A novel framework to evaluate climate model emulators for global warming projections

Manali S. Nayak[1], Kyle C. Armour[1,2], David S. Battisti[1]

[1]Department of Atmospheric and Climate Science, University of Washington, Seattle, Washington, USA
[2]School of Oceanography, University of Washington, Seattle, Washington, USA

Key Points:

- We propose a novel framework for evaluating climate model emulators for use in global warming projections
- Calibrating a widely-used emulator to match the historical warming of GCMs does not ensure agreement in future warming projections
- Emulator errors in climate feedback and ocean heat uptake efficiency can lead to incorrect warming or compensate to produce accurate warming

Corresponding author: Manali Nayak, `manalin@uw.edu`

Abstract

Climate model emulators were extensively used in the IPCC's Sixth Assessment Report due to their computational efficiency and consistency with key climate metrics. The emulators were calibrated to historical observations and used for future climate projections without systematic evaluation of their robustness. Here, we develop a framework to evaluate emulator performance against global climate model (GCM) large ensembles. This is demonstrated by constraining a two-layer energy balance model (EBM) to historical simulations of four GCMs and comparing their 21st century warming projections. The EBM matches projected warming in three of the four GCMs but exhibits substantial spread across ensemble members. It fails to reproduce the time-evolving global climate feedback in GCMs, with compensating biases between feedbacks and ocean heat uptake efficiency allowing seemingly accurate projections for incorrect reasons. Our results underscore the importance of evaluating the accuracy and physical realism of climate model emulators before using them for warming projections.

Plain Language Summary

Global climate models (GCMs) are essential tools for predicting the climate response to human-driven activities, but they are computationally expensive and known to contain biases. Climate model emulators use simplified representations of climate processes to replicate GCM behavior and are advantageous due to their computational efficiency and ability to be calibrated to match climate observations. The recent Intergovernmental Panel on Climate Change Sixth Assessment Report (IPCC AR6) used emulators extensively, including for global warming projections. The emulators were calibrated to historical observations but were not robustly evaluated for their accuracy in simulating future warming. In this study, we develop a framework based on the AR6 methodology to evaluate emulator performance by constraining them using historical simulations from four state-of-the-art GCMs and comparing their 21st century warming projections. We demonstrate the framework for a widely used emulator and show that it captures average warming well in three of the four GCMs but exhibits substantial spread in future warming. Moreover, the emulator fails to capture key physical processes, in some cases achieving the correct average warming for incorrect reasons. Our results demonstrate the need to evaluate climate model emulators against GCMs before relying on them for warming projections in future IPCC Reports.

1 Introduction

The Intergovernmental Panel on Climate Change (IPCC) periodically publishes reports synthesizing the state of our scientific understanding of past climate change, projecting future changes, assessing potential impacts, and informing adaptation and mitigation strategies. While global climate models (GCMs) have historically served as the primary tool for projecting global temperature, the IPCC Sixth Assessment Report (AR6) augmented GCM simulations with additional lines of evidence (Lee et al., 2021). A major innovation of AR6 was the extensive use of climate model emulators (also known as reduced-complexity climate models; Nicholls et al., 2020) which contain simplified representations of key climate processes that can replicate aspects of GCM behavior and historical climate observations.

Emulators provide two main advantages over GCMs. First, they are computationally efficient, allowing their use in a wide range of applications in AR6 including: attributing warming to specific forcing agents; estimating the contribution of non-CO_2 emissions to remaining carbon budgets; projecting global mean sea level rise; and projecting global warming over the 21st century and beyond (IPCC, 2021). Second, while the GCMs participating in Phase 6 of the Coupled Model Intercomparison Project (CMIP6; Eyring et al., 2016) on average have too-high climate sensitivity and global warming over recent decades (Meehl et al., 2020; Forster et al., 2020; Zelinka et al., 2020; Fredriksen et al., 2023; Forster

et al., 2021), emulators can be calibrated to match the assessed ranges of both climate sensitivity and historical temperature trends (Forster et al., 2021). Moreover, emulators have been shown to replicate the climate response of GCMs under idealized CO_2 forcing scenarios (Geoffroy et al., 2013a, 2013b; Forster et al., 2021). These features provided confidence in the use of emulators for temperature projections in AR6 and suggest that they will play a key role in future IPCC Assessment Reports.

One climate model emulator, known as the Finite Amplitude Impulse Response (FaIRv1.3) model (Smith et al., 2018), was used extensively throughout AR6. Its temperature response is based on a two-layer energy balance model (EBM) (Held et al., 2010; Millar et al., 2017) that approximates global mean surface warming on two timescales: a fast timescale for the surface components of the climate system and a slower timescale associated with the deep ocean response. AR6 (Forster et al., 2021) calibrated a FaIR ensemble for consistency with the assessed distributions of equilibrium climate sensitivity (ECS, i.e., the equilibrium global surface temperature change in response to CO_2 doubling), as well as observed global warming and ocean heat uptake, and then used it for applications including global warming projections.

The far-reaching implications of the use of emulators for climate projections motivate the need for a robust framework to evaluate their performance. As noted above, Geoffroy et al. (2013a, 2013b) showed that the two-layer EBM can accurately replicate the response of GCMs to abrupt CO_2 quadrupling (*abrupt4xCO$_2$*) when its parameters are calibrated correctly. However, Jackson et al. (2022) found that calibration to an *abrupt4xCO$_2$* simulation does not guarantee an accurate emulation of GCM warming under historical forcing. Given that the two-layer EBM was designed to emulate the temperature response to CO_2 forcing alone, these results raise concerns about its use in emulating the response to historical and future forcing, which is strongly influenced by non-CO_2 greenhouse gases, aerosols, and natural forcing agents. This raises a key question about the use of emulators in general: How reliable are their future projections when calibrated to the observational record, which is strongly influenced by non-CO_2 forcings?

We propose a framework for rigorously evaluating climate emulators against GCM simulations. We use the two-layer EBM as an example because of its ability to capture essential features of the climate system despite its structural simplicity and because it comprises the physical climate component of FaIR, which was extensively used in AR6. However, this framework is relevant for other emulators as well. We constrain a two-layer EBM ensemble to match historical simulations of each of four CMIP6 GCM large ensembles — analogously to the AR6 approach of constraining it to the observational record — and compare the future warming projection of the EBM to that of the GCM it is calibrated to.

2 Data and Methods

2.1 CMIP6 Output

We evaluate the two-layer EBM using the following CMIP6 GCMs (number of ensemble members in parentheses): CanESM5 (25), HadGEM3-GC31-LL (5), IPSL-CM6A-LR (11), and MIROC6 (50). These models are chosen because each provides multiple *historical* and Shared Socioeconomic Pathway (SSP) forcing ensemble members, enabling calculation of the forced response over the period 1850-2100, and the output needed to calculate effective radiative forcing (ERF) through their participation in the Radiative Forcing Model Intercomparison Project (RFMIP; Pincus et al., 2016). We use the CMIP monthly-mean variables *tas* (near-surface air temperature), *rsut*, *rsdt rlut*, (top-of-the-atmosphere (TOA) shortwave upwelling, shortwave downwelling, and longwave upwelling fluxes, respectively) from *historical* (1850-2014) and *SSP2-4.5* (2015-2100) simulations. We calculate radiative imbalance at the TOA as *rsdt - rsut - rlut* and calculate globally and annually averaged anomalies relative to the mean over 1850-1900.

The ERF quantifies the impact of forcing agents on the TOA energy budget, including the radiative response to rapid atmospheric adjustments but excluding the response to changes in surface temperature (Myhre et al., 2014). In the RFMIP protocol, each model has a *piClim-histall* simulation wherein the atmosphere component of the GCM is run with sea-surface temperatures (SST) and sea-ice concentrations (SIC) fixed to pre-industrial values, while all radiative forcing agents are prescribed to vary following the CMIP6 *historical* (1850-2014) and *SSP2-4.5* (2015-2100) simulations. The availability of *piClim-histall* simulations extending to 2100 limited our analysis to the four GCMs listed above. For each GCM, we calculate the ERF as the difference in the annual and global TOA radiation anomaly between the *piClim-histall* simulation and an average over a 30-year *piClim-control* simulation in which SSTs, SICs, and forcing agents are fixed to pre-industrial values. Following Hansen et al. (2005), we apply a correction to the ERF by subtracting the TOA radiative response to warming over land and sea ice, which is considered part of the feedback rather than part of the forcing. The Supporting Information provides further details and shows the ERF time series diagnosed from the four GCMs (Text S1; Fig. S1).

2.2 Two-layer Energy Balance Model (EBM): Structure and Calibration to GCMs

The version of the FaIR emulator used in AR6 (v1.6.2) simulates the global temperature response to forcing using a two-layer energy balance model (Held et al., 2010):

$$C\frac{dT}{dt} = F + \lambda_{eq}T - \varepsilon\gamma(T - T_0), \tag{1}$$

$$C_0\frac{dT_0}{dt} = \gamma(T - T_0), \tag{2}$$

where T is the temperature anomaly of the upper layer, which represents the quickly-responding surface components of the climate system including the atmosphere, land, sea ice and the ocean mixed layer; T_0 is the temperature anomaly of the lower layer, which represents the slowly-responding components of the climate system including the deep ocean; C and C_0 are, respectively, the heat capacities of the upper and lower layers; F is the ERF; λ_{eq} is the equilibrium radiative feedback parameter; γ is the heat exchange coefficient, representing the strength of coupling between the two layers; and ε is the ocean heat uptake efficacy, which captures the evolution of the global radiative feedback with changing ocean heat uptake (Winton et al., 2010; Held et al., 2010; Geoffroy et al., 2013b).

