relevance of our results.

We agree with Nicholls et al. (2021) that close emulation of the historical period is not sufficient to guarantee reliable emulation of future temperature changes. Late twentieth-century warming is suppressed by strong aerosol cooling (Smith and Forster 2021) and opposing errors in the emulation of GHG and aerosol forcings give a misleading impression of the accuracy of emulator performance. Further, opposing trends in GHG and aerosol forcings during the twenty-first century can cause a large divergence between AOGCM and EBM2 projections. Nicholls et al. (2021) found that many climate model emulators do not reliably emulate future projections from AOGCMs for high emissions scenarios. Our results also suggest that strong mitigation scenarios may not be reliably emulated.

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

Incorporating time varying feedbacks and an unforced pattern effect into EBM2 could reduce emulation errors and improve the reliability of future projections. 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) and future warming could be affected by future 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). To include 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)).

Improvements in the emulations by optimization of the and parameters could be implicitly compensating for errors arising from being unable to cleanly separate forcing and climate feedbacks in AOGCMs, as forcing estimates are dependent on the method used (Forster et al. 2013; Sherwood et al. 2015; Larson and Portmann 2016; Fredriksen et al. 2021). 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).

We optimized the and parameters by minizing the RMSE for temperature. Using the Hector emulator, Dorheim et al. (2020) show that minimizing errors for temperature and ocean heat flux produces more physically plausible parameter tunings than minimizing errors in temperature projections alone. Our initial investigations minimizing RMSE for temperature and N, however, showed that the emulation of historical temperatures was substantially worse than minimizing RMSE for temperature alone. Incorporating time varying feedbacks may mitigate this issue. Machine learning could also provide new techniques for calibrating and designing climate model emulators (Strobach and Bel, 2020; Watson-Parris, 2020).

There are several reasons why some AOGCMs are closely emulated and others not. First, some AOGCMs have greater symmetry in their responses to GHG and aerosol forcings (Figure 2) and EBM2 assumes symmetric responses to opposing forcings. Second, optimization of the and parameters (for temperature) yields closer emulations of N for some AOGCMs (Figure 3). Third, if EBM2 has a good representation of time varying feedbacks and the evolution of pattern effects in a AOGCM, model structural error is smaller. Finally, with small ensemble sizes, some of the variation in emulation errors arises from chance.

One approach for managing the variability in emulation errors between AOGCMs is to use a multi-model ensemble. Multi-model ensembles can be used to estimate structural uncertainty (e.g., Tebaldi and Knutti, 2007) and typically offer improved skill over individual climate models (e.g., Hagedorn et al. 2005). Our AOGCM ensemble is small, however, and we find that the ensemble mean of AOGCM emulations does not perform as well as the best AOGCM (Figure 4).

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 characterize uncertainty in emulating past climate. For future projections, AOGCMs remain an essential tool for estimating out-of-sample prediction errors, as done in this study, and enable the use of optimization techniques for emulator calibration.

Acknowledgments

LSJ, ACM, TA and PMF were supported by the European Union's Horizon 2020 research and innovation 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.

Data Availability Statement

CMIP6 data were downloaded from the ESGF; publically available from https://esgf-node.llnl.gov/search/cmip6/. Code will be publically available with a DOI in a Zenodo repository.

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.

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.

Byrne, B., & Goldblatt, C. (2013). Radiative forcing at high concentrations of well-mixed greenhouse gases. *Geophys. Res. Lett.*, 41, 152–160, doi:10.1002/2013GL058456.

Cummins, D. P., Stephenson, D. B., & 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.

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.

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.

Hagedorn, R., Doblas-Reyes, F. J. & Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. Tellus 57A, 219–233.

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.

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., 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.

Smith, C. J., & Forster, P. M. (2021). Suppressed Late-20th Century Warming in CMIP6 Models Explained by Forcing and Feedbacks. *Geophysical Research Letters*, *48*(19). https://doi.org/10.1029/2021GL094948.

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.

Strobach, E., & Bel, G. (2020). Learning algorithms allow for improved reliability and accuracy of global mean surface temperature projections. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-14342-9.

Tebaldi, C., & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. Phil. Trans. R. Soc. A (2007) 365, 2053–2075, doi:10.1098/rsta.2007.2076.

Watson-Parris D. (2021). Machine learning for weather and climate are worlds apart. *Phil. Trans. R. Soc. A.* **379:** 20200098, doi.org/10.1098/rsta.2020.0098.

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.