For AR6, a FaIR ensemble was formed using prior probability distributions of model parameters based on best estimates of the ERF, climate response parameters, and carbon cycle parameters (Smith et al., 2021). Ensemble members were then selected based on their agreement with observed global warming, ocean heat uptake, and atmospheric CO_2 concentrations. In this study, we follow a similar approach by constraining a 1000-member ensemble of the two-layer EBM using individual GCM ensemble members, treating each GCM realization as an analogue to the observational record used to constrain the AR6 FaIR ensemble.

Following Dvorak et al. (2022), we randomly draw EBM parameter values (C, C_0, γ, ε) for the prior ensemble from probability distributions centered on estimates obtained by calibrating the EBM to *abrupt4xCO$_x$* simulations of CMIP5 models (Geoffroy et al., 2013a, 2013b), with standard deviations for all prior parameter distributions expanded by 50% (Text S2; Fig. S2). We use a uniform prior distribution of ECS from 1°C to 10°C and global radiative feedback values at equilibrium are taken to be $\lambda_{eq} = -F_{2\times}/\text{ECS}$, where $F_{2\times} = 3.71\ Wm^{-2}$ is the radiative forcing under doubling of CO_2 from pre-industrial levels (Myhre et al., 1998; Smith et al., 2020), which is similar to the value of $F_{2\times}$ in each GCM considered here (Armour et al., 2024).

For each GCM, we generate a prior EBM ensemble with a size equal to 1000 times the number of GCM ensemble members. We then force each EBM ensemble member by the ERF time series diagnosed from that GCM over the period 1850-2020. Figure 1a uses MIROC6 as an example to illustrate the procedure for generating the prior EBM ensemble's historical warming. The ensemble mean temperature anomaly of the EBM only agrees with that of the GCM during the first few decades and increasingly overestimates warming over the 20th century. The increasing spread and deviation of the EBM ensemble from the GCM arise from combinations of prior parameter values that produce global mean warming and ocean heat uptake changes that are inconsistent with those of the GCM.

We then produce a constrained EBM ensemble for each GCM (illustrated for MIROC6 in Fig. 1b) by retaining ensemble members that satisfy the condition

$$\sqrt{(\frac{\delta T}{\sigma_T})^2 + (\frac{\delta N}{\sigma_N})^2} < 1.65, \tag{3}$$

where δT and δN are calculated as mean differences in global surface temperature and ocean heat uptake, respectively, between each EBM member and the corresponding GCM member over 2000-2020 relative to 1850-1900; and σ_T and σ_N are the standard deviations of the annual mean global surface temperature and ocean heat uptake, respectively, calculated across each GCM's ensemble members (using anomalies over 2000-2020 relative to 1850-1900), thus representing uncorrelated internal variability across ensemble members. The value of 1.65 corresponds to a confidence level of 90%. The period 2000-2020 is similar to that used for the constraint in AR6, chosen because it captures the climate response to greenhouse gas forcing while minimizing the influence of internal variability, volcanic eruptions, and large changes in tropospheric aerosol forcing. For comparison (black dashed lines in Figs. 1a,b), we also run the EBM with parameter values obtained by calibrating the EBM to *abrupt4xCO$_2$* simulations of the GCMs (Armour et al., 2024).

Table S1 and Fig. S2 provide the posterior parameter values for all GCM calibrations. Across the four GCMs, the posterior distributions for λ_{eq} and γ become more tightly constrained relative to their priors, while posterior distributions for ε, C and C_0, remain similar to their priors (Fig. S2; see Section 4 for a discussion). The posterior values of λ_{eq} obtained from the historical simulation constraint (Eq. (3)) are consistently more negative than those obtained from calibrating the EBM to the *abrupt4xCO$_2$* simulations, while values of γ, ε, and C_0 are generally larger (with the exception of MIROC6).

Figure 1b illustrates that the ensemble-mean temperature evolution of the constrained EBM (with posterior parameter values) accurately matches that of MIROC6 up to the year 2020, with a substantially reduced spread compared to the prior ensemble (Fig. 1a). The historical temperature anomaly from the EBM simulation run with parameters calibrated to the *abrupt4xCO$_2$* simulation also agrees with the GCM ensemble mean throughout 1850-2020 (Fig. 1a), despite discrepancies in parameter estimates (Table S1). These features are true of the EBM constrained by the historical simulations of each of the other three GCMs as well. We are now in a position to answer the question we posed above about whether constraining the EBM to replicate global warming and ocean heat uptake over the historical period guarantees accurate projections of global temperature over the 21st century.

3 Results

3.1 Temperature projections in the constrained EBM ensemble

Having obtained EBM ensembles constrained to match each GCM's historical global warming and ocean heat uptake, we seek to evaluate whether their warming projections remain consistent with those of the corresponding GCMs through 2100 under *SSP2-4.5* forcing. We use posterior parameter sets obtained in Section 2.2 to perform EBM simulations using each

ERF time series over 1850-2100 (Figs. 1c-f). We find that EBM ensemble mean warming is close to (within two standard deviations of) GCM ensemble mean warming for three of the four GCMs, with the exception being IPSL-CM6A-LR for which the EBM underestimates mean end-of-century warming roughly 0.4°C. The GCM standard deviation is a measure of the spread across the ensemble members (Fig. 1). Importantly, the spread of warming projected by each EBM ensemble spans the warming projected by its corresponding GCM.

We also use the parameter sets from calibration to the *abrupt4xCO$_2$* simulations for EBM simulations driven by each ERF time series over 1850-2100. The projected warming substantially overestimates that found in the GCMs beyond the early 21st century, with the exception of MIROC6 (Fig. 1c-f). This is in line with expectations as the EBM calibration to the *abrupt4xCO$_2$* simulation produced higher values of ECS compared to the posterior mean values obtained from the EBM constrained over the historical period (Table S1). That the EBM calibrated to *abrupt4xCO$_2$* simulations produces a good match to GCM warming over the historical period but biased projections of 21st century warming is in broad agreement with the results of Jackson et al. (2022). This illustrates the importance of evaluating climate model emulators using realistic historical and future forcing simulations of GCMs.

When constrained to match historical warming and ocean heat uptake, the EBM produces an accurate match to projected 21st century warming in three out of the four GCMs used in this study. However, the EBM ensemble spread increases substantially after the constraint window (i.e., after 2020), and the fraction of EBM ensemble members that remain within the GCM spread by 2100 is small (Fig. 1c-f). For example, only 35% of the EBM ensemble members fall within the MIROC6 range while only 10% fall within the range of HadGEM3-GC31-LL. The large spread in the EBM ensembles arises despite being constrained using exact GCM values of global warming and ocean heat uptake (to within internal variability) and forced with the exact ERF time series derived directly from each GCM, and thus devoid of any uncertainty. These results indicate that substantial uncertainty in future warming would persist even if we had perfect observations of historical warming, ocean heat uptake, and radiative forcing because the historical record is too short to fully constrain key physical climate parameters that are relevant for future warming (e.g., λ_{eq} and ε in the case of the two-layer EBM; see Section 4).

3.2 The roles of global climate feedback and ocean heat uptake efficiency

Since the ensemble mean warming in all EBM historical calibrations is similar to that in most corresponding GCMs, it is important to evaluate whether agreement in future warming is achieved through correct representation of the physical processes or whether it results from compensating biases in the EBM parameters. We examine this by considering the EBM within the standard energy imbalance framework (Gregory & Mitchell, 1997; Gregory et al., 2002):

$$N = F + \lambda T, \tag{4}$$

where N is the anomalous global TOA energy imbalance; F is the ERF diagnosed from each GCM and applied exactly to each corresponding EBM ensemble; and λ is the global net radiative feedback, which may be time dependent and thus differs from the equilibrium feedback λ_{eq}. T is taken to be the upper layer temperature response in the EBM, and the global near-surface air temperature anomaly in the GCM. To a good approximation, the net TOA energy flux is equal to the energy absorbed by the ocean (Raper et al., 2002; Gregory & Mitchell, 1997). Under transient warming, ocean heat uptake is approximately proportional to global surface temperature change such that $N = \kappa T$, where the ocean heat uptake efficiency κ is a measure of the rate at which heat is removed from the upper ocean to depth. Figure 2 shows the time series of the net radiative feedback and ocean heat

uptake efficiency calculated from the ensemble-mean T and N for each GCM and EBM, respectively, as

$$\lambda = \frac{N - F}{T}, \tag{5}$$

$$\kappa = \frac{N}{T}. \tag{6}$$

All GCMs exhibit a decrease in the magnitude of λ over time, consistent with expectations that the effective climate sensitivity will increase as equilibrium is approached (e.g., Senior & Mitchell, 2000; Andrews et al., 2015; Armour, 2017). This weakening of λ has been attributed to its dependence on evolving spatial patterns of SST (Armour et al., 2013; Rose et al., 2014; Gregory & Andrews, 2016; Dong et al., 2019, 2020; Andrews et al., 2015), particularly associated with initially-delayed and later-enhanced warming of the eastern Pacific and Southern Oceans relative to the global mean. In contrast, λ calculated in the EBM does not show much variation in time. This EBM behavior can be understood by considering λ diagnosed from the EBM parameters: adding Eqs. (1) and (2) yields the net radiative imbalance at the TOA (N), which can then be substituted into Eq. (5) to derive

$$\lambda = \lambda_{eq} - (\varepsilon - 1)\gamma(1 - \frac{T_0}{T}). \tag{7}$$

As evident from Eq. (7), the temporal variation in λ within the EBM arises due to a changing ratio of deep ocean to surface temperature anomalies (T_0/T). Because T_0 evolves negligibly compared to T over 1850-2100 in the EBM simulations (Fig. S3), λ is relatively constant. Moreover, the factor $(\varepsilon - 1)$ is small since ε is close to unity, further suppressing the time dependence of λ. Consequently, the EBM produces a nearly-constant λ that aligns with that of the GCM only during the constraint window (2000-2020), and is unable to capture the strong negative λ of the GCM earlier in the record or the changes in λ over the 21st century. However, the EBM is still capable of matching GCM projected warming because most of the GCM time variation in λ occurs before the period 2000-2020 such that using a nearly-constant λ over the 21st century does not substantially bias EBM warming projections.

In contrast, the EBM effectively captures the time variation of κ across all GCMs, particularly the late 20th-century increase followed by a gradual decline in the 21st century. The behavior in the late 20th century is consistent with observations (Cael, 2022; Watanabe et al., 2013) and has been attributed to natural forcings, particularly the eruption of Mount Pinatubo in 1991 (Shi et al., 2025). The EBM evolution of κ can be understood by considering

$$\kappa = \frac{F}{T} + \lambda \approx \gamma(1 - \frac{T_0}{T}), \tag{8}$$

where this approximation for κ holds on timescales longer than that of upper-layer adjustment ($C\,dT/dt \cong 0$). In the EBM, the accurate time variation of κ appears to be driven primarily by the F/T term in Eq. (8), implying that the common ERF causes κ diagnosed from the EBM to closely track that diagnosed from the GCM. Hence, offsets in κ such as those in IPSL-CM6A-LR (Fig. 2f) and HadGEM3-GC31-LL (Fig. 2h) likely originate from an offset in λ (Eq. (7)). Equation (8) also provides insight into the overall time dependence of κ: over the late 20th century, κ becomes larger as the ratio of deep ocean to surface temperature (T_0/T) decreases due to rapidly increasing forcing, while over the 21st century, κ becomes smaller as T_0/T increases as the system equilibrates to more-constant forcing.

We also find that λ and κ in the EBM calibrated to the GCM *abrupt4xCO$_2$* simulations mirror the temporal evolution of the EBM ensemble mean λ and κ from the EBM constrained

by the historical GCM simulations, although λ in these simulations is consistently weaker than the EBM mean (Fig. 2). In general, the EBM inadequately represents the temporal evolution of λ as seen in the GCMs, but emulates the evolution of κ reasonably well.

3.3 Attribution of errors

Although the EBM's representation of λ and κ is similar across GCM calibrations, it remains unclear whether trade-offs between these variables affect the skill of the EBM in projecting future temperatures. To understand the contributions of λ and κ to the temperature anomalies of each EBM ensemble, we use the 40-year running means of λ and κ diagnosed from each ensemble member (see Section 3.2) to reconstruct the global temperature time series for that member by rearranging Eqs. (5) and (6) to give

$$T = \frac{F}{\kappa - \lambda}. \tag{9}$$

From Eq. (9), we obtain the probability distributions of the mean warming over 2080-2100 for both the EBM and GCM in Fig. 3 (left column). Note that, given λ and κ, Eq. (9) provides an exact reproduction of the warming in each EBM or GCM ensemble member by construction. To evaluate the contribution of λ and κ in setting the end-of-century warming of the EBM, we use Eq. (9) for each EBM ensemble member but replace either λ or κ with its corresponding value from the GCM ensemble member to which it was constrained, giving the probability distributions T_λ (purple) and T_κ (green), respectively (right column in Fig. 3).

For MIROC6 and CanESM5 (compare Figs. 3a,b and Figs. 3c,d), the EBM values of λ and κ already closely match those of the corresponding GCM, and thus replacing the EBM values with those of the GCM does not result in a considerable difference in projected warming. However, for IPSL-CM6A-LR (compare Figs. 3e,f) where the EBM underestimates the GCM warming, we find that both T_λ and T_κ are roughly halfway (2.97°C) between the mean temperature anomalies of the EBM (2.83°C) and GCM (3.13°C). This indicates that errors in both λ and κ contribute to errors in warming simulated by the EBM for this GCM. These errors appear to compensate for each other to provide a good EBM match to GCM warming over the constraint window (2000-2020), but lead to projected warming that is too low. However, in the case of HadGEM3-GC31-LL (Fig. 3g,h), the EBM accurately matches GCM warming, but T_λ shows lower warming (by 0.16°C) while T_κ shows higher warming (by 0.08°C). This indicates that providing an accurate value for either λ or κ does not improve the warming projection of the EBM relative to the GCM (on the contrary, it worsens it), and that the EBM is able to emulate GCM warming with inaccurate values for λ and κ because their errors compensate for each other over the 21st century. Overall, while the EBM simulates global warming reasonably well, our framework for evaluating it against GCMs reveals cases where it fails to do so due to errors in λ and κ (e.g., IPSL-CM6A-LR) or where it produces accurate emulations due to compensation of these errors (e.g., HadGEM3-GC31-LL).

4 Discussion and Conclusions

Climate model emulators played a key role in IPCC AR6 due to their computational efficiency and consistency with assessed ranges for key climate metrics such as ECS, historical global warming, and ocean heat uptake. These emulators were subsequently used to project future warming without a systematic evaluation of their accuracy. In this study, we develop a framework to assess the reliability of future warming projections from climate model emulators by evaluating them against state-of-the-art GCM ensembles.

We demonstrate this framework using a two-layer EBM which was widely used in AR6. For consistency with AR6 methodology, we constrain the EBM parameters to match historical

simulations of four CMIP6 GCMs and compare their 21st century warming projections. We find that the EBM closely reproduces GCM ensemble mean warming in three out of four cases, with the exception being IPSL-CM6A-LR where the EBM underestimates end-of-century warming by nearly 0.4°C. However, in all cases, the EBM ensemble spread is wide enough to span the median warming of its corresponding GCM, suggesting that the EBM constrained to historical observations could similarly provide a useful bound on future warming. The spread of the constrained EBM ensemble increases steadily after the calibration period (2000-2020), with fewer than 40% of the EBM members within the GCM spread of internal variability by 2100. This suggests that irreducible uncertainty in future warming would remain even if the EBM could be constrained by historical observations without any uncertainty.

To better understand the source of errors in warming emulation, we diagnose the net climate feedback parameter λ and ocean heat uptake efficiency κ from both the EBM and the GCMs. Unlike the GCMs, the EBM simulates a nearly-constant λ because it depends only on the ratio of deep ocean to surface temperature change (the ratio T_0/T) which remains small. This exposes a structural limitation of the EBM: it was designed to replicate λ changes under a slow evolution of changing global ocean heat uptake within GCMs equilibrating to constant CO_2 forcing through the use of the heat uptake efficacy ε (Winton et al., 2010); it is unable to represent decadal-scale variations in λ associated with more-rapid changes in SST patterns (i.e., those driven by non-CO_2 forcings or internal variability, which likely dominate changes in λ over the historical period). Non-CO_2 forcing agents are known to generate unique radiative feedbacks due to their differing spatial distributions (Hansen et al., 2005; Marvel et al., 2016; Zhou et al., 2023). Assigning unique efficacy parameters to individual forcing agents could improve the emulation of λ and projected warming in the two-layer EBM, but this has not been tested here.

How well the EBM can emulate GCM projected warming is also sensitive to the choice of constraint window used. Applying the constraint (Eq. (3)) over the period 1980-2000 results in a significantly poorer emulation of future warming, with a much larger spread in the EBM ensemble (Fig. S4). The reason for this difference is that when the EBM is constrained over 1980-2000, λ is constrained to be more negative compared to using the period 2000-2020 (compare Fig. S5 and Fig. 2). Since the EBM is unable to produce much time variation in λ, this overly negative value of λ persists over the 21st century leading to weaker projected warming. The 2000-2020 period is therefore a more suitable choice for constraining the EBM because it avoids the late 20th century period in which λ is temporarily very negative, which could be driven in part by factors such as large tropospheric aerosol or volcanic forcing (Gregory & Andrews, 2016).

Fig. 2 shows that both λ and κ in the EBM arrive closest to their corresponding GCM values during the constraint window (2000-2020). Considering timescales longer than that of ocean mixed layer adjustment ($C\,dT/dt \approx 0$), and considering the fact that the deep layer does not warm considerably through the early 21st century ($T_0 \approx 0$), Eqs. (1) and (2) give $N \approx \gamma T$. Since the EBM is constrained using T and N, γ is well constrained. Similarly, κ is also constrained via Eq. (6). Rearranging Eq. (1) gives $T = -F/(\lambda_{eq} - \epsilon\gamma)$, indicating that the quantity $(\lambda_{eq} - \epsilon\gamma)$ is constrained when F is specified to match that of the GCMs. Importantly, the historical period provides no direct constraints on λ_{eq} or ϵ individually, resulting in a large spread in projected 21st century warming. The constrained quantity $(\lambda_{eq} - \epsilon\gamma)$ is equivalent to $(\lambda - \kappa)$ (Eq. (7)), which is well constrained only during the constraint window (2000-2020). The parameter ε is not well constrained because the historical record is too short for the deep layer to warm substantially (Figs. S2 and S3). While constraining the EBM using the period 1980-2000 worsens its ability to emulate GCM projected warming (Fig. S4), extending the constraint window into the 21st century may provide a better constraint on ε, and thus on λ_{eq}, as deep layer warming becomes more apparent.

Finally, we assess the contribution of errors from λ and κ to the end-of-century warming in the EBM. We find that even when the EBM ensemble mean matches the GCM warming, the agreement does not necessarily reflect an accurate representation of physical processes. For example, in the case of IPSL-CM6A-LR, errors in both λ and κ lead to a substantial underestimation of end-of-century warming. In contrast, for HadGEM3-GC31-LL, compensating errors in λ and κ coincidentally result in accurate emulation of future warming. These findings underscore the importance of evaluating the physical basis of emulator projections rather than measuring accuracy solely on the basis of agreement with GCM temperature output. In just the four GCMs considered here, we found that the same emulator (the two-layer EBM) performs differently, raising important questions: what distinguishes the GCMs for which emulators fail to accurately reproduce future warming and can we build additional confidence that emulators provide accurate warming projections when constrained by observations? Expanding the availability of RFMIP-style simulations extending through 2100 from more modeling groups would enable the application of this emulator evaluation framework using a wider range of GCMs.

Our proposed framework offers a systematic approach to evaluate emulators against GCMs, closely mirroring the methodology used in the IPCC AR6. We encourage researchers to apply this framework to evaluate other emulators to further our understanding of their strengths and limitations. Evaluating emulators in this manner is imperative to ensure knowledge of their uncertainties and potential limitations before their application to project future warming.

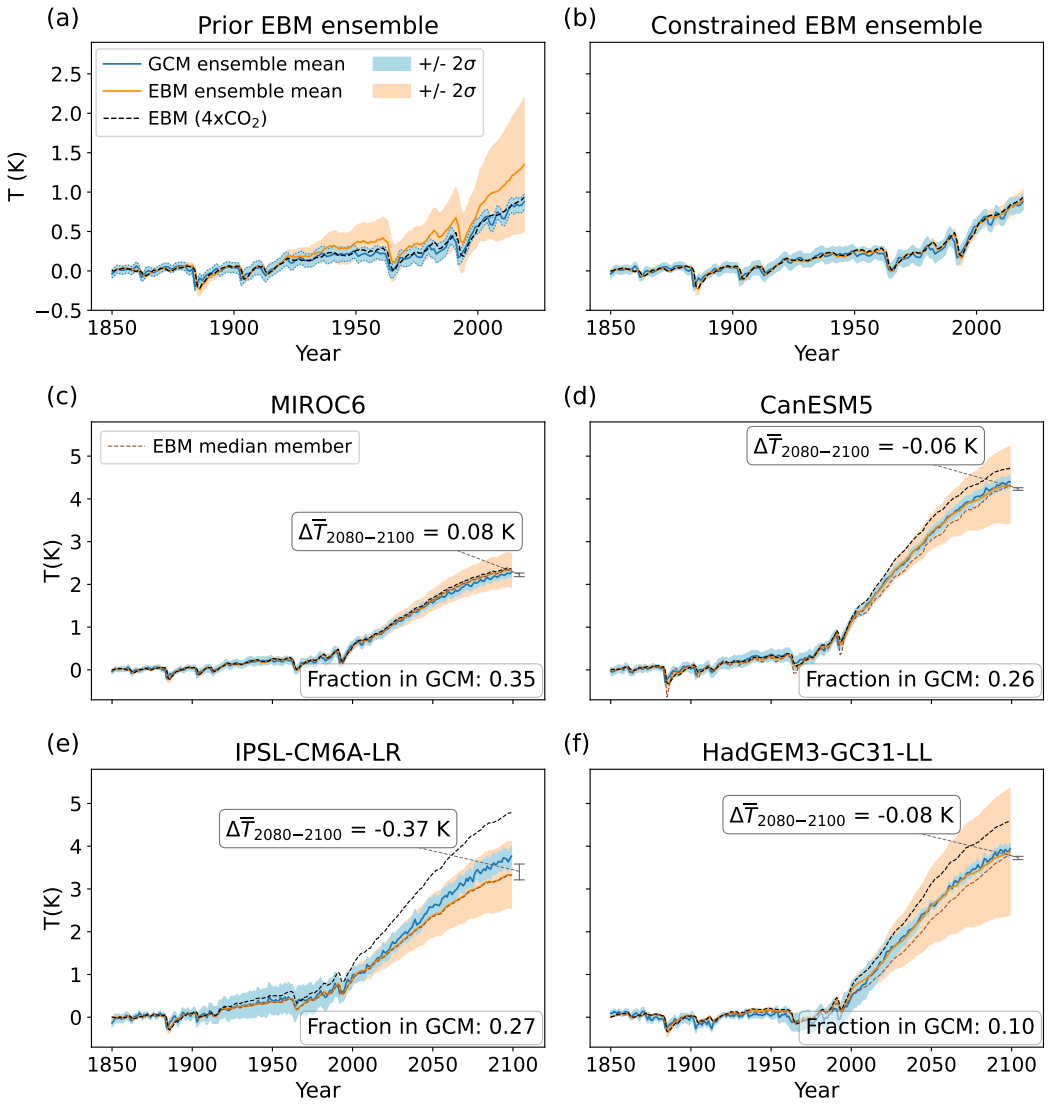

Figure 1. Time series of global and annual mean surface temperature anomaly over 1850-2020 for (a) prior and (b) constrained EBM ensemble compared with MIROC6. Lines denote MIROC6 ensemble mean (blue) and EBM ensemble mean (orange), with light blue and light orange shading representing two standard deviations around the GCM and EBM ensemble means, respectively. (c-f) Time series of global and annual mean surface temperature anomaly over 1850-2100 for the EBM constrained to 4 CMIP6 GCMs. The brown dashed line represents the median member (corresponding to the member with the median value of warming over 2080-2100) of the EBM ensemble. The temperature anomaly for the EBM with parameters fit to the $abrupt4xCO_2$ simulation of each GCM is shown in the dashed black line. In all panels, EBM ensemble standard deviations are calculated at each year as there is no representation of internal variability. GCM ensemble standard deviations are calculated across the 21-year running mean time series of the annual temperature anomaly of each member, with the values at 1860 and 2090 prescribed for the first and last ten years, respectively. Values $\Delta\overline{T}$ report the difference in ensemble mean EBM and GCM warming averaged over 2080-2100, and "Fraction in GCM" reports the fraction of EBM ensemble members falling within the $\pm 2\sigma$ range of GCM ensemble members averaged over 2080-2100.

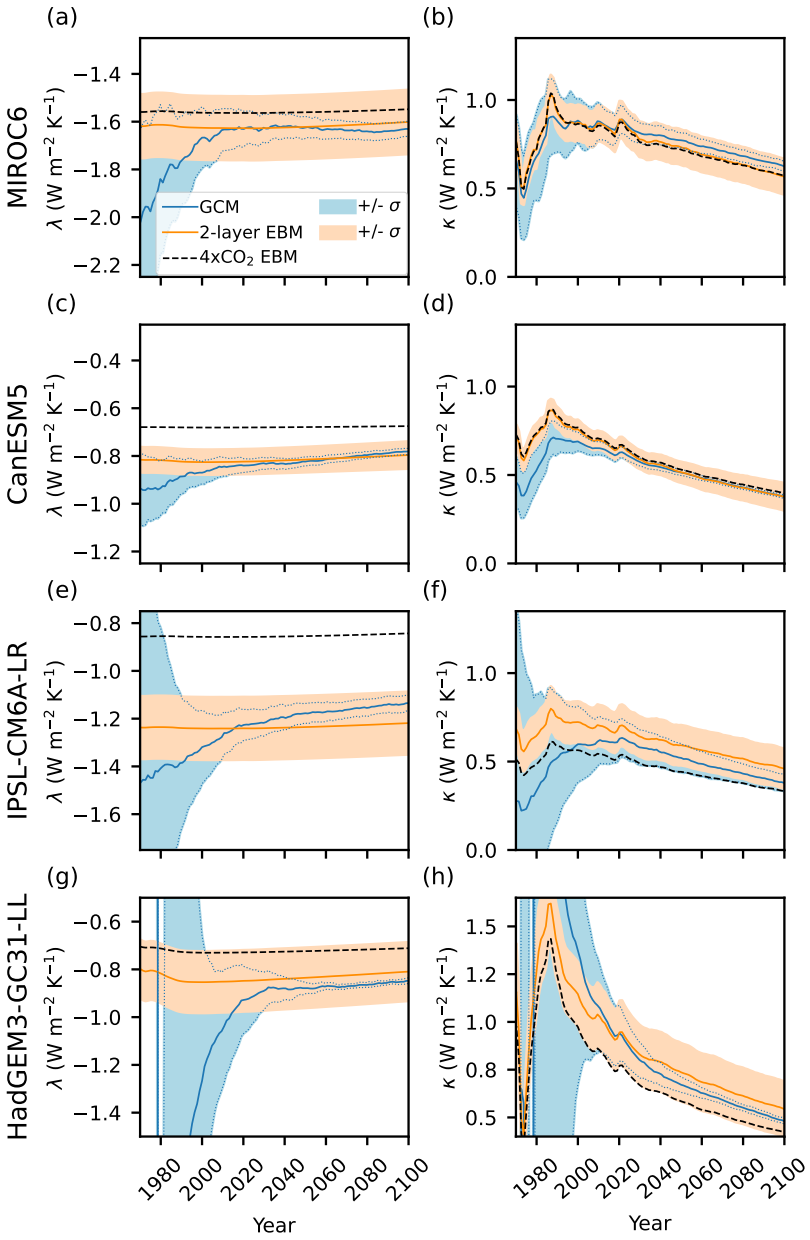

Figure 2. 40-year running means of (left column) radiative feedback parameter (λ) and (right column) ocean heat uptake efficiency (κ) calculated as the ensemble mean of each GCM (blue) and constrained EBM (orange). The blue and orange shading represent one standard deviation about the mean for the GCM and EBM respectively, with the blue dotted lines bounding the spread in the GCM feedbacks for clarity. The evolution of λ and κ fit to the *abrupt4xCO*$_2$ simulation of each GCM is shown in the dashed black line.

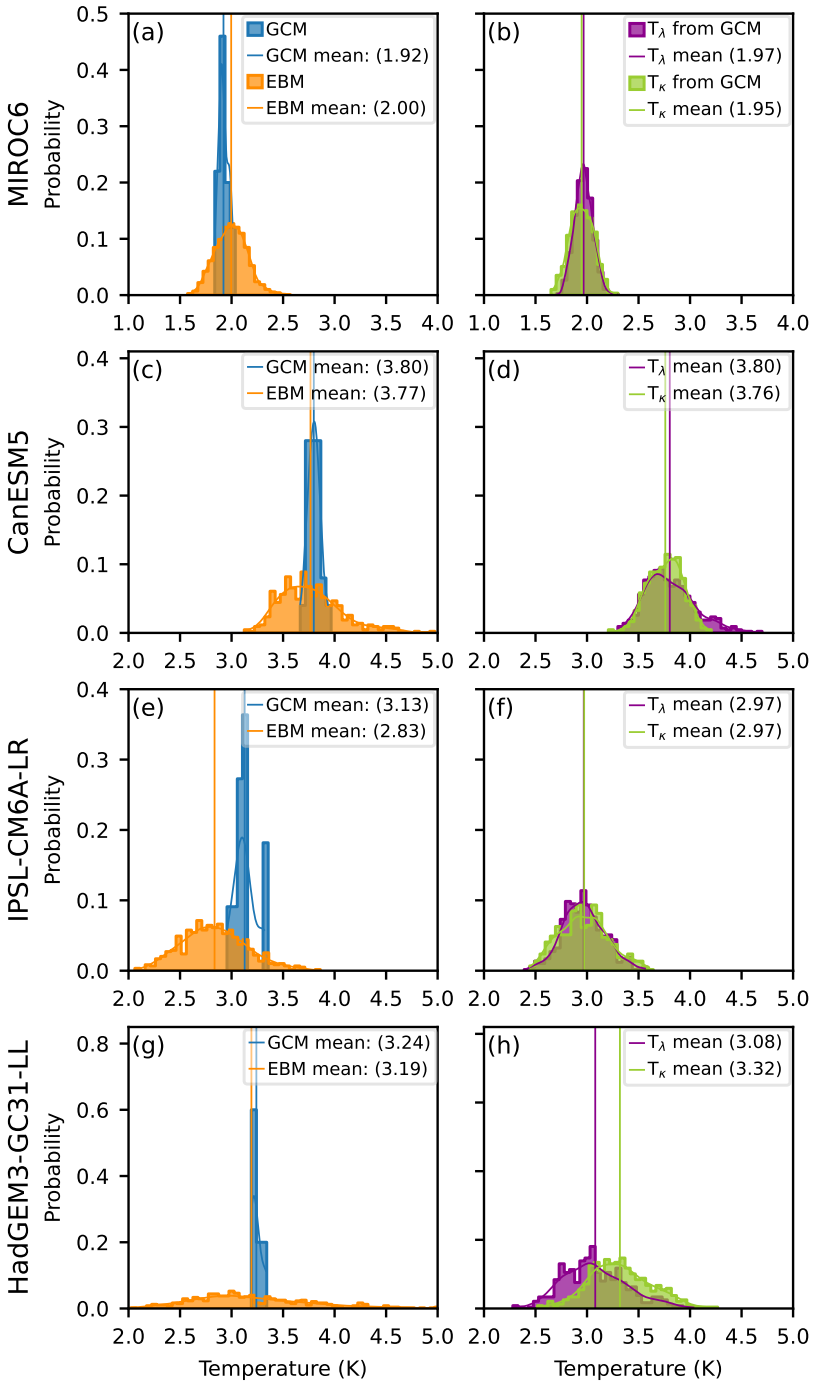

Figure 3. Probability distributions of average temperature anomalies over the last 20 years of the simulation period (2080-2100). (Left column) Reconstructed temperature anomalies from the radiative feedback parameter and ocean heat uptake efficiency from each GCM (blue) and constrained EBM (orange). (Right column) Reconstructed temperature anomalies obtained by substituting either the radiative feedback parameter (purple) or the ocean heat uptake efficiency (green) from the GCM, while keeping the remaining parameter from the EBM. Vertical lines indicate the mean for each distribution (values in legend).

Open Research Section

The original CMIP6 model data sets used in this study are accessible at the Earth System Grid Federation (ESGF) portal (https://aims2.llnl.gov/search/cmip6/).

Acknowledgments

M.N., K.C.A., and D.S.B were supported by National Science Foundation (NSF) Award AGS-2203543. M.N. and K.C.A. were supported by National Oceanic and Atmospheric Administration (NOAA) Modeling, Analysis, Predictions and Projections Program Award NA20OAR4310391. K.C.A. was supported by NSF Award AGS-1752796 and a Calvin Professorship in Oceanography. M.N. and D.S.B. were supported by the Tamaki Foundation. We thank Dargan Frierson, Camille Hankel, and Cristian Proistosescu for insightful discussion of this work.

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.

Armour, K. C. (2017). Energy budget constraints on climate sensitivity in light of inconstant climate feedbacks. *Nature Climate Change*, *7*(5), 331–335.

Armour, K. C., Bitz, C. M., & Roe, G. H. (2013). Time-varying climate sensitivity from regional feedbacks. *Journal of Climate*, *26*(13), 4518–4534.

Armour, K. C., Proistosescu, C., Dong, Y., Hahn, L. C., Blanchard-Wrigglesworth, E., Pauling, A. G., ... others (2024). Sea-surface temperature pattern effects have slowed global warming and biased warming-based constraints on climate sensitivity. *Proceedings of the National Academy of Sciences*, *121*(12), e2312093121.

Cael, B. (2022). Ocean heat uptake efficiency increase since 1970. *Geophysical Research Letters*, *49*(19), e2022GL100215.

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.

Dong, Y., Proistosescu, C., Armour, K. C., & Battisti, D. S. (2019). Attributing historical and future evolution of radiative feedbacks to regional warming patterns using a Green's function approach: The preeminence of the Western Pacific. *Journal of Climate*, *32*(17), 5471–5491.

Dvorak, M., Armour, K., Frierson, D., Proistosescu, C., Baker, M., & Smith, C. J. (2022). Estimating the timing of geophysical commitment to 1.5 and 2.0°C of global warming. *Nature Climate Change*, *12*(6), 547–552.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958.

Forster, P. M., Maycock, A. C., McKenna, C. M., & Smith, C. J. (2020). Latest climate models confirm need for urgent mitigation. *Nature Climate Change*, *10*(1), 7–10.

Forster, P. M., Storelvmo, T., Armour, K., Collins, W., Dufresne, J.-L., Frame, D., ... Zhang, H. (2021). The Earth's energy budget, climate feedbacks, and climate sensitivity. In V. Masson-Delmotte et al. (Eds.), *Climate change 2021: The Physical Science Basis. contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (p. 923-1054). Cambridge, UK and New York, NY, USA: Cambridge University Press. doi: 10.1017/9781009157896.009

Fredriksen, H.-B., Smith, C. J., Modak, A., & Rugenstein, M. (2023). 21st century scenario forcing increases more for CMIP6 than CMIP5 models. *Geophysical Research Letters*, *50*(6), e2023GL102916.

Geoffroy, O., Saint-Martin, D., Bellon, G., Voldoire, A., Olivié, D. J., & Tytéca, S. (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*(6), 1859–1876.

Geoffroy, O., Saint-Martin, D., Olivié, D. J., Voldoire, A., Bellon, G., & 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*(6), 1841–1857.

Gregory, J. M., & Andrews, T. (2016). Variation in climate sensitivity and feedback parameters during the historical period. *Geophysical Research Letters*, *43*(8), 3911–3920.

Gregory, J. M., & Mitchell, J. F. (1997). The climate response to CO_2 of the Hadley Centre coupled AOGCM with and without flux adjustment. *Geophysical Research Letters*, *24*(15), 1943–1946.

Gregory, J. M., Stouffer, R., Raper, S., Stott, P., & Rayner, N. (2002). An observationally based estimate of the climate sensitivity. *Journal of Climate*, *15*(22), 3117–3121.

Hansen, J., Sato, M., Ruedy, R., Nazarenko, L., Lacis, A., Schmidt, G., ... others (2005). Efficacy of climate forcings. *Journal of Geophysical Research: Atmospheres*, *110*(D18).

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*(9), 2418–2427.

IPCC. (2021). *Climate change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (V. Masson-Delmotte et al., Eds.). Cambridge, UK and New York, NY, USA: Cambridge University Press.

Jackson, L. S., Maycock, A. C., Andrews, T., Fredriksen, H.-B., Smith, C. J., & Forster, P. (2022). Errors in simple climate model emulations of past and future global temperature change. *Geophysical Research Letters*, *49*(15), e2022GL098808.

Lee, J.-Y., Marotzke, J., Bala, G., Cao, L., Corti, S., Dunne, J., ... Zhou, T. (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* (p. 553-672). Cambridge, UK and New York, NY, USA: Cambridge University Press.

Marvel, K., Schmidt, G. A., Miller, R. L., & Nazarenko, L. S. (2016). Implications for climate sensitivity from the response to individual forcings. *Nature Climate Change*, *6*(4), 386–389.

Meehl, G. A., Senior, C. A., Eyring, V., Flato, G., Lamarque, J.-F., Stouffer, R. J., ... Schlund, M. (2020). Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 earth system models. *Science Advances*, *6*(26), eaba1981.

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

Myhre, G., Highwood, E. J., Shine, K. P., & Stordal, F. (1998). New estimates of radiative forcing due to well mixed greenhouse gases. *Geophysical Research Letters*, *25*(14), 2715–2718.

Myhre, G., Shindell, D., Bréon, F.-M., Collins, W., Fuglestvedt, J., Huang, J., ... others (2014). Anthropogenic and natural radiative forcing. *Climate Change 2013-The Physical Science Basis*, 659–740.

Nicholls, Z. R., Meinshausen, M., Lewis, J., Gieseke, R., Dommenget, D., Dorheim, K., ... others (2020). Reduced Complexity Model Intercomparison Project Phase 1: introduction and evaluation of global-mean temperature response. *Geoscientific Model Development*, *13*(11), 5175–5190.

Pincus, R., Forster, P. M., & Stevens, B. (2016). The radiative forcing model intercomparison project (RFMIP): Experimental protocol for CMIP6. *Geoscientific Model*

Development, *9*(9), 3447–3460.

Raper, S. C., Gregory, J. M., & Stouffer, R. J. (2002). The role of climate sensitivity and ocean heat uptake on AOGCM transient temperature response. *Journal of Climate*, *15*(1), 124–130.

Rose, B. E., Armour, K. C., Battisti, D. S., Feldl, N., & Koll, D. D. (2014). The dependence of transient climate sensitivity and radiative feedbacks on the spatial pattern of ocean heat uptake. *Geophysical Research Letters*, *41*(3), 1071–1078.

Senior, C. A., & Mitchell, J. F. (2000). The time-dependence of climate sensitivity. *Geophysical Research Letters*, *27*(17), 2685–2688.

Shi, J.-R., Zanna, L., & Adcroft, A. (2025). The impact of natural external forcing on ocean heat uptake efficiency since the 1980s. *arXiv preprint arXiv:2504.06366*.

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

Smith, C. J., Kramer, R. J., Myhre, G., Alterskjær, K., Collins, W., Sima, A., ... others (2020). Effective radiative forcing and adjustments in CMIP6 models. *Atmospheric Chemistry and Physics*, *20*(16), 9591–9618.

Smith, C. J., Nicholls, Z. R. J., Armour, K., Collins, W., Forster, P., Meinshausen, M., ... Watanabe, M. (2021). The Earth's energy budget, climate feedbacks, and climate sensitivity supplementary material [Book Section]. 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* (p. 1-35). Cambridge, UK and New York, NY, USA: Cambridge University Press.

Watanabe, M., Kamae, Y., Yoshimori, M., Oka, A., Sato, M., Ishii, M., ... Kimoto, M. (2013). Strengthening of ocean heat uptake efficiency associated with the recent climate hiatus. *Geophysical Research Letters*, *40*(12), 3175–3179.

Winton, M., Takahashi, K., & Held, I. M. (2010). Importance of ocean heat uptake efficacy to transient climate change. *Journal of Climate*, *23*(9), 2333–2344.

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., ... Taylor, K. E. (2020). Causes of higher climate sensitivity in CMIP6 models. *Geophysical Research Letters*, *47*(1), e2019GL085782.

Zhou, C., Wang, M., Zelinka, M. D., Liu, Y., Dong, Y., & Armour, K. C. (2023). Explaining forcing efficacy with pattern effect and state dependence. *Geophysical Research Letters*, *50*(3), e2022GL101700.

Supporting Information for "A novel framework to evaluate climate model emulators for global warming projections"

Manali S. Nayak[1], Kyle C. Armour[1,2], David S. Battisti[1]

[1]Department of Atmospheric and Climate Science, University of Washington, Seattle, Washington, USA

[2]School of Oceanography, University of Washington, Seattle, Washington, USA

Contents of this file

1. Table S1
2. Text S1 and S2
3. Figures S1 to S5

Introduction

This supplement provides Table S1, which presents the posterior values of the EBM parameters after constraining to each GCM considered in this study. Values from calibrating the EBM to the *abrupt4xCO*$_2$ simulations of those GCMs are also given. Text S1 provides details on the correction factor used in the calculation of the effective radiative forcing (ERF). Text S2 provides details on the probability distributions defined for the EBM prior ensemble. Figs. S1-S5 support the results discussed in the main paper. The effective ra-

diative forcing (ERF) obtained from all global climate models (GCMs) and the prior and posterior distributions of the parameters in the two-layer energy balance model (EBM) from each GCM calibration are shown in Figures S1 and S2. Figure S3 demonstrates the slower deep layer temperature evolution (T_0) in comparison to the upper layer (T) for MIROC6. Figures 2 and 3 in the main paper are reproduced in Figures S4 and S5 where the EBM calibration to each of the GCMs is done over the period 1980-2000.

Text S1: CMIP6 output for ERF and calculation of correction factor

For each GCM, the average of three available *piClim-histall* simulations is taken for all variables (*tas, rsut, rsdt, rlut*) prior to analysis to reduce variability. Following the calculation of the ERF (see Section 2.1 in the main paper), a correction factor is subtracted from ERF to account for the radiative response to surface warming in the fixed sea-surface temperature simulations. The correction factor is calculated as $\lambda_{eq}T$, where T is the global temperature anomaly and λ_{eq} is the equilibrium climate feedback, which is estimated by calibrating the two-layer EBM to the *abrupt4xCO*$_2$ simulation of each model (Armour et al., 2024). We calculate T as the difference between the annual and global mean near-surface air temperatures in the *piClim-histall* and *piClim-control* simulations. The ERF time series for all four GCMs is shown in Fig. S1.

Text S2: Parameter distributions for EBM prior ensemble

The EBM parameter values are randomly drawn from probability distributions centered on estimates from Geoffroy et al. (2013a, 2013b), with standard deviations expanded by 50%. Normal distributions are defined for C (mean = 8.2 $\mathrm{W\,yr\,m^{-2}\,^{\circ}C^{-1}}$, σ = 1.35 $\mathrm{W\,yr\,m^{-2}\,^{\circ}C^{-1}}$), C_0 (mean = 109 $\mathrm{W\,yr\,m^{-2}\,^{\circ}C^{-1}}$, σ = 78 $\mathrm{W\,yr\,m^{-2}\,^{\circ}C^{-1}}$), and γ (mean =

0.67 $\mathrm{W\,m^{-2}\,{}^{\circ}C^{-1}}$, $\sigma = 0.225$ $\mathrm{W\,m^{-2}\,{}^{\circ}C^{-1}}$), while a lognormal distribution is defined for ε (mean = 1.28, $\sigma = 0.375$) to ensure that values are larger than 1, consistent with the expectation that radiative feedbacks will weaken as equilibrium is approached (Armour, 2017; Andrews et al., 2015; Dong et al., 2019, 2020). The distribution of γ is truncated to exclude values less than 0.1 $\mathrm{W\,m^{-2}\,{}^{\circ}C^{-1}}$, while C_0 is truncated to exclude values less than 10 $\mathrm{W\,yr\,m^{-2}\,{}^{\circ}C^{-1}}$.

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.

Armour, K. C. (2017). Energy budget constraints on climate sensitivity in light of inconstant climate feedbacks. *Nature Climate Change*, *7*(5), 331–335.

Armour, K. C., Proistosescu, C., Dong, Y., Hahn, L. C., Blanchard-Wrigglesworth, E., Pauling, A. G., ... others (2024). Sea-surface temperature pattern effects have slowed global warming and biased warming-based constraints on climate sensitivity. *Proceedings of the National Academy of Sciences*, *121*(12), e2312093121.

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.

Dong, Y., Proistosescu, C., Armour, K. C., & Battisti, D. S. (2019). Attributing historical and future evolution of radiative feedbacks to regional warming patterns using a

Green's function approach: The preeminence of the Western Pacific. *Journal of Climate*, *32*(17), 5471–5491.

Geoffroy, O., Saint-Martin, D., Bellon, G., Voldoire, A., Olivié, D. J., & Tytéca, S. (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*(6), 1859–1876.

Geoffroy, O., Saint-Martin, D., Olivié, D. J., Voldoire, A., Bellon, G., & 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*(6), 1841–1857.

Table S1. Posterior parameter estimates for each GCM.

C	C_0	λ_{eq}	ECS	γ	ε
$\text{W yr m}^{-2}\,^{\circ}\text{C}^{-1}$	$\text{W yr m}^{-2}\,^{\circ}\text{C}^{-1}$	$\text{W m}^{-2}\,^{\circ}\text{C}^{-1}$	°C	$\text{W m}^{-2}\,^{\circ}\text{C}^{-1}$	unitless
MIROC6					
8.3	139	-1.46	2.58	0.76	1.30
7.0, 9.6	77, 200	-1.30, -1.61	2.31, 2.85	0.61, 0.91	1.19, 1.40
8.9	*175*	*-1.38*	*2.60*	*0.65*	*1.32*
CanESM5					
7.9	105	-0.70	5.42	0.55	1.30
6.8, 9.2	39, 171	-0.61, -0.77	4.77, 6.06	0.41, 0.69	1.19, 1.42
8.0	*80*	*-0.65*	*5.64*	*0.52*	*1.07*
IPSL-CM6A-LR					
8.2	117	-1.11	3.40	0.65	1.28
6.9, 9.5	56, 178	-0.97, -1.25	2.96, 3.84	0.48, 0.82	1.18, 1.39
8.2	*63*	*-0.75*	*4.56*	*0.41*	*1.33*
HadGEM3-GC31-LL					
8.2	127	-0.67	5.68	0.75	1.30
6.9, 9.5	63, 190	-0.55, -0.79	4.64, 6.72	0.54, 0.95	1.20, 1.40
8.0	*77*	*-0.63*	*5.55*	*0.51*	*1.22*

Note. For each GCM, ensemble means of posterior parameter estimates are shown in the first row. The second row provides values for one standard deviation above and below the mean. The third row shows parameter values derived from the *abrupt4xCO*$_2$ simulations (Armour et al., 2024) in italics.

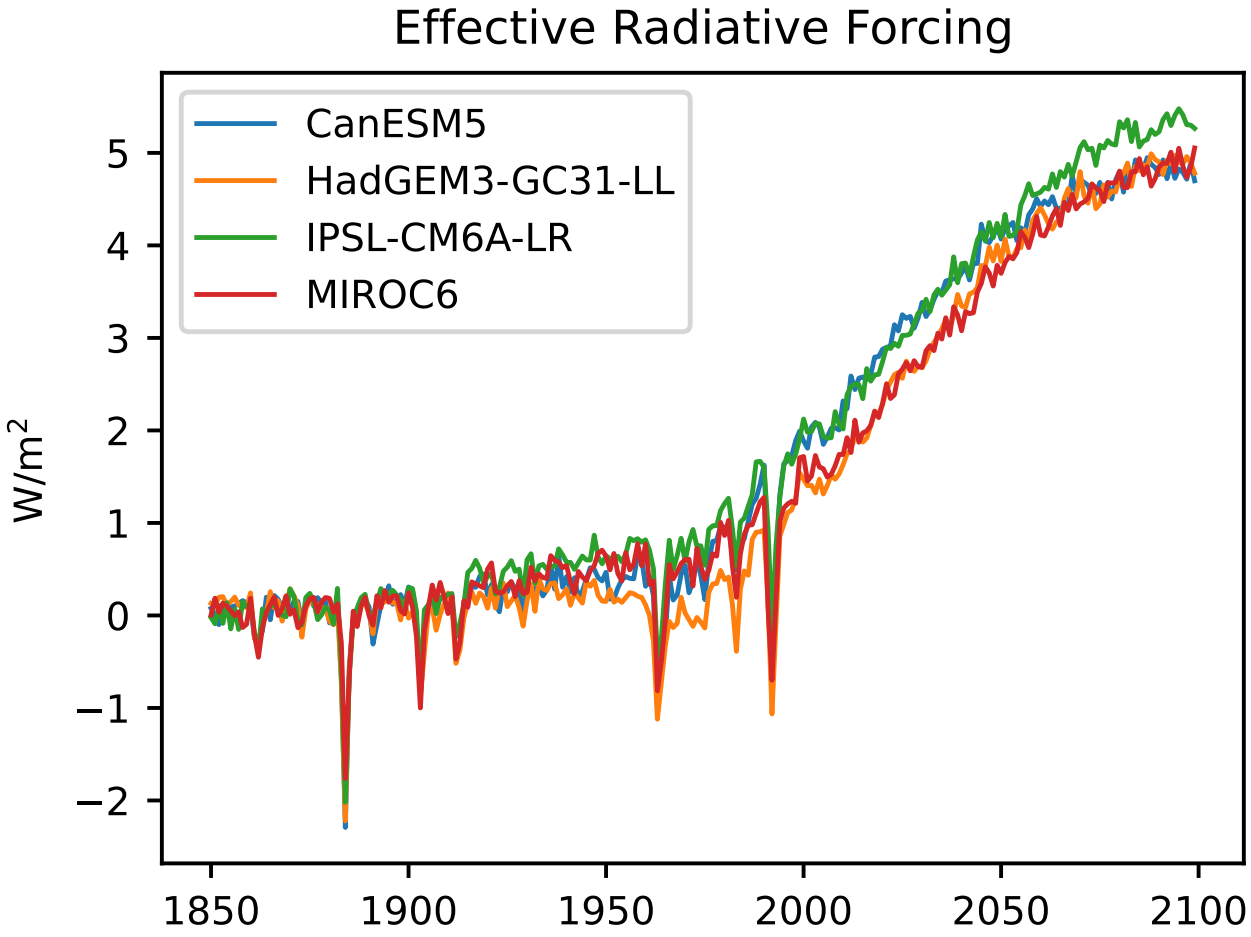

Figure S1. Time series of effective radiative forcing (ERF) estimated from RFMIP *piClim-histall* simulations for each model.

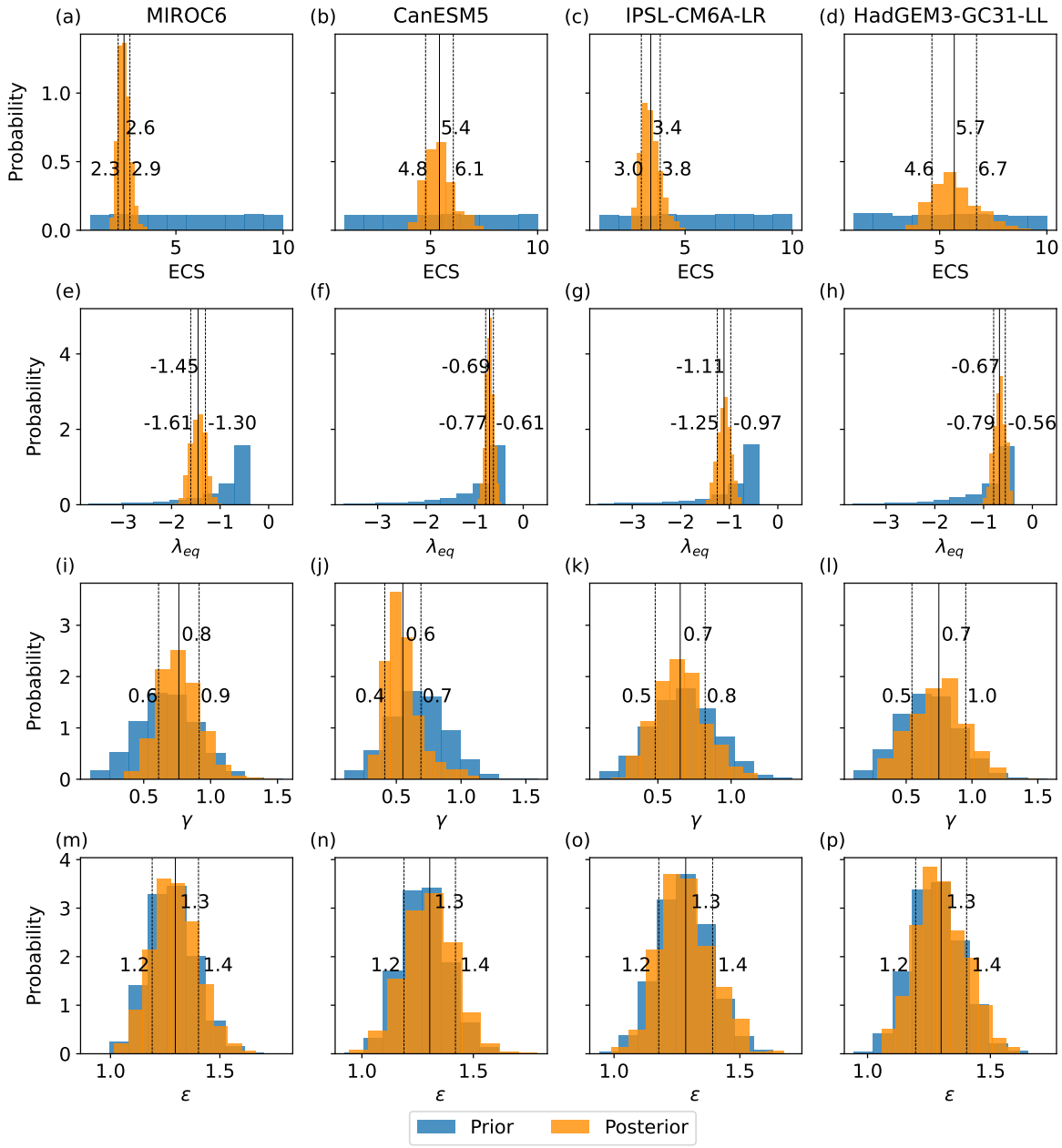

Figure S2. Prior (blue) and posterior (orange) distributions of (a-d) equilibrium climate sensitivity, (e-h) equilibrium radiative feedback parameter λ_{eq} (i-l) coefficient of vertical heat exchange γ and (m-p) heat uptake efficacy ϵ for the models MIROC6, CanESM5, IPSL-CM6A-LR and HadGEM3-GC31-LL. The mean estimate is shown by the solid black line and one standard deviation on either side of the mean is given by the dashed black lines. Note that λ_{eq} is skewed towards lower magnitude values as it depends inversely on ECS.

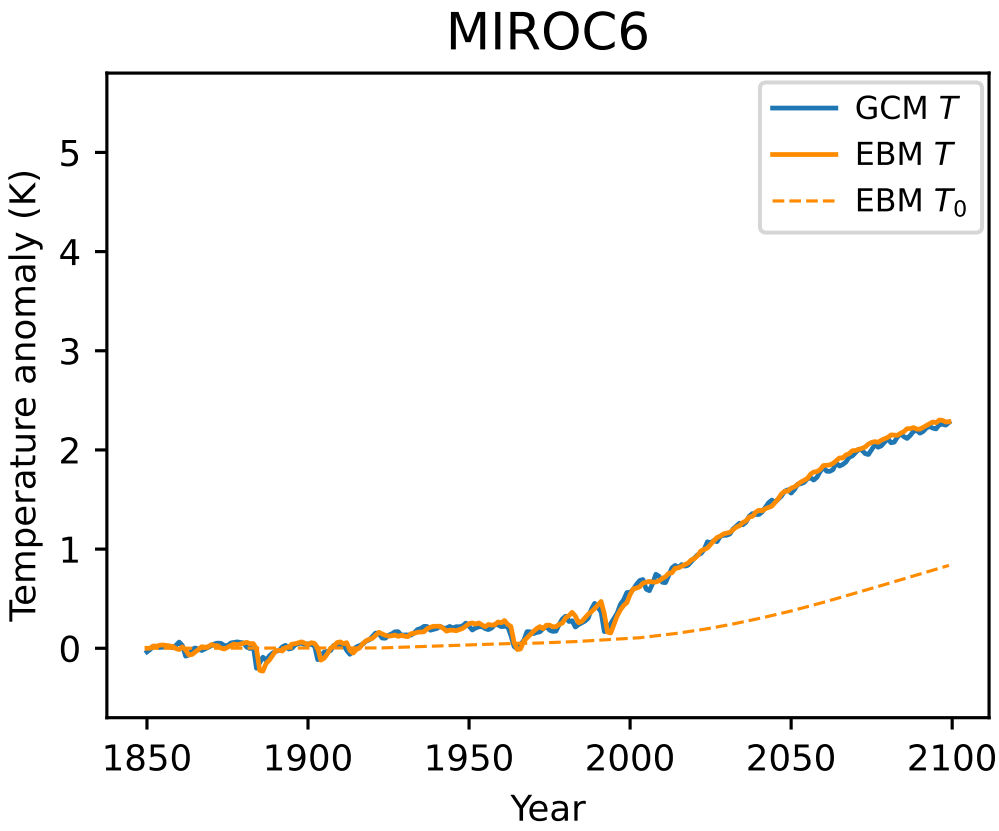

Figure S3. Time series of the deep layer temperature anomaly T_0 (dashed orange line) in comparison with the top layer (solid orange line) in the calibrated EBM, and the near-surface air temperature anomaly in the GCM (solid blue line) for MIROC6. All time series are ensemble means.

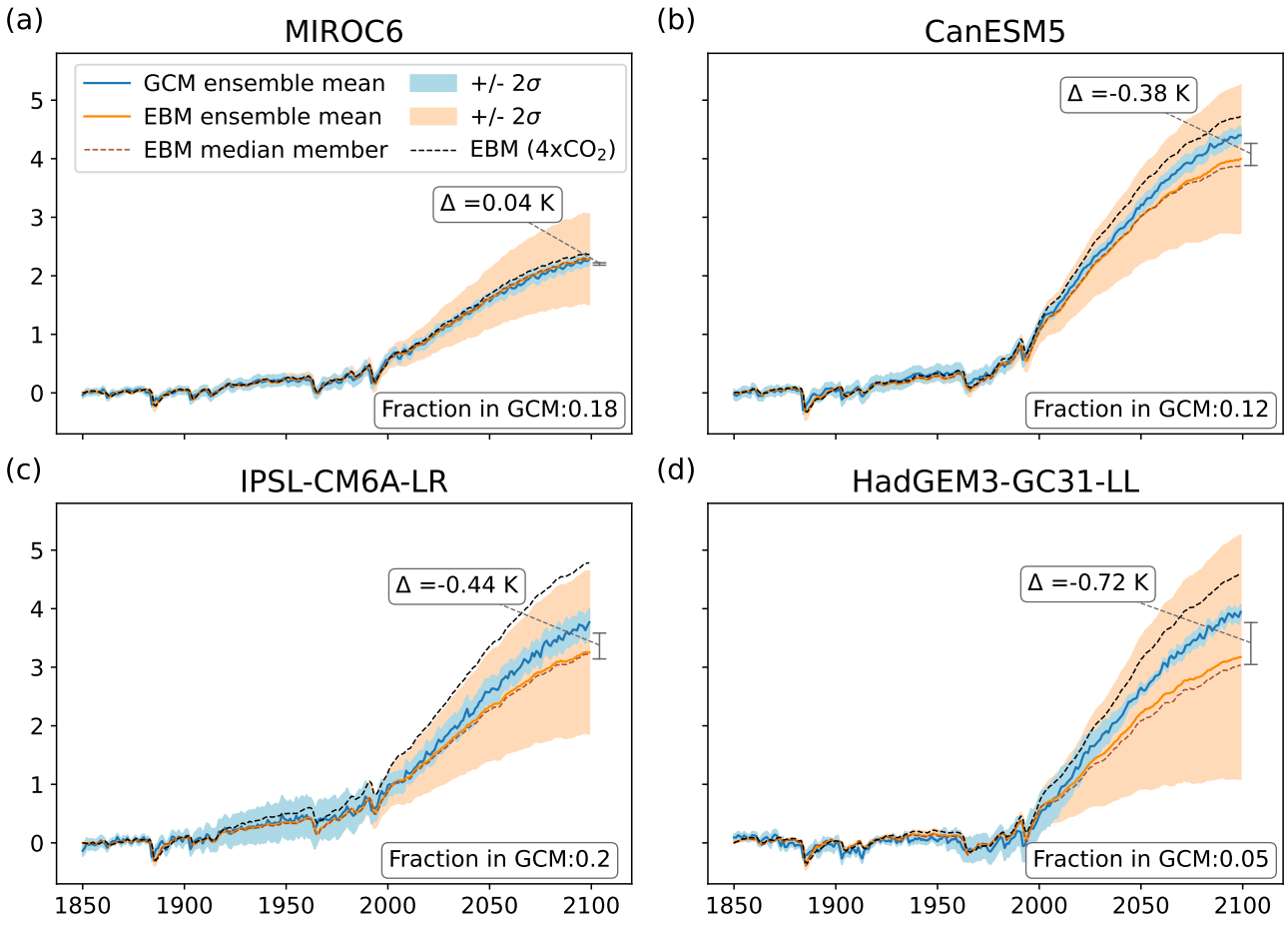

Figure S4. Similar to Figure 1, global and annual mean temperature anomaly time series for the EBM constrained to 4 CMIP6 GCMs, except the constraint is applied during the period 1980-2000. EBM ensemble mean (orange line) is shown in comparison to the corresponding GCM ensemble mean (blue line) with two standard deviations across the ensemble members shaded around the ensemble mean in light orange and blue, respectively. The brown dashed line represents the median member of the EBM ensemble.

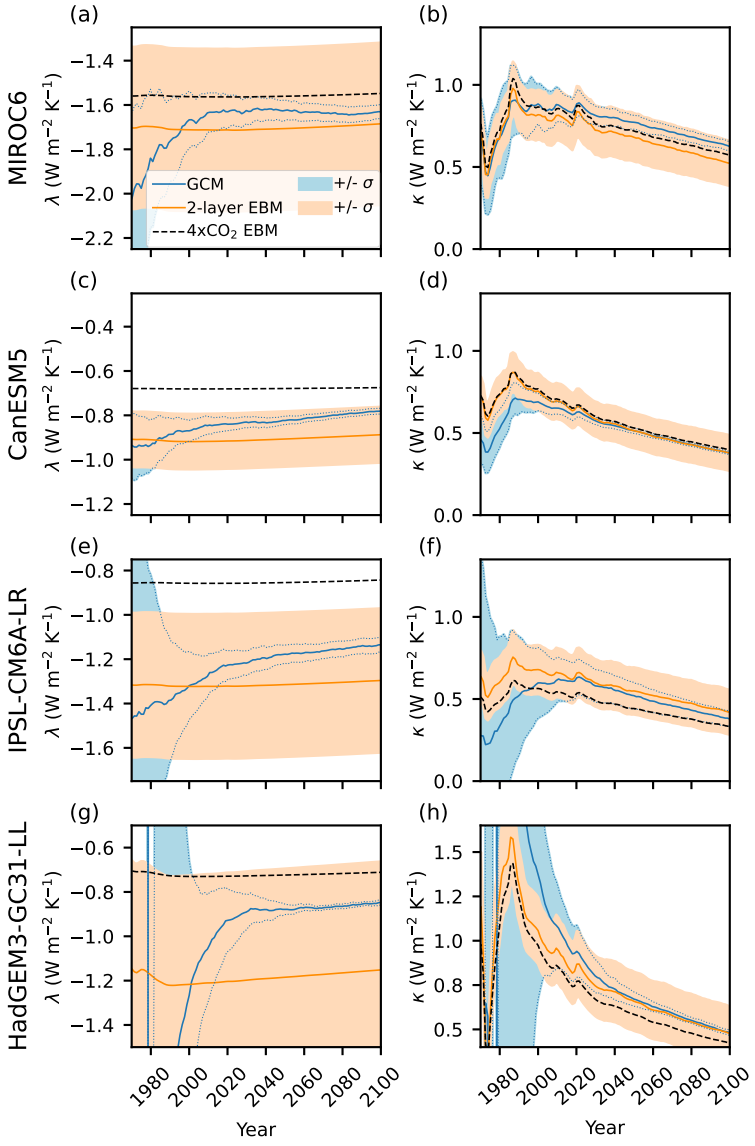

Figure S5. Similar to Figure 2, 40-year running means of (a-d) radiative feedback parameter λ and (e-h) ocean heat uptake efficiency κ calculated as the ensemble mean of each GCM (blue) and calibrated EBM (orange), except the calibration is done during the period 1980-2000. The blue and orange shading represent one standard deviation about the mean for the GCM and EBM respectively, with the blue dotted lines bounding the spread in the GCM feedbacks for clarity. The evolution of λ and κ fit to the *abrupt4xCO*$_2$ simulation of each GCM is shown in the dashed black line.